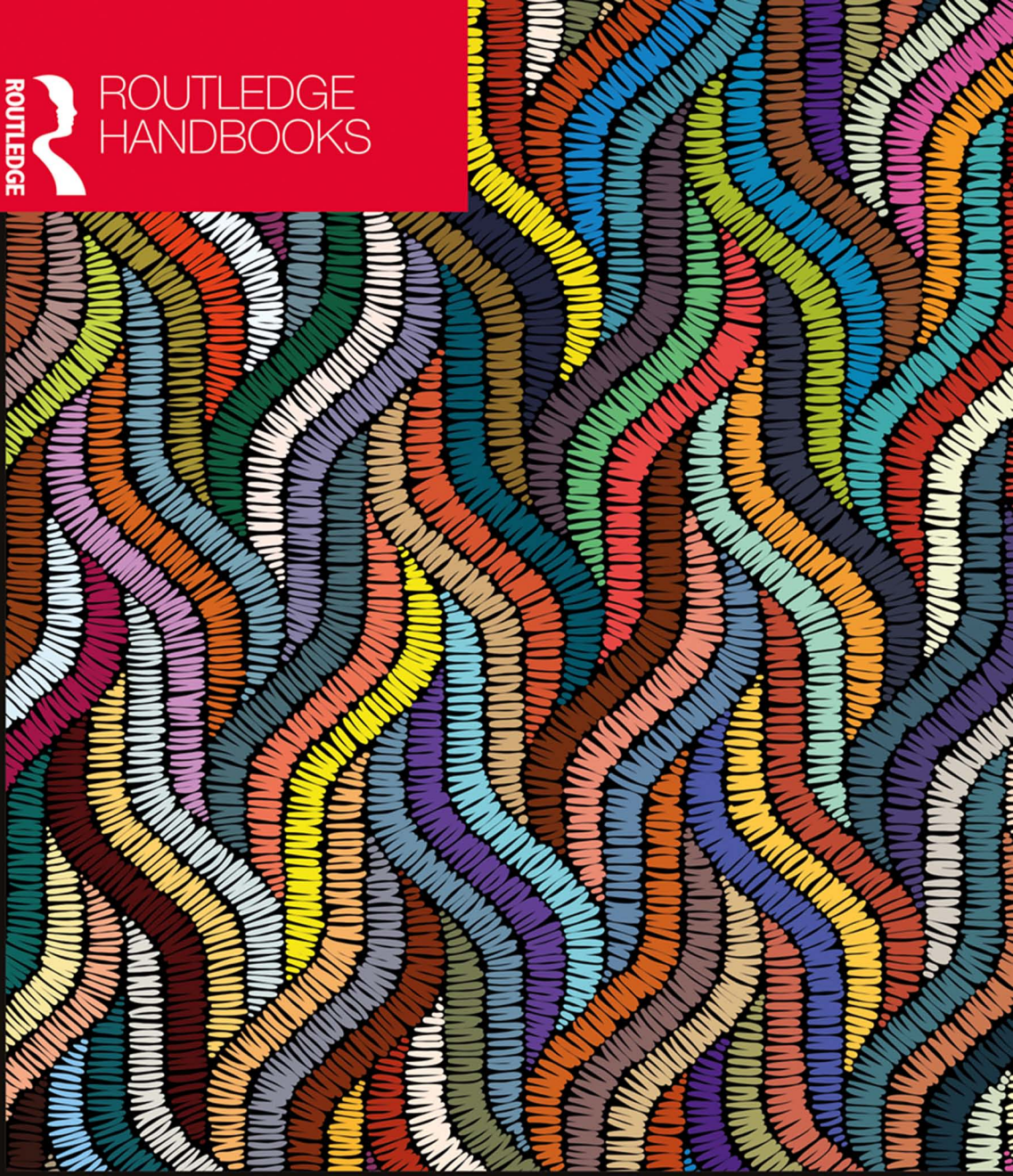
ROUTLEDGE HANDBOOKS



The Routledge Handbook of Language Testing

Second Edition

Edited by Glenn Fulcher and Luke Harding

The Routledge Handbook of Language Testing

This second edition of *The Routledge Handbook of Language Testing* provides an updated and comprehensive account of the area of language testing and assessment.

The volume brings together 35 authoritative articles, divided into ten sections, written by 51 leading specialists from around the world. There are five entirely new chapters covering the four skills: reading, writing, listening, and speaking, as well as a new entry on corpus linguistics and language testing. The remaining 30 chapters have been revised, often extensively, or entirely rewritten with new authorship teams at the helm, reflecting new generations of expertise in the field. With a dedicated section on technology in language testing, reflecting current trends in the field, the *Handbook* also includes an extended epilogue written by Harding and Fulcher, contemplating what has changed between the first and second editions and charting a trajectory for the field of language testing and assessment.

Providing a basis for discussion, project work, and the design of both language tests themselves and related validation research, this *Handbook* represents an invaluable resource for students, researchers, and practitioners working in language testing and assessment and the wider field of language education.

Glenn Fulcher is Professor of Applied Linguistics and Language Assessment at the University of Leicester (UK). He has served as president of the International Language Testing Association and as editor of the journal *Language Testing*. His Routledge book *Re-examining Language Testing* was joint winner of the SAGE/ILTA book award, together with the first edition of this *Handbook*.

Luke Harding is Professor in Linguistics and English Language at Lancaster University (UK). His research interests are in applied linguistics and language assessment, particularly assessing listening and speaking, World Englishes and English as a Lingua Franca, and language assessment literacy and professional ethics. He is currently co-editor of the journal *Language Testing*.

Routledge Handbooks in Applied Linguistics

Routledge Handbooks in Applied Linguistics provide comprehensive overviews of the key topics in applied linguistics. All entries for the handbooks are specially commissioned and written by leading scholars in the field. Clear, accessible and carefully edited *Routledge Handbooks in Applied Linguistics* are the ideal resource for both advanced undergraduates and postgraduate students.

THE ROUTLEDGE HANDBOOK OF CORPUS APPROACHES TO DISCOURSE ANALYSIS
Edited by Eric Friginal and Jack A. Hardy

THE ROUTLEDGE HANDBOOK OF WORLD ENGLISHES
Second Edition
Edited by Andy Kirkpatrick

THE ROUTLEDGE HANDBOOK OF LANGUAGE, GENDER AND SEXUALITY
Edited by Jo Angouri and Judith Baxter

THE ROUTLEDGE HANDBOOK OF PLURILINGUAL LANGUAGE EDUCATION
Edited by Enrica Piccardo, Aline Germain-Rutherford and Geoff Lawrence

THE ROUTLEDGE HANDBOOK OF THE PSYCHOLOGY OF LANGUAGE LEARNING AND TEACHING
Edited by Tammy Gregersen and Sarah Mercer

THE ROUTLEDGE HANDBOOK OF LANGUAGE TESTING
Second Edition
Edited by Glenn Fulcher and Luke Harding

For a full list of titles in this series, please visit www.routledge.com/series/RHAL

The Routledge Handbook of Language Testing

Second Edition

Edited by Glenn Fulcher and Luke Harding

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

Second edition published 2022
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

and by Routledge
605 Third Avenue, New York, NY 10158

Routledge is an imprint of the Taylor & Francis Group, an informa business

© 2022 selection and editorial matter, Glenn Fulcher and Luke Harding; individual chapters, the contributors

The right of Glenn Fulcher and Luke Harding to be identified as the authors of the editorial material, and of the authors for their individual chapters, has been asserted in accordance with sections 77 and 78 of the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this book may be reprinted or reproduced or utilised in any form or by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying and recording, or in any information storage or retrieval system, without permission in writing from the publishers.

Trademark notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

First edition published by Routledge 2012

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Fulcher, Glenn, editor. | Harding, Luke, editor.
Title: The Routledge handbook of language testing / edited by Glenn Fulcher and Luke Harding.
Description: Second edition. | London ; New York : Routledge, 2021. | Series: Routledge handbooks in applied linguistics | Includes bibliographical references and index.
Identifiers: LCCN 2021021439 | ISBN 9781138385436 (hardback) | ISBN 9781032116501 (paperback) | ISBN 9781003220756 (ebook)
Subjects: LCSH: Language and languages—Ability testing. | LCGFT: Essays.
Classification: LCC P53.4 .R68 2021 | DDC 401/.93—dc23
LC record available at https://lccn.loc.gov/2021021439

ISBN: 978-1-138-38543-6 (hbk)
ISBN: 978-1-032-11650-1 (pbk)
ISBN: 978-1-003-22075-6 (ebk)

DOI: 10.4324/9781003220756

Typeset in Bembo
by Apex CoVantage, LLC

Contents

List of figures *ix*
List of tables *x*
List of contributors *xi*
Acknowledgements *xviii*

Editorial 1
Glenn Fulcher and Luke Harding

SECTION 1
Validity **15**

1 Conceptions of validity 17
Carol A. Chapelle and Hye-won Lee

2 Articulating a validity argument 32
Michael T. Kane

3 Inference and prediction in language testing 48
Steven J. Ross

SECTION 2
The uses of language testing **61**

4 Social dimensions of language testing 63
Richard F. Young

5 Designing language tests for specific purposes 81
Carol Lynn Moder and Gene B. Halleck

6 Revisiting language assessment for immigration and citizenship: The case of US citizenship and the Naturalization Test 96
Antony John Kunnan

SECTION 3
Classroom assessment and washback **117**

7 Classroom-based assessment 119
Janna Fox, Nwara Abdulhamid, and Carolyn E. Turner

8 Washback: Looking backward and forward 136
Liying Cheng and Nasreen Sultana

9 Assessing young learners 153
Yuko Goto Butler

10 Dynamic assessment 171
Marta Antón and Próspero N. García

11 Diagnostic assessment in language classrooms 187
Eunice Eunhee Jang and Jeanne Sinclair

SECTION 4
Assessing the language skills **207**

12 Assessing speaking 209
Fumiyo Nakatsuhara, Nahal Khabbazbashi, and Chihiro Inoue

13 Assessing listening 223
Elvis Wagner

14 Assessing writing 236
Ute Knoch

15 Assessing reading 254
Tineke Brunfaut

SECTION 5
Test design and administration **269**

16 Test specifications 271
Yan Jin

17 Evidence-centered design in language testing 289
Chengbin Yin and Robert J. Mislevy

18 Accommodations and universal design 306
Jamal Abedi

19 Rater and interlocutor training 322
Larry Davis

SECTION 6
Writing items and tasks 339

20 Item writing and item writers 341
Dongil Shin

21 Writing integrated tasks 357
Lia Plakans

22 Test-taking strategies and task design 372
Andrew D. Cohen

SECTION 7
Prototyping and field tests 397

23 Prototyping new item types 399
Susan Nissan and Elizabeth Park

24 Pre-operational testing 415
Benjamin Kremmel, Kathrin Eberharter, and Franz Holzknecht

25 Piloting vocabulary tests 430
John Read

SECTION 8
Measurement theory in language testing 445

26 Classical test theory 447
James Dean Brown

27 Item response theory and many-facet Rasch measurement 462
Gary J. Ockey

28 Reliability and dependability 477
Xun Yan and Jason Fan

29 Scoring performance tests 495
Evelina D. Galaczi and Gad S. Lim

SECTION 9
Technology in language testing 511

30 Validity and the automated scoring of performance tests 513
Xiaoming Xi

31 Computer-based testing 530
Yasuyo Sawaki

32 Corpus linguistics and language testing 545
Sara T. Cushing

SECTION 10
Ethics, fairness, and policy 561

33 Ethics and fairness 563
F. Scott Walters

34 Standards in language proficiency measurement 578
Bart Deygers

35 Quality management in test production and administration 597
Nick Saville and Sarah McElwee

36 Epilogue: language testing: where are we heading? 622
Luke Harding and Glenn Fulcher

Index *633*

Figures

3.1	Scatterplot of grade-point average with truncated aptitude	54
3.2	Scatterplot of grade-point average with low-aptitude trainees	55
3.3	Parallel growth model for proficiency and aptitude	56
6.1	An example of the literacy test in Italian	103
10.1	Sample CDA learner profile	179
14.1	Process of writing assessment	237
14.2	Levels of writing development	247
16.1	A comparison of specification-driven and effect-driving testing	283
17.1	Toulmin's general structure for arguments	290
17.2	Extended Toulmin diagram for assessment	292
23.1	Schematic table	404
23.2	Summary task	405
25.1	Sample item from the WAT (translated from Dutch)	438
27.1	ICCs for three items from the 1PL Rasch model	466
27.2	ICCs for 1PL, 2PL, and 3PL	469
27.3	Category response curves for writing example	471
28.1	Some sources of error in a test	483
28.2	Two sample multiple-choice items measuring vocabulary	484
28.3	Scatterplots of test scores on two administrations	485
32.1	Output of CLAWS tagger	548
32.2	Output of USAS tagger	549
32.3	Output from Stanford parser	550
32.4	Output of concordance on *truth*	551
33.1	Worksheet grid	574
35.1	Core processes and outputs of the assessment cycle	606
35.2	The assessment cycle	608
35.3	The assessment cycle showing periodic test review	613

Tables

1.1	Characteristics defining validity as presented by Messick (1989)	19
1.2	Types of evidence pertaining to test score interpretations and actions from Messick (1989) and in the *Standards* (1999, 2014)	20
1.3	Types of claims in validity arguments and the inferences leading to each type	22
6.1	Applications, naturalizations, and denials, 2015–2019	109
6.2	Naturalizations from six leading countries by year and percentage, 2015–2019	109
10.1	Pre-scripted prompts for classroom DA	177
10.2	Functions of the scores and learner profile automatically generated by the C-DA tests	179
14.1	Dimensions and considerations in writing assessment	238
14.2	Continuum of writing constructs in different assessment types	245
14.3	Continuum of level of detail in scoring	245
14.4	Continuum of score and feedback provision related to different assessment types	245
16.1	The SAT essay domain	275
16.2	A summary of the rubric of the SAT Essay Test	276
16.3	Specification of some features of the CET-SET Band 4	276
17.1	Summary of ECD layers in the context of language testing	291
17.2	Design pattern attributes and relationships to assessment argument	293
17.3	A design pattern for assessing cause-and-effect reasoning reading comprehension	294
17.4	Steps taken to redesign TOEFL iBT and TOEIC speaking and writing tests and guided by layers in evidence-centered design	299
27.1	Test taker response on a multiple-choice listening test	464
28.1	Classification of test takers on two test administrations	487
28.2	An example of classification of test takers for computing kappa coefficient	488
34.1	PISA reading performance levels	580
34.2	Kaulfers's aural performance scale	582
34.3	FSI/ILR speaking descriptors (1968)	583
34.4	ACTFL alignment with ILR scales and CEFR levels	584
34.5	Organization of CEFR levels	585
34.6	Overview of the CLB and their relationship with the CEFR levels	587
34.7	CSE overview and CEFR link	588
34.8	Minimum English language proficiency requirements for international students in arts and humanities	589
34.9	Score equivalence guides compared	589

Contributors

Nwara Abdulhamid, PhD, completed her doctoral study of high-stakes testing, curricular alignment, and washback on teaching and learning in Applied Linguistics at Carleton University. She currently engages in research regarding language education, teacher education and educational assessment, and relationships between misalignment and washback in high-stakes testing contexts.

Jamal Abedi is a professor at the School of Education of the University of California, Davis, and a research partner at the National Center for Research on Evaluation, Standards, and Student Testing. His research interests include accommodations and classification for English language learners and comparability of alternate assessments for students with significant cognitive disabilities.

Marta Antón is Professor of Spanish at Indiana University-Purdue University Indianapolis. Her main research focus is the application of sociocultural theory to second language learning and teaching in the classroom, including teacher–learner interaction, peer interaction, and dynamic assessment. Other interests are classroom-based assessment, authentic assessment, and Spanish sociolinguistics.

James Dean Brown is Professor Emeritus in the Department of Second Language Studies at the University of Hawai'i at Mānoa. He has spoken and taught courses in places ranging from Albuquerque to Zagreb and has published numerous articles and books on language testing, curriculum design, research methods, and connected speech.

Tineke Brunfaut is Professor in Linguistics and English Language at Lancaster University, UK. Her main research interests are language testing and reading and listening in a second/foreign language. In recent research, she has focused on cognitive and affective factors and processes in language testing, methodology in language testing research, and language test development.

Yuko Goto Butler is Professor of Educational Linguistics in the Graduate School of Education at the University of Pennsylvania. She is also the director of the Teaching English to Speakers of Other Languages (TESOL) program at Penn. Her research interests include language assessment and second and bilingual language learning among children.

Carol A. Chapelle is Distinguished Professor in Liberal Arts and Sciences at Iowa State University. Recent books include *Validity Argument in Language Testing: Case Studies of*

Validation Research (Cambridge, 2021; with E. Voss), *Argument-Based Validation in Testing and Assessment* (Sage, 2021), and *The Handbook of Technology and Language Learning and Teaching* (Wiley, 2017; with S. Sauro).

Liying Cheng is Professor and Director of the Assessment and Evaluation Group (AEG) at the Faculty of Education, Queen's University. Her primary research interests are the impact of large-scale testing on instruction, the relationships between assessment and instruction, and the academic and professional acculturation of international and new immigrant students, workers, and professionals. Her seminal research on washback focuses on the global impact of large-scale testing.

Andrew D. Cohen, Professor Emeritus, University of Minnesota, has conducted research for many years on language learner strategies, including a focus on how learners respond to tests of reading (cloze and multiple choice) and oral language skills (e.g., skills in pragmatics). He uses verbal report techniques to explore test response strategies.

Sara T. Cushing is Professor of Applied Linguistics at Georgia State University. She has published research in the areas of language assessment, second language writing, and teacher education. She has been invited to speak and conduct workshops on second language assessment throughout the world, most recently in Vietnam, Colombia, and Thailand.

Larry Davis is a research scientist in the Center for Language Education, Assessment, and Research (CLEAR) in the research and development division of Educational Testing Service. His interests broadly encompass the assessment of speaking, including creation of rubrics; rater cognition; automated evaluation of speech; and assessment of spoken interaction.

Bart Deygers is an assistant professor of second language didactics and assessment at Ghent University, Belgium. In recent research, he has focused on language testing policy for migration purposes, the use of the CEFR in migration policy, and the impact of language requirements on low-literate migrants.

Kathrin Eberharter holds a research position at the University of Innsbruck. Her main interests are in the area of assessing L2 writing and speaking, with a focus on task development, rating scale development, rater cognition, and training. She is also interested in cognitive processes in language testing and language assessment literacy.

Jason Fan is the deputy director and senior research fellow at the Language Testing Research Centre at the University of Melbourne. His research interests include test validity and validation, research methods, and language assessment literacy.

Janna Fox is Professor Emerita of Applied Linguistics and Discourse Studies in the School of Linguistics and Language Studies, Carleton University. Her research interests include language assessment, within and external to the classroom; washback and impact on teaching and learning; and sociocultural, transdisciplinary perspectives in validation research.

Glenn Fulcher is Professor of Applied Linguistics and Language Assessment at the University of Leicester (UK). He has served as president of the International Language

Testing Association and as editor of the journal *Language Testing*. His Routledge book *Re-examining Language Testing* was joint winner of the SAGE/ILTA book award together with the first edition of this *Handbook*.

Evelina D. Galaczi is Head of Research Strategy at Cambridge Assessment English and has worked in language education for over 30 years as a researcher, assessment specialist, teacher, and materials writer. Evelina's expertise lies in speaking assessment, interactional competence, test development, and the use of technologies in assessing productive skills.

Próspero N. García is associate professor of Spanish applied linguistics at Rutgers University, Camden. His research focuses on sociocultural psychology applied to heritage and second language acquisition and pedagogy, language evaluation and assessment, teacher's cognition, and technology-enhanced language learning.

Gene B. Halleck retired as Professor of Linguistics/TESL at Oklahoma State University. She has designed a set of diagnostic tests, the Oral Proficiency Tests for Aviation, to accompany an aviation English curriculum designed for the US Federal Aviation Administration.

Luke Harding is Professor in Linguistics and English Language at Lancaster University (UK). His research interests are in applied linguistics and language assessment, particularly assessing speaking and listening, World Englishes and English as a Lingua Franca, and language assessment literacy and professional ethics. He is currently co-editor of the journal *Language Testing*.

Franz Holzknecht is a researcher in applied linguistics and language testing at the University of Applied Sciences in Special Needs Education, Zurich, Switzerland. His main research interests are in language testing, particularly in the areas of sign language testing; testing second language listening; and language test development.

Chihiro Inoue is Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her main research interests are in designing tasks and rating scales for L2 speaking and the effects of task-related variables on learner language.

Eunice Eunhee Jang is Professor at Ontario Institute for Studies in Education, University of Toronto. Her research areas include diagnostic assessment for supporting students and teachers, validation of English proficiency descriptors-based assessment for K–12 English language learners and scenario-based language assessment, evaluation of school effectiveness in challenging circumstances, and technology-infused assessment system design and validation.

Yan Jin is Professor of Applied Linguistics of the School of Foreign Languages at Shanghai Jiao Tong University, China. Her research interests are in the development and validation of large-scale and high-stakes language assessments. She has been chair of the National College English Testing Committee since 2004.

Michael T. Kane is the Messick Chair in validity at the Educational Testing Service. His research interests are validity theory, generalizability theory, and standard setting. Dr. Kane

holds a PhD in education, an MS in statistics from Stanford University, and a BS in physics from Manhattan College.

Nahal Khabbazbashi is Senior Lecturer in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her main research interests are the assessment of speaking, the use and impact of technology in assessment, and multimodal communication.

Ute Knoch is Associate Professor in Language Assessment and the Director of the Language Testing Research Centre at the University of Melbourne. Her research interests are in the areas of writing assessment, rating processes, rating scale development, test validation, assessing languages for academic and professional purposes, and placement testing.

Benjamin Kremmel is head of the Language Testing Research Group Innsbruck (LTRGI) at the University of Innsbruck, Austria, where he teaches and researches language learning, teaching, and assessment. His research interests include vocabulary assessment, language assessment literacy, diagnostic language testing, and the interface between language testing and SLA.

Antony John Kunnan (PhD, UCLA) is a principal assessment scientist at Duolingo, Inc. He specializes in the areas of fairness, validation, and ethics. He is the author and editor of over 90 publications; he recently authored a book titled *Evaluating Language Assessments* (Routledge, 2018). He is a past president of the International Language Testing Association and the founding president of the Asian Association for Language Assessment. He is also the founding editor of *Language Assessment Quarterly*.

Hye-won Lee is Senior Research Manager at Cambridge Assessment English. She holds a PhD in Applied Linguistics and Technology from Iowa State University, with specialization in technology-enhanced language assessment and argument-based validation. Her current work focuses on assessing speaking via video call and defining language ability in data-driven diagnostic assessment.

Gad S. Lim is Director of Assessment at Michigan Language Assessment. His areas of interest include the design of learning-oriented assessment, the development and setting of standards, and the application of these in education policy and reform.

Sarah McElwee is Principal Research Manager in the Research and Thought Leadership group at Cambridge Assessment English, University of Cambridge. Her research interests include consequential validity; cognitive factors in testing, particularly related to younger learners; and the role of validation in product lifecycle management for assessment.

Robert J. Mislevy is the Frederic M. Lord Chair in Measurement and Statistics at ETS and Professor Emeritus of Measurement, Statistics, and Evaluation at the University of Maryland, with affiliation in second language acquisition. His research applies developments in statistics, technology, and cognitive science to practical problems in educational assessment.

Carol Lynn Moder is Professor of Linguistics and TESOL at Oklahoma State University. She has designed ESP testing and curriculum materials for International Teaching Assistants,

English for Academic Purposes, and aviation English. She designed aviation English assessments for the international training division of the US Federal Aviation Administration and for Ordinate Corporation. She served as a consultant to ICAO in 2005.

Fumiyo Nakatsuhara is Reader in Language Assessment at the Centre for Research in English Language Learning and Assessment (CRELLA), University of Bedfordshire. Her main research interests include the nature of co-constructed interaction in speaking tests, task design, rating scale development, and the relationship between listening and speaking skills.

Susan Nissan is Senior Director at Educational Testing Service (retired) in Princeton, New Jersey. In addition to assessment development, her research interests include test design, test validity, standard setting, assessment literacy, and the relationships between assessment and learning.

Gary J. Ockey is a professor of Applied Linguistics and Technology at Iowa State University. He is interested in the use of technology for aiding in the assessment of second language oral communication and the applications of various quantitative statistical approaches to facilitate the understanding and practice of second language education and assessment.

Elizabeth Park is an assessment specialist at Educational Testing Service. She designs English language proficiency assessments and professional development programs for teachers of English as a second/foreign language.

Lia Plakans is a professor of Multilingual Education at the University of Iowa in the US. Her research interests include integrated skills assessment, assessing reading and writing, test use, and bilingual assessments in school settings.

John Read is an emeritus professor in Applied Language Studies at the University of Auckland. His research interests include the design of second language vocabulary tests, academic literacy and the post-admission assessment of university students, and testing English for occupational purposes, particularly in aviation and the health sciences.

Steven J. Ross is a professor of second language acquisition at the University of Maryland. His recent research interests are assessment of second language pragmatics, validity theory, language assessment, and longitudinal research methods. Professor Ross completed a PhD in second language acquisition at the University of Hawai'i Manoa in 1995.

Nick Saville is Director of Research and Thought Leadership (Cambridge Assessment English, University of Cambridge) and Secretary-General of the Association of Language Testers in Europe. His research interests include assessment and learning in the digital age, the use of ethical AI, language policy and multilingualism, the CEFR, and learning-oriented assessment.

Yasuyo Sawaki is Professor of Applied Linguistics at Waseda University in Tokyo, Japan. She has research interests in diverse topics in second/foreign language assessment, including test validation, assessment and instruction of summary writing, diagnosing language ability, and using technology in language assessment.

Dongil Shin is Professor of Applied Linguistics in the Department of English Language and Literature, Chung-Ang University, Korea. Since completing his doctorate at University of Illinois at Urbana-Champaign, he has researched language testing (policies), (critical) discourse analysis, and language ideologies and subjectivities, mostly in Korean contexts. He tries to explore critical discursive approaches to language testing in newly emerging multilingual (Korean) societies.

Jeanne Sinclair earned her PhD in 2020 from the Ontario Institute for Studies in Education (OISE) of the University of Toronto. She currently serves as senior psychometrician with the Nursing Community Assessment Service (British Columbia) and as postdoctoral fellow (OISE/McMaster University) researching child development and educational stratification.

Nasreen Sultana has earned PhD in Education from Queen's University, Kingston, with a dissertation exploring the relationship between washback and curriculum alignment. Currently, she is a full-time teaching and learning consultant at Conestoga College, Kitchener, Ontario. Nasreen's areas of interest include washback, intercultural pedagogy, equity, diversity, and inclusion.

Carolyn E. Turner is Associate Professor in the Department of Integrated Studies in Education, McGill University (retired). Her research interests include language assessment in educational and multi-lingual health professional settings, the impact of high-stakes tests on teaching and learning, learning-oriented assessment, empirically based rating scales, and mixed-methods research designs.

Elvis Wagner is Associate Professor of TESOL at Temple University, where he coordinates the PhD in Applied Linguistics program and the World Languages Education program. His research interests include the assessment of L2 oral ability, specifically focusing on how L2 listeners process and comprehend unscripted, spontaneous spoken language.

F. Scott Walters is Associate Professor in the Office of International Education at Seoul National University of Science and Technology. His research interests include the development and validation of L2 pragmatics tests, conversation analysis, and assessment-literacy development.

Xiaoming Xi is Chief of Product, Assessment, and Learning, VIPKID International. Her areas of interest include validity, validation frameworks, fairness frameworks, test design, speaking assessment, human scoring, rater issues, automated scoring and feedback, and the use of AI technology in learning and assessment.

Xun Yan is an associate professor of Linguistics, Second Language Acquisition and Teacher Education (SLATE), and Educational Psychology at the University of Illinois at Urbana-Champaign. His research interests include speaking and writing assessment, psycholinguistic approaches to language testing, and language assessment literacy.

Chengbin Yin specializes in Second and Foreign Language Education and Measurement, Statistics, and Evaluation at the University of Maryland, College Park. Her research interests are assessment design from a socio-cognitive perspective and measurement models in educational assessment.

Richard F. Young is Professor Emeritus of English and Second Language Acquisition at the University of Wisconsin-Madison. His research is concerned with the relationship between language and social context, with a particular focus on variation in interlanguage, morphology, talking and testing, language and interaction, and discursive practice in language teaching and learning.

Acknowledgements

We would first like to pay tribute to Fred Davidson, one half of the original editorial team. Fred's work on the first edition laid extremely strong foundations for the second. It was an honour to carry on this work.

Second, we thank our brilliant contributors. There were 51 authors involved in this second edition (not including us), and everyone has been a pleasure to work with. The atmosphere around this project has always been convivial and collaborative, and we thank all contributors for making our task easy.

We would also like to thank Hannah Rowe and Eleni Steck at Routledge, who have assisted at various stages of this project, Kate Fornadel at Apex CoVantage, and Geisa Davila Perez (Lancaster University) for some last-minute editorial assistance.

Finally, on a more personal note, Glenn thanks Jenny for sharing all the lovely things in life. And Luke thanks Rachael, Iris, and Saffron for being great company during the various lockdowns of 2020–2021, for their constant support, and for always making life fun.

Editorial

The first edition of *The Routledge Handbook of Language Testing* was published in 2012. In 2016, it was the joint winner of the SAGE/International Language Testing Association prize for the best book in language testing. The international selection committee provided the following citation to justify the award:

> "[T]he editors have succeeded in assembling a set of contributors with an unparalleled level of expertise in their respective areas, and with distinctive talents in communication. The strength of this extremely well-edited collection lies in the interweaving of theoretical and practical aspects of language testing through nine broad themes, and in the structuring of individual contributions to provide a historical perspective, a discussion of current issues and contributions, and a consideration of future directions. The volume stands not only to have a wide impact on best practice in the field, but also in the development of language assessment literacy in other professionals who find themselves involved in activities of language assessment."

Adoption by teachers and learners alike has seen the *Handbook* become the standard reference text in the field, but theory and research do not stand still, and in the last decade, the field has continued to flourish and expand with the use of new technologies, assessment purposes, and contexts of score use. The rapid expansion of research in language assessment literacy has also provided information about what is required in reference works that underpin successful pedagogy for language teachers, future language testers, and other stakeholders. In preparing for this second edition, we therefore not only took into account changes in research and assessment literacy needs, but also engaged with the publishers in a survey of users to discover what changes they would like to see. Perhaps the largest change is the introduction of Section 4 on assessing the language skills, as skills assessment remains a key aspect of many language testing programmes around the world, but there are numerous modifications to both content and focus throughout the volume.

DOI: 10.4324/9781003220756-1

We also welcome many new authors, reflecting the ever-expanding and international nature of language testing research. Some of the contributors to the first volume have retired, and we have sadly seen the passing of Alan Davies, a towering academic thinker who has left an outstanding legacy to the language testing field and literature. In the words of Marcus Aurelius, "Flows and changes are constantly renewing the world, just as the ceaseless passage of time makes eternity ever young" (Meditations 4, 36), and so it is with research and our understanding of assessment theory and practice. We have therefore strived to include a diverse and exciting authorship that retains the insight of established scholars alongside the novelty of recent innovations and insights from the next generation of researchers.

What has not changed is the editorial philosophy that made the first edition such a great success. We have always believed that the role of editors is to provide structure and direction to the volume and to aid authors in executing their arguments coherently and cohesively. The editor's role is not to direct content, arguments, or conclusions, although they may make recommendations that help strengthen chapter coverage. All too often, editors wish to shape publications in their own image, but this is not how fields develop or individuals learn. As J. S. Mill would argue, attempts to control reveal a presumption of infallibility, and for true progress, all facts and interpretations must be heard and discussed with open critical minds. Only in this manner can we ultimately provide warrants to justify an emerging consensus that at some future point will again be challenged, and so on endlessly (Mill, 1859: 25–26). Such is the nature of learning and progress.

This is, of course, the rationale for future editions of this *Handbook*, but we will refrain from crystal-ball gazing in the editorial and leave that to the predictions offered by our authors and to our Epilogue to this volume. We therefore turn to the updated content, which we believe provides the same authoritative, stimulating, and pedagogically useful platform as the first edition for investigating the field of language testing and assessment.

Content

Section 1 Validity

In Chapter 1, Chapelle and Lee provide an evolutionary account of validity theory and the practice of validation. As they make clear, there are many disparate ways of understanding validity. The one that researchers choose is either quite random, depending on their background and reading, or very deliberate, based on a commitment to an underpinning philosophy. Chapelle and Lee are unequivocal in adopting an argument-based approach as the lens through which to view the evolution of theory and practice, which carries with it an instrumental concern with the day-to-day practice of conducting validity research. They argue that it provides both a common language and the specific tools that the field needs to move forward. It is unquestionably the case that at the present time, the argument-based approach to validity and validation is in the ascendency, and it is therefore right and proper that this chapter appears first in the second edition of the *Handbook*. But readers should be aware that it is but one option in the current marketplace of ideas (Fulcher, 2015: 108–12).

Chapter 2 is the most natural progression from the position presented in Chapter 1, representing as it does the clearest possible statement of an argument-based approach to validation by its principal proponent, Michael T. Kane. He sets out the two fundamental building blocks of validation. First comes the interpretative/use argument (IUA), which establishes the proposed uses of test scores and the interpretations of those scores required for their

use. Second is the validity argument that evaluates the plausibility of the IUA by evaluating the evidence and rationales provided to support score use. The chapter explains how a test developer would go about building a confirmatory research agenda and a critical evaluator conduct studies that investigate alternative plausible interpretations of score meaning. The reader should note that Kane focuses on *observable attributes*, which are contrasted with *traits*. The latter are underlying theoretical abilities that account for observed behaviour, but in argument-based approaches, the preference is to focus on what can be observed and how consistent behaviours associated with target domains can be rendered into a score. Kane illustrates his approach with examples that will be familiar to language teachers and testers, making this chapter a highly accessible introduction to mainstream argument-based validation.

In Chapter 3, Ross begins his discussion with an acknowledgement that the argument-based approach to validation has done a great deal to help structure validation evidence in such a way that it can be evaluated in relation to counter-explanations. This is an important observation, as it pinpoints the main contribution made by the appropriation of Toulmin's argument structure to language test validation. But drawing inferences from evidence is often problematic, and Ross carefully analyses the pitfalls that validation researchers face with respect to both quantitative and qualitative data. Anyone who still believes that the outcomes of statistical analysis represent a universally generalisable truth should read and reread this chapter. Its treatment of how we make inferences from scores and, more generally, in research is one of the most masterful in the language testing literature, alongside Bachman (2006). Readers of the first edition of the *Handbook* will note that we have moved this revised chapter into the section on validity because of the major contribution this revised chapter makes to validation theory and practice.

Section 2 The uses of language testing

When we invited Richard F. Young to revise his chapter for the new edition of the *Handbook*, he declined – quite understandably – on account of now being retired. After some consideration, we decided to include the original chapter again (with Richard's consent, of course), with a few minor alterations. Following Cathie Elder's review of this chapter, we have also moved it into the first position within this section, as Chapter 4. There has been plenty of research on the consequences of language tests in the years since the first edition; much of this research is picked up in other chapters in the volume written by Deygers, Walters, and Cheng and Sultana, among others. Similarly, there have been important contributions to our understanding of interactional competence in language contexts. Again, this research is discussed in a new chapter by Nakatsuhara, Khabbazbashi, and Inoue. The reason that we decided to include this chapter again is that we believe it is utterly unique. Young's discussion is wide ranging and, given that it was written more than a decade ago, eerily prescient. Consider, for example, his final paragraph, in which he imagines what the future of language testing might look like: "an image of two psychometricians, experts in the field of educational measurement, sitting in front of a computer monitor scratching their heads as a waterfall of data pours down the screen." But we include this chapter, not because it is a historical artefact (after all, the interested reader could go back to the 2012 edition and read it there). Rather, we include it because it makes sense within this current collection and draws together various themes that are still highly relevant for language testing and assessment. Leaving this chapter out would have diminished the volume. If you haven't read it yet, we strongly encourage you to do so now.

It is arguably the case that language testing for specific purposes is the most high-stakes assessment that we find. In Chapter 5, Moder and Halleck consider the history and practice of aviation English language testing. Perhaps two critical issues from the many involved come to the fore. The first is the interaction, negotiation, and inevitable compromise between language testers, institutions that make policy, and very powerful stakeholders. The second is the requirement that domain-specific language tests be firmly grounded in an analysis of the language used in the workplace context as defined in the purpose of the test. This is the cutting edge of interaction between applied linguistic investigation and language assessment. Moder and Halleck demonstrate the critical role that language testing plays in maintaining safety for the public in very high-stakes social contexts, and their discussion generalises to many other situations in which the risk associated with error is particularly grave. During the COVID-19 pandemic, this language of error, risk, and consequence has become the staple of news items, and their consideration is no less serious in language assessment than it is in the field of medicine.

There is no topic in language testing more controversial than the use of tests for making decisions about immigration and citizenship. Despite many well-reasoned studies critiquing the use of language tests for these purposes – from both within and outside the field of language testing – the practice continues and appears to be proliferating across numerous countries. Is there a fair and ethical way to conduct language assessment for immigration and citizenship purposes? In Chapter 6, Kunnan revisits his contribution to the first edition of the *Handbook* by considering one particular gatekeeping exam in detail: the United States Naturalization Test. Kunnan provides a vivid description of the historical antecedents to the introduction of the test before scrutinising the current version, evaluating it against principles of fairness and justice. The test does not emerge well from this critique, and Kunnan's chapter demonstrates how a structured and systematic approach to fairness and justice can provide just the right tools for highlighting a flawed and potentially harmful testing regime.

Section 3 Classroom assessment and washback

For a long time, there has been an imbalance in language testing research, with a strong focus on large-scale international tests at the expense of a much more common activity: classroom-based language assessment. In classrooms around the world, teachers regularly set assessments, mark tests, provide feedback, and prepare learners for high-stakes exams. And yet there is so much we don't know or understand about assessment practices in these everyday contexts. In Chapter 7, Fox and Abdulhamid join Turner (who was the sole author of the chapter in the first edition) to provide an up-to-date overview of classroom-based assessment, charting what has changed over the past decade and what issues remain the same. This chapter illustrates that classroom-based language assessment has come a long way quickly, with newer unifying frameworks such as learning-oriented assessment helping drive a classroom-based assessment "turn." At the same time, technology is having an important impact on classroom assessment as digital tools provide some relief for teachers (e.g., automated writing evaluators), as well as a range of new challenges. The general paradigm shift toward a greater blending of external, large-scale testing and teacher assessment is a key feature of this chapter. In another ten years' time, we might expect to see technology mediating a much smoother alignment between classroom assessment and standardised assessment, though it will be interesting to observe how this is done and whose needs and objectives are prioritised: those of learners, teachers, policy makers, or test providers?

Washback is an area of language testing that has received sustained attention since Alderson and Wall's landmark paper in 1993, "Does washback exist?" Many of the hypotheses raised in that article have since been investigated empirically, and the findings have clarified our understanding. A fair summary would seem to be yes, washback exists, but it's complicated. Cheng and Sultana explore some of this complexity in Chapter 8, focusing on literature which has been published in the period since the first edition of the *Handbook*. The authors focus on three main trends in washback research over the past decade: the expansion of washback studies into previously under-researched educational contexts, the connection of washback in the classroom with broader social/educational contexts, and the adoption of a wider range of conceptual/theoretical frameworks to conduct washback research. A key theme that emerges in this chapter is the importance of "alignment": that is, the extent to which teaching, curriculum, and assessment are in agreement. An increasing number of studies now show that changing a test is not enough, on its own, to bring about positive washback. In fact, *negative* washback is the more likely result in contexts in which there is no alignment between the test, the curriculum goals, and classroom practices. We hope that policy makers in language education contexts will read this chapter and note that the introduction of a test, on its own, rarely solves problems and quite often creates new ones.

Since the first edition of the *Handbook*, assessment of young learners has expanded rapidly. We see this in the range of commercial tests now on the market targeted at children and teenagers, and we also see it in the number of research articles and new books focusing on young learners. The increased understanding of how best to assess young learners is a very welcome development, particularly given that learners under 18 are likely to comprise a large proportion of the number of language learners worldwide. In Chapter 9, Butler brings great expertise to this topic and provides a thorough and comprehensive overview of the current state of play in assessing this test-taker population. Butler highlights the challenges of dealing with great degrees of variability within the characteristics of young learners and also points toward the rapid shift of younger generations toward digital technologies. Previous reservations about test taker "computer familiarity" seem long gone in many contexts; in their place are concerns that test developers may not be able to keep up with the digital communicative practices of learners in this age range. As other authors in this volume note, however, technology in language testing also has its limitations. At the end of the chapter, Butler cautions that technology cannot replace teachers and emphasises the importance of teachers' "diagnostic competence" in young learner assessment. Young learner assessment looks set to remain a site of conceptual and practical advances in the coming years; we will watch these developments with interest.

Dynamic assessment has been a "hot" topic in language assessment for some time now. It offers a very radical departure from traditional testing practice, focusing on learning potential and the key role played by mediation. This places it at odds with more psychometric views which see language ability as residing solely within the mind of the individual learner. However, the dynamic assessment approach embodies an important critique of these more traditional testing practices. Scholarship in dynamic assessment forces us to consider that, if language ability develops through social interaction, then such interaction should play a primary role in our assessment practices. In Chapter 10, Antón and García update Antón's first-edition version, integrating research findings that have propelled dynamic assessment in new directions. Computerized dynamic assessment (CDA) features strongly; the authors describe fascinating new research in which new technology (such as messaging apps) is harnessed for its interaction and mediation potential. In the years to come, we expect to see the principles of dynamic assessment employed in many more assessment settings.

Diagnostic assessment is a feature of many professional contexts, most notably healthcare, but also IT support, car mechanics, engineering, and so on. Language testing has grappled with its own approach to diagnostic assessment for some time, but in recent years, there has been considerable progress on the topic through work on theory building and on technical issues (particularly in the area of cognitive diagnostic modeling – CDM). In Chapter 11 – an update of Jang's sole-authored contribution to the 2012 edition – current issues in diagnostic assessment are surveyed, and a range of novel directions are discussed. Jang and Sinclair provide an exciting glimpse into what diagnostic assessment might look like in the future, one in which machine learning could play an important role in capturing and processing data. Still, even with the most sophisticated methods for tracking and measuring learning, diagnostic assessment is not truly *diagnostic* without well-developed feedback systems. Here, Jang and Sinclair remind us that feedback utilization is a core component of effective diagnostic assessment and that evidence of use of feedback is essential for supporting validity arguments for diagnostic assessment. Our field needs more research on the effectiveness of specific, innovative diagnostic procedures. We hope that future researchers will read Jang and Sinclair's article and be inspired by the potential that diagnostic assessment offers.

Section 4 Assessing language skills

In the introduction to Chapter 12, Nakatsuhara, Khabbazbashi, and Inoue wisely quote Lado's observation that the ability to speak in a second language is the most prized objective of language learning. This is as true today as it has always been. Yet the assessment of speaking also remains as problematic as it was in the 1960s, despite all that we have learned since then, and is expertly documented in the historical section of this chapter. The authors consider the new range of task types available to us and the constructs of interest that each is claimed to reveal for scoring. As the scoring mechanism embodies the construct, it is not surprising that a variety of approaches have been proposed. The variety of competing claims about approaches also makes assessing speaking a controversial area of research and practice. In looking to the future, the authors state that technology has played a large role in shaping the speaking construct and will likely continue to do so. We also wonder whether this will be driven by commercial need, as it was during the COVID-19 pandemic, and how the machine scoring of speaking might model our humanity in the coming decades.

In another new addition to the *Handbook*, Chapter 13, Wagner provides a comprehensive overview of current challenges and debates in second language listening assessment. It is well understood among practitioners that listening is a complex skill to assess, both in terms of the practicalities of developing listening tests (sourcing or creating recordings, administering tests in contexts where resources are scarce) and in defining and operationalising the construct itself. But there is now a lot of research on listening assessment, and we were pleased to see Wagner "bust" the myth that listening is an under-researched skill. In fact, research on listening has grown over the past two decades, and in specific areas like the use of video in listening assessment and the role of speaker accent, there are now established canons of research to guide test developers in making good decisions about their own tests. The key problem is transforming that research into practice: designing innovative, authentic tasks that move beyond the traditional scripted, monologic, multiple-choice format. Wagner covers these issues, among many others, concluding, in a similar way to the previous chapter, that technology is likely to have a transformative approach on listening in the future.

Writing assessment has a long history, and one could be forgiven for thinking that we know a lot about what writing is, how it develops, and how to assess it. But as with many issues in language testing, the reality is more complicated. Writing is a constantly moving target for assessment; literacy practices evolve, and writing technologies change. In Chapter 14, Knoch carefully charts the different methods for conceptualising a writing construct and discusses a range of challenges related to scoring performances and interpreting the meaning of results. As with other chapters, technology looms large in new methods of automated scoring and feedback provision and in newer affordances raised by technology, such as the increasing prevalence of collaborative writing on platforms such as Google Docs. One of the most interesting contributions in this chapter, though, is conceptual. Knoch draws together research on written corrective feedback and writing assessment, arguing for greater merging of work in these two areas. As the field increasingly turns its attention to learning-oriented approaches to assessment, we might anticipate that feedback will become a central concern in writing assessment, even for large-scale international exams.

Based firmly in reading theory and the history of psycholinguistic research into reading processes, Brunfaut offers a masterly survey of approaches to the assessment of second language reading over the decades. Covering reading in both a first and a second language, she traces the evolution of reading assessment from the early days to the present. The discussion is illustrated with well-chosen references to some of the most influential reading tests currently in use. The summary of research into reading assessment is comprehensive and impressive and thoroughly supports the description of a range of research methodologies that practitioners can use in both creating and investigating the validity of reading tests. Brunfaut concludes Chapter 15 with directions for future research, which should provide an excellent starting point for anyone wishing to develop a reading assessment project that would add to our understanding of the field.

Section 5 Test design and administration

In Chapter 16 on test specifications (specs), Yan Jin provides a description of the role and purpose of test blueprint documents for both high- and low-stakes tests. The piercing analysis reveals the critical role of specs within the two paradigms: specifying purpose, content, and structure, and creating parallel forms. In the context of low-stakes assessment, their role in defining curriculum content and enhancing local understanding of learning goals is explained with clarity. The role of specs is supported with excellently chosen examples to illustrate their use across contexts. Yan Jin draws upon Fred Davidson's work in the previous edition of this volume as well as that of others but adds to our understanding of the richness of possibility in spec use. Pointing to the future, it is suggested that test specs may also be used to articulate intended consequences. What we like about this chapter is its inclusivity, the mastery over the history of the field, and the foresight offered through such a thorough understanding as well as practical knowledge.

Evidence-centered design (ECD) has increasingly been used in both test design, and the evaluation of test use and retrofit, as the chapter by Yan Jin in this volume attests. In Chapter 17 Yin and Mislevy provide us with a clear definition of ECD and its design components, along with an explanation of how these relate to argument-based approaches and a number of other validity models. It is arguably the flexibility of ECD that has led to its widespread use by test producers in North America, and its modular approach allows reuse of design elements across tests. While ECD can sometimes seem complex, coming to

terms with the design components can provide test designers with a conceptual framework for their activity.

Chapter 18 deals with accommodations to tests and assessments and is an updated version of Abedi's chapter in the first edition of this book. Accommodations are changes or alterations to a test in order to compensate for some disadvantage that may lower test scores as a result of a construct-irrelevant disability. These accommodations may compensate for problems such as a hearing or sight impediment or, in the case of language learners taking content tests (e.g., mathematics or history), provide them with access to the material without language ability impacting the test score. It is arguably the case that this is where validity theory meets practice head on. The reason is that any accommodation should reduce the impact of construct-irrelevant factors (the disability or taking the test in a second language) so that the score reflects ability on the construct; yet, if the accommodation alters the definition of the construct, as would be the case if a text in a reading comprehension test were read aloud to a blind test taker, the score meaning is fundamentally altered. Secondly, if the accommodation would raise the scores of test takers who would not normally be eligible for an accommodation, questions may be raised regarding the "fairness" of the practice. This may be the case with allowing additional time to dyslexic candidates, for example, if the same accommodation had a similar impact on scores of non-dyslexic candidates. Abedi also refers to the concept of universal design, which is becoming increasingly important to create accessible materials and avoid litigation on the grounds of discrimination.

Rater variability is a perennial topic in language testing research. As long as there are human raters, there will be interest in rater cognition, rater behaviour, and rater bias. The results of such studies are often fascinating, but they also lead to one inevitable conclusion: the need for rater training. Variability is seen as a threat to fairness. But how do we train effectively? Is training worth the effort? What does being an expert rater entail? And is variability actually all that bad? Davis provides a comprehensive overview of what we currently know about rater training (as well as interlocutor training, relevant to those tests with examiner-interlocutors), pointing to evidence-based findings and best-practice approaches. One of the most intriguing parts of Chapter 19 is the acknowledgement that strong uniformity is not always desirable. To some extent, training raters to be interchangeable with other raters only paves the way for automated scoring systems, which do uniformity much better and at a far lower cost. At the same time, we would not want to see a return to a system such as that practiced in the original Cambridge Proficiency in English exam, in which – as described by Davis – reputable individuals made judgments of the acceptability of language according to personal taste. Nevertheless, we believe there is always room for a human approach to language assessment, and what is more human than variability? The best kind of training, then, might not be the one that brings all raters into line, but the one which allows raters to form a community: to share their interpretations of rubrics and to increase their understanding of the nature of the construct.

Section 6 Writing items and tasks

There is a very serious scarcity of research and scholarship in our field about one of the most fundamental activities of language testing: item writing. While there are signs that this is changing, particularly as graduate student projects gravitate toward this obvious gap in the literature, there remain some difficult challenges in conducting research on item writing practices: security, ethics, and the general reluctance of item writers to share their "real" experiences on the job for fear of saying the wrong thing. In Chapter 20, Shin addresses this

point but takes the issue much further. Shin effectively opens up a range of possibilities for further research into item writing and item writers, pointing both to the micro-level issues at the heart of the item writing process and also the broader socio-political issues which guide the creation or selection of items for test construction. One of the most vivid passages in this chapter posits that item writers can be both oppressed (through the precarity of their working practices) and oppressors (possessing considerable power to select and write items which are then presented to test takers). In passages such as this, Shin's chapter lifts the study of item writing practices beyond the more mundane questions of test assembly and creates a new focus for language testing research: the item writer as a complex social actor and item writing as an important site of critical inquiry.

For a long time, it has been argued that an item or task should test a particular ability or skill in isolation so that scores are not contaminated. By "contamination," the critics meant (for example) that an ability to read should not cause a score on a speaking or writing test to vary. Recently, however, there has been a renewed interest in integrated items. The rationale is that such items more precisely reflect the kinds of holistic tasks that are undertaken in the "real world." Students do not simply write essays from prompts. They read texts, listen to lectures, discuss ideas, and then come to the writing task. In this revised Chapter 21, Plakans addresses the issues of complex constructs, how they are operationalized in integrated tasks, and the problems surrounding score interpretation. She provides new illustrations of integrated task types and updates us on the important research that has been carried out in recent years. The evolution and increased adoption of these task types will make integrated assessment a prolific area of research for many years to come.

In Chapter 22, Cohen updates his cutting-edge discussion of test-taker strategies. Central to the discussion is how, and to what extent, strategy use is part of the construct being assessed or something that fundamentally undermines our ability to make meaningful inferences about ability from test responses and scores. In the descriptive and historical parts of the chapter, Cohen adds to what we have learned about test-taking strategies, including the *social* element of their nature. He adds extensively to the list of new studies on test-taking strategies, providing an overview of advances in the field. The major new contribution to the debate comes with the new concept *test-deviousness strategies* and the role of these strategies in subverting valid score interpretation. This is a more accurate term than *test-wiseness* for strategies designed to result in higher scores without a corresponding increase in the construct of interest. In debunking much of what has proved less than useful in strategy research and setting us on a firmer path, Cohen has produced a chapter that is essential reading for anyone planning to undertake research in this area.

Section 7 Prototyping and field tests

Language test development has many commonalities with other areas of human endeavour that involve design. There is first an idea or thinking stage, followed by a stage in which those ideas are made more concrete through plans, specifications, and models. Developing a prototype is a fundamental step – often, this is the first point at which an idea is made tangible: an important means of communicating the design to a wider audience. In language testing, prototyping can often be a challenging step in the design process – full of trial and error – but it is also the point at which language test development can be exciting, as a new task comes to fruition, and theoretical potential is realised. In Chapter 23, Nissan and Park provide a clear overview of how these processes work at Educational Testing Service, giving some real examples of prototyping and talking through the various methods involved in

determining how to revise a task and whether to take it through to the next stage. As with other chapters in this collection, Nissan and Park point toward the increasing importance of digital technology in prototyping. This chapter will be of great interest to anyone involved in test design or anyone who wants to understand the role of design in tests.

Chapter 24 by Kremmel, Eberharter, and Holzknecht replaces and updates Chapter 20 in the first edition of the *Handbook* by Kenyon and MacGregor. They have ensured that the processes and practices so successfully described in the first chapter are still present, but they have added to the discussion with their own test development experience in Europe. This chapter, like the one before it, is essential reading for test developers. The simple reason, as the authors explain, is that there is a dearth of literature on pre-operational processes that must be carried out before a test is used to make important decisions. All too often, tests and assessments are written quickly and go straight into operational use without all the quality controls that this chapter so adeptly describes. Together with the contribution of Nissan and Park, we are provided with comprehensive test development guidance.

Chapter 25 – written by John Read – retains the dual focus that it had in the first edition of the *Handbook*. On the one hand, it is about the practice and process of piloting tests and so complements the new chapter by Kremmel *et al*., and the updated chapter by Nissan and Park. However, it does this from the perspective of vocabulary testing. Vocabulary testing continues to evolve as a very distinct discipline within language testing, and the more widespread use of corpora (see Cushing, this volume) has only added to the number of vocabulary studies being conducted. Vocabulary tests remain widely used as placement instruments and for research into how the brain stores lexical items and meaning. It is therefore crucial that these tests are piloted with care, and as John Read observes, the literature really does not tell us how piloting is carried out. With reference to the word associates format, he guides the reader through the process of piloting a vocabulary test and, in the process, provides us with a generic template for piloting other tests as well. The new chapter is updated with current literature, making it an invaluable resource for all vocabulary test researchers.

Section 8 Measurement theory in language testing

Perhaps the oldest psychometric toolkit available to language testing is CTT, or classical test theory. In Chapter 26, J. D. Brown reminds us that these tools are not obsolete. CTT is in widespread and perhaps pervasive use because it provides test developers with a means to carefully monitor the contribution of each test item to the overall test score distribution. The assumptions underlying modern test theory are essentially the same as those in CTT, and it is arguably the case that CTT offers insights that we have lost in some of the statistical determinism that can accompany more recent innovations in statistical analysis. This treatment of the "classical" toolbox is not only highly relevant to current practice; it is also essential for all language testers to gain an understanding of how we use statistical analysis in language testing research and design.

Item response theory (IRT) and many-facet Rasch measurement (MFRM) are now commonplace methods in language testing research. Aspiring testers are encouraged to learn these techniques in order both to be able to use these statistics in their research and development work and to develop a critical statistical literacy for reading and evaluating the research literature of the field. For those with little background in these methods, a daunting question is often "Where do I start?" Ockey's updated Chapter 27 in the second edition of the *Handbook* is a model of clarity. This is a perfect entry point for anyone interested in the

fundamental concepts in IRT and MFRM and provides an excellent discussion of issues for those already familiar with these approaches. Ockey includes coverage of recent literature and provides direction for interested readers to continue exploring the area. We encourage anyone with an interest in IRT and MFRM to dip into Ockey's chapter for an up-to-date overview.

In a similar way to Ockey's chapter, Chapter 28 on reliability by Yan and Fan renders a topic difficult to understand into a model of clarity. Their treatment of essential concepts in all measurement is completely accessible, and so when they go on to look at the five commonly used reliability coefficients, the reader is able to understand the type of measurement error that each one addresses with ease. But the chapter is not just a comprehensible overview of reliability. The authors offer new insights into the theoretical place of reliability within language testing and educational measurement with a nuanced treatment of its role and meaning in measurement paradigms. Far too often, reliability is treated simply as the application of psychometric technique. The authors' understanding of this topic is far too deep for such a fallacy to remain unchallenged. This chapter is set to become a classic in the field and should be read by researchers in educational measurement and behavioural science research more generally.

The Measurement section is rounded off by Chapter 29, in which Galaczi and Lim provide a thoughtful consideration of the very human factors that go into making judgments about performance. Notwithstanding current research into automated scoring (see the chapter by Xi, as well as the discussion here), it remains the case that communication takes place between complex biological beings, and deciding whether that communication is successful for a particular purpose is a task best suited to human inference. The authors set out some of the threats to sound inference and agreement between humans and describe the range of scoring options that have been devised to address concerns. But they argue, correctly, that there is no single solution and that there is no such thing as a scoring mechanism that is universally "good." Nevertheless, the research that has led us to where we are allows Galaczi and Lim to set out suggestions for practice that will be of benefit to all who are tasked with designing a scoring system for a performance test.

Section 9 Technology in language testing

Automated scoring systems are on the rise in language testing, extending from their use in high-stakes testing to more low-stakes classroom applications (e.g., automated writing evaluators [AWEs]). Their utility is obvious: they can bypass the need for human raters, allowing for the processing of large volumes of speaking and writing performances, at great speed and with a very quick turnaround for results. This is not just an economic argument; human raters bring with them biases that can impact scoring, and automated systems can maintain consistency in a way that is difficult for humans. Automated scoring, though, brings unique challenges with respect to validity. As Xi puts it in Chapter 30, "an automated scoring system is not just a case of replacing the human rater; rather, the system interacts with the other assessment components in complex ways." In her updated chapter, Xi outlines the validity issues associated with automated scoring, highlighting key questions that need to be addressed in supporting their use in operational testing. Xi also raises a fascinating discussion point: what sort of AI (artificial intelligence) literacy is required for stakeholders to engage with automated scoring systems, to understand their workings, and to critique their design? Relating to this point is the general issue of transparency. It remains to be seen how producers of automated scoring systems will grapple with balancing transparency and

proprietary knowledge. In the interests of stakeholders, we hope that the balance is toward the former.

Working through the chapters that were revised or newly written for this second edition, there was one that stood out as a common cross-reference: computer-based testing by Sawaki. Digital technology and computer-based testing are major themes across the whole collection. This is not necessarily related to the influence of the pandemic; most first drafts of chapters were produced prior to its onset, demonstrating that technology was already a major concern for language testers, weaving its way into a range of topics. However, the pandemic has increased the relevance of the focus on computer-based testing and amplified many of the issues that were already in train. Against this backdrop, Sawaki was given the unenviable task of encapsulating the current state of play in just one chapter. We think she has done a truly admirable job. Chapter 31 moves through the historical trajectory of computer-based testing before tackling some current issues. There is coverage here of more conventional comparative research (computer-based versus paper-based), as well as the novel and innovative uses of technology through virtual reality and spoken dialog systems. Similar to other authors, Sawaki sees potential for technology to connect assessment with learning more effectively. The chapter ends with a call for more collaboration with others outside the field to best harness the potential of technology. We agree with this sentiment and will look forward to future interdisciplinary work that explores computer-based testing from all angles: educational, technical, social, and ethical.

In deciding on new chapters for the second edition of the *Handbook*, one non-negotiable inclusion was a chapter on corpus linguistics and language testing. In the past 10 to 15 years, corpus approaches have become so commonplace in the practices of language testers that, as Cushing says in this chapter, "it would be almost unthinkable now to develop a large-scale language test without referring to corpus data." There are various ways in which corpora can be used in language testing. In Chapter 32, Cushing explains very clearly the various common methods of corpus linguistics and demonstrates the utility of corpus approaches for language testers. Some of the key applications include describing the domain of language use, determining features of performance across levels, and providing evidence to support extrapolation inferences. Corpora are not without their limitations, though, and language testers need to become not only corpus conversant, but also corpus critical. Cushing further points out that corpus techniques are closely interconnected with automated scoring systems. For that reason, language testers are encouraged to work together with computational linguists to create better corpora to inform scoring models.

Section 10 Ethics, Fairness, and Policy

The updated Chapter 33 by Walters incorporates fairness and ethics, which were separate chapters in the first edition of the *Handbook*. Building on the previous chapter by Davies, Walters incorporates a discussion and definition of the various ethical positions that language testers have adopted. The discussion of fairness then sits within the framework established. The chapter also treats a currently unresolved philosophical question: how is fairness different from validity? To close the chapter, Walters still leaves the reader with several carefully constructed exercises that help them relate the concepts and frameworks to their own assessment contexts.

Standards are endemic in all walks of life. Language education and testing have attempted to emulate what happens in industry. While analogy is a tenuous basis for a rationale, the fact is that creating and promoting language standards has become an industry. In this

masterly review and analysis, Deygers classifies standards into three types: educational performance indicators, proficiency frameworks, and institutionalized language tests. Chapter 34 describes each in turn and presents us with use, research, and critique. This is an area that cuts across policy, theory, research, and practice. So it is not surprising that it is highly controversial. Deygers is a sure guide to potential pitfalls and opportunities.

Language tests are developed across a variety of contexts. Many well-known tests are produced by large commercial or nonprofit organisations, with teams of staff working on assessment development, rating, research, and marketing. Other tests are developed by small teams of teachers with little support from their institutions. Each case presents different challenges for maintaining quality. In the case of smaller teams, it may be a lack of resources (both material and human) with few opportunities for conducting post hoc analysis of results. For larger providers, it is the complexity of the organisation itself: keeping track of who is doing what, ensuring communications are clear, avoiding "silo thinking," and so on. Saville and McElwee provide an updated version of an important chapter focusing on quality management in language testing. Combining the approaches of management science with language testing's typical focus on validity, Chapter 35 provides a useful template for test developers of all sizes to consider their own quality control procedures.

In the final chapter, we provide an epilogue to the volume as a whole. The first edition of the *Handbook* did not contain an epilogue; Fulcher and Davidson were satisfied to let the chapters have the final say. We maintain this stance, and our epilogue is not intended to critique or question the arguments put forth in the collection. Rather, we wanted to use the epilogue to reflect on themes and issues that we noticed emerging across the collection of chapters from our vantage point as editors. Or, to use a musical analogy, we wanted to riff on the multi-layered tracks laid down by our collaborators. From this perspective, we identify four key themes: (1) the increasing role of technology, (2) connecting assessment with learning, (3) grappling with complexity, and (4) theorising the socio-political nature of language testing. And having identified and discussed these themes, we offer some predictions for the direction of travel in the field as a whole. We end the epilogue by reiterating a call made by Fulcher and Davidson (2007) for a shift toward effect-driven testing. Understanding the scope and nature of the field, however, is the necessary precursor for predicting such effects. A volume of this kind tells us where we have been and where we need to go.

References

Aurelius, M. (2013). *Marcus Aurelius: Meditations, Books 1–6*. Oxford: Oxford University Press.

Bachman, L. (2006). Generalizability: A journey into the nature of empirical research in applied linguistics. In M. Chalhoub-Deville (ed.), *Inference and Generalizability in Applied Linguistics: Multiple Perspectives*. Amsterdam: John Benjamins, 165–207.

Fulcher, G. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. London and New York: Routledge.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.

Mill, J. S. (1859). On Liberty. In J. Gray (ed.), *John Stuart Mill On Liberty and Other Essays*. Oxford: Oxford University Press.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Section 1
Validity


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1
Conceptions of validity

Carol A. Chapelle and Hye-won Lee

Introduction

For many test users, validity is seen as an essential quality of a language test because to them "a valid test" means "a good test." Accordingly, test users frequently explain their search for a new test or their choice of a test by citing their desire for a test that "has been validated." In using the passive voice "has been validated," users reveal their assumption that validity is a quality of a test that is bestowed by testing experts to make a test measure what it is supposed to. This conception of validity, which requires test users to take little or no responsibility for validity, is at odds with the majority opinion of specialists in educational measurement. In fact, to claim that validity refers simply to demonstrating that a "test measures what it purports to measure" or that it is an inherent property of a test is to ignore at least seventy years of research on validity theory and test validation as well as the consensus Technical Recommendations and Standards that have existed since 1954 (Sireci, 2009: 28).

If validity is not that "a test measures what it purports to measure," what is it? Validity is defined by language testers in more than one way today, but in seeking an authoritative answer to the question, most would agree that the standards Sireci referred to should be taken into account. Sireci referred to the *Standards for Educational and Psychological Measurement*, published jointly by the American Psychological Association (APA), the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). In the *Standards*, validity is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of test scores" (AERA, APA, and NCME, 2014: 1). If validity entails demonstrating the basis for the interpretation of test scores and justifying their use, the question is, How does one go about doing this? In other words, what are the codes of practice for doing validation work? Specifically, what counts as credible evidence and compelling theoretical support for the validity of interpretations and uses of language tests?

In language testing, practices and rationales are outlined in books and journal articles, which draw upon sources in educational measurement, including the *Standards*, multiple editions of a volume entitled *Educational Measurement*, and other books and journal

DOI: 10.4324/9781003220756-3

articles on educational and psychological testing. Language testers apply the general concepts and approaches to the specific challenges of language assessment. In doing so, they recontextualize knowledge and practices to establish the field of language assessment. This chapter outlines the conceptions of validity that have appeared in language testing over the past sixty years. It outlines critical issues that remain in need of exploration and notes areas providing impetus for development. Finally, current contributions in validation are illustrated through the expanded roles of validation work across many facets of language testing.

Historical perspectives

From the first books on modern language testing in the 1960s, most have had something to say about validity, although precisely how validity is conceived and presented varies. Reflecting changes in both educational testing and the needs of language testers, multiple conceptions of validity are evident over the past sixty years, but the proposals for different frameworks for presenting and investigating validity in language testing increased considerably in the middle of the first decade of the 2000s. Despite the diversity of frameworks, the overall narrative of research in this area is that it transitions from what seems to rest on a question about what the test measures to a more complex investigation of multiple dimensions of test score meaning and use.

Types of validities

Early language testing books. such as Valette's in the 1960s, stated that for a test to be valid, "It must measure what it is supposed to measure" (Valette, 1967: 30). This and other textbooks at least through the 1980s introduced various types of validity, including "content validity," "face validity," "construct validity," and "criterion-related validity," sometimes with two subtypes: "concurrent validity" and "predictive validity" (Harris, 1969; Heaton, 1975; Hughes, 1989; Finocchiaro and Sako, 1983; Oller, 1979). These types of validities were borrowed from educational measurement, in which Newton and Shaw (2014) have documented expressions for dozens of types of validities coined by researchers in a variety of disciplines. Language testers used these terms to refer to concerns about the quality of tests. For example, Lado (1961) outlined procedures for validation by combining his own findings from contrastive linguistics with the contemporary work in educational measurement (e.g., Anastasi's, 1954 *Psychological Testing*). Similarly, Oller (1979) presented his construct theory along with validation concepts grounded in statistical analysis as explicated in educational and psychological testing. The terms introduced for types of validity in these books remain in some textbooks and appear regularly in the test manuals of commercial tests. In addition, as late at 2005, some language testers were adding to the types of validities with terms such as "cognitive validity," "scoring validity," and "context validity" within a framework constructed of types of validities (Weir, 2005).

Types of evidence for validity

In 1990, Bachman's *Fundamental Considerations in Language Testing* introduced applied linguists to validity as it had been outlined in the third edition of *Educational Measurement*. In that volume, Messick (1989) defined validity as "an overall evaluative judgment of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick, 1989: 13). The focus

is the test score, or the results of the test, because they are used to make an interpretation about test takers' ability and, typically, a decision. "Test score" refers to the results of a testing or assessment process even though such results might be in the form of a descriptor (rather than a number), and they might refer to more than one single score, as in the case of diagnostic information. As Messick (1989: 18) explained, the concept of validity as unitary was not new. Instead, his paper explicated the shortcomings of "types of validities" and justified a definition of validity that had emerged in the field over the previous twenty years. He reminded readers that the 1985 *Standards* "states flatly that validity is a unitary concept 'referring to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores'" (p. 8). Table 1.1 summarizes four aspects of Messick's conception of validity that have been important for language testing. Each characteristic is stated in terms of both what validity refers to and what it does not refer to.

Table 1.1 depicts a more complex conception of validity and process of validation than what had appeared in the early language testing books. The *Standards* in 1999 and 2014 reflected Messick's perspective and also made recommendations for validation practice as he did by indicating the types of evidence that can be used by researchers to support the interpretations and uses of test scores. These are summarized in Table 1.2, each with a note about what findings would be expected to reveal about the validity of a test's interpretation for a particular use.

Messick's conception of validity has been influential in language assessment. In a review of validity perspectives appearing in research articles in the two main language testing journals, Chapelle and Voss (2013) found Messick's paper is regularly cited as background for the validation research. The types of evidence Messick identified have been used to introduce collections of language testing research (e.g., Chapelle and Douglas, 1993; Cumming, 1996; Kunnan, 1998). Messick's emphasis on construct validity as central prompted Bachman's (1990) substantial discussion of the construct of communicative language ability, which has been used widely in language assessment. Messick's inclusion of the relevance and utility, value implications, and social consequences of testing is also credited with what Davies and Elder (2005) retrospectively called the social turn in the conception of validity in language assessment. The social view in language assessment (McNamara and Roever, 2006) welcomes critical inquiry as part of the validation process in order to be able to analyze the values and social consequences underlying test interpretation and use. Critical language testing (Shohamy, 2001) does not accept the status quo, but instead questions the assumptions about language underlying a test, the effects of those assumptions, and who benefits and loses because of these assumptions.

Table 1.1 Characteristics defining validity as presented by Messick (1989)

Validity refers to . . .	*Validity does NOT refer to . . .*
□ The interpretations and uses of tests.	□ A property of a test.
□ One unitary conception with construct validity as central.	□ Multiple validities such as "content validity," "criterion-related validity," or "face validity."
□ Construct validity, the relevance and utility, value implications, and social consequences of testing.	□ This scope for validity contrasts with the view that validity refers only to technical considerations.
□ A judgment based on an ongoing process of inquiry.	□ Product-oriented perspective of a validated test – one for which the research has been completed.

Table 1.2 Types of evidence pertaining to test score interpretations and actions from Messick (1989) and in the *Standards* (1999, 2014)

Evidence based on . . .	*Reveals . . .*
• Rationales and expert judgment about test content.	• Matches and mismatches between the test developers' intended task content and expert judges' evaluation of task content (Sireci and Faulkner-Bond, 2014).
• The study of test takers' response processes.	• The extent to which test taking processes engage test takers' knowledge, skills, and abilities as intended by the design of the test tasks and the stated construct interpretation (Ercikan and Pellegrino, 2017).
• Statistical testing of the internal structure of response data.	• Whether or not test takers' response patterns are consistent with the intended interpretation of the dimensionality of the construct.
• Relationships to other variables.	• The expected relationships between the construct measured by the test and measures of other variables (Cronbach and Meehl, 1955; Campbell and Fiske, 1959).
• The consequences of testing.	• The extent to which intended impacts of testing come to fruition for the test takers, test users, and others indirectly affected by the use of the test.

In short, many language testers found Messick's presentation of validity appropriately complex in view of the challenges inherent in language assessment. Nevertheless, if any assessment literacy was to be nurtured within the field, the complexity had to be operationalized as procedures that could actually be taught and used by language testers. Some textbooks in the 1990s responded by continuing to use "types of validity" from the earlier era. For example, Alderson *et al*. (1995) referred to types of validity but pointed out that "these 'types of validity' are in reality different methods of assessing validity" (Alderson *et al*., 1995: 171). Bachman and Palmer (1996), in contrast, introduced a novel approach by conceiving validity as "test usefulness."

Test usefulness

Bachman and Palmer (1996) introduced a framework for evaluating test usefulness to interpret validity theory in a manner that could be understood and used by those responsible for developing language tests and justifying their use. Framing the issue as an evaluation of usefulness rather than justifying interpretations and uses of test scores, Bachman and Palmer were able to communicate to a wide audience, including graduate students and practitioners, that tests have to be evaluated in view of the particular uses for which they are intended. The focus on usefulness helped communicate the definition of validity that appeared in the *Standards*, and the six criteria for evaluating test usefulness tailored the complexity of Messick's definition to language assessment. The six criteria are the following:

- Reliability: the consistency of test performance that is reflected in the test scores.
- Construct validity: the appropriateness of the inferences about test takers' ability made on the basis of test scores.
- Authenticity: the correspondence between the test tasks and the tasks students will engage in outside the test context.

- Interactiveness: the extent of the involvement of test takers' engagement in completing the test tasks using the same capacities they would use in other contexts.
- Impact: positive consequences the test might have on individuals and institutions within the immediate test setting and beyond.
- Practicality: the extent to which resources support the design and development of the test, as well as its ongoing use.

Bachman and Palmer's usefulness framework helped practitioners move beyond thinking of validity as a characteristic of a test to a more contextualized, use-focused evaluation. This framework is used in papers describing test evaluation projects, test reviews, and other books on language testing, including a widely disseminated book published by TESOL in the United States for teachers and others wishing to evaluate tests for adoption (Stoynoff and Chapelle, 2005). The usefulness framework also forecasted key concepts that are played out in the more praxis-oriented approach to validity that appeared in the next and most recent edition of *Educational Measurement*. Rather than qualities of test usefulness, an argument-based approach to validity is structured through the claims concerning test interpretation and use.

Argument-based validity

The editor of the fourth edition of *Educational Measurement* introduced the chapter on validation relative to Messick by indicating that the current chapter "aims at making validation a more accessible enterprise for educational measurement practitioners" (Brennan, 2006: 3). The emphasis on the detail of constructing arguments is the praxis step, which makes concrete the expression (i.e., validity argument) that had been used for many years (e.g., Cronbach, 1988). The praxis orientation to validity arguments is developed according to these principles as outlined by Kane (1992, 2001, 2006, 2013; Kane et al., 1999; see also Kane, this volume):

- A validity argument is structured through claims that one would like to make about test score meaning. Claims are justified by support for their underlying inferences.
- The first step in developing a validity argument is to create an interpretation/use argument stating the claims to be made about the test scores and their intended use.
- The interpretation/use argument guides the validation research, which investigates the extent to which inferences are supported. The supported interpretation/use argument is the validity argument.
- The validity argument specifies multiple facets of score meaning, which can include technical aspects, such as level of response consistency, as well as social aspects, such as the value and policy implications of score use.
- The conclusion for the validity argument is a claim (or claims) about the utility of the test score for a specific purpose and the intended ensuing consequences.

Kane's (2006, 2013) conception of validation calls for a reflective process requiring the test developer to have an understanding of what a test score can and should mean, as well as the knowledge to draw upon theoretical rationales and empirical methods to support score meaning and use. Kane provides examples of how interpretation/use arguments can be developed, as summarized in Table 1.3, but for any particular test, the researcher is responsible for outlining the interpretation/use argument.

Table 1.3 Types of claims in validity arguments and the inferences leading to each type[1]

General Claim[2] about Test Scores	*Inference Required to Accept the Claim in a Validity Argument and its Meaning*
• Test score use results in positive consequences.	• Consequence implication: the test user believes that the intended positive consequences will result from intended test use.
• The test scores are useful for their intended purpose.	• Utilization: the test user finds that the test scores serve well for their stated use.
• The test scores represent the performance of interest in the target context.	• Extrapolation: the test user accepts that the score meaning is appropriate in view of the target context of language use.
• The test scores reflect the intended construct.	• Explanation: the test user concludes that the score meaning is explained by the defined construct.
• Scores reflect performance consistency.	• Generalization: the test user accepts that the test has produced reliable scores in relevant research contexts.
• The test scores accurately summarize relevant performance.	• Evaluation: the test user trusts that relevant test takers' performance has been summarized accurately in scores
• Appropriate observations of test takers' performance has been obtained.	• Domain definition: the test user accepts that the analysis of the domain of interest has been carried out adequately to create relevant assessment tasks with appropriate content.

[1] The wording used for the inferences comes from Chapelle (2021).
[2] These general claims illustrate the types of statements that testers want to be able to make about tests, but in constructing a validity argument for a particular test, the tester would make specific claims about the intended interpretation and use of that test for the designated purpose.

Kane's perspective on the interpretation/use argument and validity argument has been influential in language assessment. Bachman (2005) presented validation by drawing on the same principles of argument that Kane did. The validation research for a major international language testing program, the Test of English as a Foreign Language (TOEFL), was framed on the basis of an interpretive argument as outlined by Kane (Chapelle et al., 2008), in conjunction with evidence-centered design (see Mislevy, this volume). Norris (2008) described validation work on a language testing program using the principles of Kane's argument-based approach. The language assessment textbook *Language Testing and Assessment: An Advanced Resource Book* (Fulcher and Davidson, 2007) presents validity argumentation based on this approach, and Bachman and Palmer's (2010) book *Language Assessment in Practice* presents an approach to validity argument specifically for language assessment. The use of argument-based validity in language assessment for over a decade has offered new perspectives on some perennial issues in language assessment and has raised some new issues.

Critical issues

The evolution of the concept of validity could be summarized as a relatively orderly narrative in which new ideas supersede their predecessors, and the research community is uniformly on board with changes. However, a more accurate presentation would acknowledge that new concepts do not supersede old; instead, new concepts join the old, creating a

marketplace of terms and concepts that teachers, researchers, test users, and testing companies use to talk about validity.

Co-existing perspectives on validity

The co-existing perspectives on validity among language testing researchers present a critical issue for the field today. All four of the frames of reference introduced chronologically earlier in this chapter appear in textbooks, journal articles, research reports, and technical test documentation – sometimes explicitly presented and sometimes only alluded to through the use of a technical term or unstated assumption. The field does not appear to share a concern about making explicit the validity framework guiding research in each study, perhaps due to a belief that there is only one way to do validation research. In a review of validation perspectives appearing in articles in the two primary language assessment journals, Chapelle and Voss (2013) found that the majority of published reports of validation research did not present an explicit statement of a validation approach or framework.

Raising consciousness about the difference between investigating the validity of a test and the validity of a test interpretation and use, for example, requires a more philosophical analysis of the meaning of research and interpretation of results than many language testers are prepared for. In an applied field such as language assessment, many testers spend their time on and direct their attention to the technical work of validation research. Such testers are not necessarily inclined to investigate the philosophical underpinnings that provide a basis for making decisions about the meaning of validity. Books have been written about the philosophical underpinnings of validity (Lissitz, 2009; Markus and Borsboom, 2013). Similarly, in language assessment, some philosophical discussion appears in books that take up the concepts of inferences, constructs, and values (Chalhoub-Deville *et al*., 2006; Fulcher, 2015; Kunnan, 2018). However, considerable scope remains for advancing language assessment literacy within the field to lay the groundwork for fruitful academic discussion and use of validation concepts.

Audiences for validity arguments

Validity arguments have important roles to play in communication beyond the field of language assessment. An argument is ideally made to an audience, and therefore success is determined on the basis of how well the argument succeeds with the audience (Cherryholmes, 1988). From this vantage point, the audience and their knowledge are critical parts of the validity question: Who is the audience for a validity argument, and what knowledge and values do they hold that affect their judgment? Obviously, the non-specialist test users will view a validity argument from a different perspective than will specialists in language testing. Therefore, one important issue is how to express a validity argument in a manner that communicates the depth and aptness of the support for test interpretations and uses to a general audience. Users are typically happy to count on the experts for the specifics of the technical backing in the validity argument. However, at the same time, test users need to be able to see their role in judging the appropriateness of a test for their particular context and use. This issue has never been more important than it is in today's environment, where anyone can start a language testing business without encountering any legal challenges such as licensing requirements based on technical review.

If test users depend on specialists, who is the specialist community for language assessment? Who is the audience for whom the technical quality of validity arguments are

presented, and by whom they are judged? The chronology of validity conceptions in language assessment demonstrates that language assessment researchers draw upon work in the largely North American community of educational measurement. However, language assessment is an international community that needs to speak to a world characterized by hybrid influences of culture, academic traditions, and disciplinary affiliations. The ideal strategies for communication about validity within a diverse professional context remain a challenge. Central to the challenge are the practices for communicating about the construct interpretations of language test scores.

Language constructs

Test scores are used as indicators of test takers' ability in the language, but because language encompasses a range of complex knowledge, skills, and abilities, the specific meaning of language that should be understood from a test score needs to be specified. If a test is intended to measure communicative competence, the test construct typically has to be specified as the response consistency reflective of the many sources of knowledge, strategies, and abilities, as well as contextual factors responsible for communicative performance. Tests of language for specific purposes provide the clearest examples of the need for a perspective on language that links language and context (McNamara, 1996; Douglas, 2000), but many test scores are intended to indicate the test taker's ability to use language, and language use takes place in contexts with specific affordances and demands. Sociolinguist Hymes identified certain features of context that affect the language choices speakers or writers make. Systemic functional linguistic theory embroiders this basic concept with theoretical and analytic tools for describing the specific linguistic choices made in view of the meaning potential of the language and the repertoire of language resources that the language user controls (e.g., Halliday and Hasan, 1989). Language educators work with the analytic tools of systemic functional linguistics (e.g., Byrnes et al., 2010; Purser et al., 2020) to take into account the context of language use in performance and development.

The challenge for language testers is to recontextualize such a perspective on language use in context into a construct definition that serves for score interpretation. Basic approaches to construct definition are sketched out in educational measurement by Messick (1981, 1989), who identified three ways of defining constructs: as samples of behavior in particular contexts; as signs of underlying traits; and as a combination of the two: i.e., as signs of traits as they are sampled in particular contexts. It is the latter, the interactionalist construct definition, that has been particularly relevant for interpretations of language ability in context that applied linguistics are interested in (Chapelle, 1998). Bachman's (2007) analysis of the three basic designs makes finer-grained distinctions, and Mislevy's (2009) analysis of construct definitions as models for interpreting response consistency provide additional input to the issue of how best to define language construct. All these concepts provide useful background, but argument-based validity offers some more precise tools and concepts for defining language constructs through the warrants and assumptions underlying the explanation and the extrapolation inferences. Warrants are statements about what needs to be supported by empirical evidence or theoretical rationales, and assumptions are statements about the specific theory and evidence that provide the support. These requirements may be useful in helping make the connection between validation research and theoretically grounded statements about the intended construct interpretation.

Language development

In some language testing contexts, for instance, where assessment operates on multiple levels and formal tests take one form of many to monitor progress in the ecosystem of learning (Jones and Saville, 2016), the test taker's score is intended to index a particular level of language development, and therefore the construct that should form the basis of the construct validity argument needs to be interpreted from a developmental perspective. Such score meanings are implied when scores are indexed to the such frameworks as those of the American Council on the Teaching of Foreign Languages (ACTFL) or the Common European Framework of Reference (CEFR). In other words, score meaning is not the amount of a particular construct or an indicator of the level of an ability. It refers to the position on a trajectory of language development. In such cases, any given score is to be interpreted relative to the scores below and above it on the scale in such a way that one would expect learners to move along the scale as their language developed. Ideally, such scale descriptors would make sense relative to perspectives and findings from research on second language development. In fact, however, it is difficult to find explicit links in descriptions of such scales and frameworks.

One reason for the gap that has sometimes been cited is that second language acquisition research of the past was primarily focused on development of morphosyntax, whereas descriptions of levels used in educational settings refer to broad functional abilities in contexts. This reason, however, has become less credible over the years as second language acquisition theory and research has expanded to encompass multiple perspectives. The issue today is not that second language research findings are too narrow to be useful but rather that they are so rich, diverse, and complex that it is not clear precisely how the connections can be made. This challenge is exacerbated by the lack of shared knowledge base between researchers in language assessment and those in second language acquisition. Yet, at some point, it is likely to become more difficult to maintain developmental score meanings that are completely unrelated to any empirical and theoretical bases in the study of second language acquisition.

Values and consequences of language testing

Messick (1989) challenged researchers to justify the values underlying test interpretation and use as well as to include the consequences of test use within the scope of validation research: "Given the pervasiveness and the subtlety of the impact of values and ideologies on test interpretation, we need to explore some ways and means of uncovering tacit value premises and of coping with their consequences for test validation" (Messick, 1989: 62). The issue, which is not immediately obvious to many researchers, is frequently exemplified with a discussion of the cultural biases inherent in the intelligence testing of the 1950s. The construct of intelligence was operationalized in a manner that favored examinees who grew up in affluent circumstances, and test scores were used to place those with high scores in advantageous positions (Zwick, 2006). Rectifying the situation requires researchers to examine such issues as the values underlying the construct definition and the beneficiaries of test use, as critical language testing does.

In language assessment, value-laden issues emanate from such practices as the use of the native speaker as a norm for assessment, selection of privileged varieties of a language as a standard, and interpretation of test scores for learners with different learning biographies

(e.g., heritage learners vs. classroom learners). In all these cases, the manner in which language constructs are defined and operationalized in tests holds the potential to systematically advantage particular test takers on the basis of their backgrounds, rather than on relevant aspects of their language ability. Language constructs can be defined in many different ways, and so the point of the values analysis is to examine the value implications of choosing one construct definition over another. Like defining intelligence, the problem is insidious because the constructs are defined and the test scores are used by people who themselves operate on the basis of their implicit values. Moreover, McNamara (2006) pointed out that the more praxis-oriented approaches to validation risk de-emphasizing analysis of values at precisely the time when such analyses are increasingly important in language assessment. The challenge for validation remains making explicit the values at work in the processes of test design and use.

Current contributions

Recent books, journal articles, and research reports have explored the form and utility of argument-based validity for various purposes in language assessment. The current uses of argument-based validity demonstrate its utility for organizing the many types of evidence entailed in validation work from test development through evaluation of consequences arising from testing outcomes.

Test developers draft an interpretation/use argument to state the claims that they will want to be able to make about the test. For example, test developers are concerned that the test tasks they create will elicit language performance that is relevant to the contexts in which the test scores will be used, that the scores are reliable, and that the inferences about ability are justified. These and other important claims about the test scores can best be taken into account if they are specified early in the test development process (e.g., Voss, 2012; Youn, 2015). After a test has been developed and some validation research has been conducted, the argument is used to guide the interpretation of research results to manage the integration of evidence coming from various sources.

Once test developers have established a validity argument, it is used to clarify the status of validity evidence for each intended test use and identify new research that is needed, interpret results, and communicate to others about validity. For example, the validity argument for the TOEFL iBT (Chapelle et al., 2008) developed in the mid-2000s has continued to serve as a framework for new research that has been prompted by warrants in the original validity argument in need of additional investigation. Researchers have sought evidence for the evaluation inference when automated scoring of the writing section was added (e.g., Enright and Quinlan, 2010), for the extrapolation inference (Brooks and Swain, 2014 for the speaking section; Llosa and Malone, 2019; Klebanov *et al*., 2019 for the writing section) and for the utilization inference when section scores are used (Sawaki and Sinharay, 2018).

External evaluators consisting of prospective test users and testing researchers have also used argument-based validity to create the validity argument needed to provide structure to their evaluation. In this case, the evaluators work with existing validation reports that present research on, for example, the reliability of the scores for various populations of test takers and the standard setting process yielding recommended cutscores. Cheng and Sun (2015) acted as external evaluators when they created a validity argument based on technical reports for the Ontario Secondary School Literacy Test, a standardized test of reading and writing skills in content areas used in Ontario. They used the validity argument to summarize the existing evidence for the validity of score interpretation and use for

English language learners. Another example is Llosa's (2005) research, in which she acted as an external evaluator to create a validity argument for the English Language Development Classroom Assessment to clarify the intended claims being made about the test scores. Having stated the intended claims and warrants, she was able to identify potential rebuttals that would weaken the validity argument, if supported. Therefore, the rebuttals were used to develop the research questions to be investigated. The validity argument contextualized the rebuttals, showing why they posed threats to the validity of the test interpretation and use.

External evaluators can adopt the position of prospective test users or test reviewers reporting on the strengths and weakness of a test for a particular use. Validity arguments developed by an evaluator for existing tests have been used to compare two tests as Johnson (2011) did by summarizing the evidence and conclusions about two assessment types for placement into writing courses. Test reviewers who structure the review by using a validity argument first create the validity argument implicitly expressed by the test developer. To illustrate, Wang *et al*. (2012) constructed a validity argument for the Pearson Test of English Academic, consisting of "desirable" claims and warrants, and gathered evidence from documents provided by the test developer to support or raise questions about each of the warrants. In doing so, the reviewers identified some potential gaps in support for the validity argument, particularly in areas of test use and consequences, and suggested the test developer should sponsor studies to address the missing pieces.

Recent language assessment research and practice demonstrate that validity argument frameworks serve as rhetorical devices with a common set of concepts and language. In the domain of language assessment, where the foundational concept of validity has been difficult to grasp throughout its evolution, such a common language and logic appear to be welcomed by test developers, testing researchers, and test evaluators who need to explain their bases for validity claims without first inventing a means of doing so. Throughout the various ways of conceptualizing validity, the imperative remains for language testers to interpret multiple types of evidence in view of its meaning for validity. Perhaps because of the management capacity of argument-based validity, validity arguments have been useful.

Future directions

The impetus to improve the clarity and usability of conceptions of validity in language assessment comes from the unrelenting need for methods of evaluating and justifying testing practices that affect millions of test takers each year. In this context, the attraction of argument-based validity is the clarity it requires in statements about the evidence needed to support the aspects of the argument as well as the logic that must be demonstrated in the connections among the types of evidence. Over the past decade or more, argument-based frameworks have proven useful for discussion of validity issues in language assessment. It appears that this trend is likely to continue in the near future as recent work suggests the expansion of use of concepts from argument-based validity. In language testing, Bachman and Dambock's (2017) recent book presents argument-based validity concepts to language classroom teachers, and Kunnan's (2018) book develops concepts of fairness and justice in test use (see also Kunnan, this volume). In educational and psychological testing, Chapelle's (2021) book presents an argument-based approach to a broad audience of testing specialists. The book is intended not to simplify what is inherently a complex, multifaceted process, but to explain and standardize the terms and concepts in order to welcome more professionals into the discourse community of testers who are able to reason about test interpretation and use in this way.

Further reading

Bachman, L. F. and Palmer, A. S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

This book takes readers through the process of building an assessment use argument based on four claims. It is aimed at readers wishing to take validity argumentation into account from the beginning stages of test design without getting into the detail of a more complex validity argument. It emphasizes the use of validity arguments for justifying test design decisions to stakeholders.

Chapelle, C. A. (2021). *Argument-based Validation in Testing and Assessment*. Thousand Oaks, CA: Sage Publishing.

Chapelle defines the concepts in assessment required to develop and use validity arguments by situating them historically and introduces the language required to create and understand complex validity arguments. She illustrates how the basic concepts in assessment are made more precise and contextually relevant by developing them into claims, inferences, warrants, and assumptions in validity arguments.

Chapelle, C. A., Enright, M. E. and Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice* 29: 3–13.

This paper describes four points of distinction that can be made between the approach to validation described by Messick and the more praxis-oriented argumentation outlined by Kane. It draws upon experience in developing a validity argument for a high-stakes language test that attempted both approaches.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50: 1–73.

This article supersedes previous presentations of validity argument by Kane. It presents the historical context in which argument-based validity took shape, introduces terms, and presents the concepts of validity argument. This is an important reference for anyone working on a validity argument.

Kunnan, A. J. (2018). *Evaluating Language Assessments*. New York and London: Routledge.

The volume develops the issues entailed in the wider political, economic, social, legal, and ethical contexts in which validity arguments are intended to work. Through the lens of fairness of test use and justice, the book presents the types of qualitative and quantitative research that can be used in validation studies. The book demonstrates the importance of theory and practice in language assessment for the fair use of language tests and achievement of social justice.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: Macmillan Publishing Co, 13–103.

This is the seminal paper on validity, which consolidates and extends work on theory and practice in validation of educational and psychological assessments while contextualizing validation within philosophy of science. It defines validity in terms of interpretations and uses of tests while including relevance and utility, value implications, and consequences of testing. It contains extensive discussion and examples of validation research in addition to definition and discussion of concepts that remain fundamental conceptions of validity.

References

AERA, APA and NCME. (2014). *Standards for Educational and Psychological Testing*, 6th edn. Washington, DC: AERA.

Alderson, J. C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

Anastasi, A. (1954). *Psychological Testing*. New York, NY: Palgrave Macmillan.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

———. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2: 1–34. https://doi.org/10.1207/s15434311laq0201_1.

———. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche and D. Bayliss (eds.), *What Are We Measuring? Language Testing Reconsidered*. Ottawa, Canada: University of Ottawa Press, 41–71.

Bachman, L. F. and Dambock, B. (2017). *Language Assessment for Classroom Teachers*. Oxford, UK: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

———. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

Brennan, R. L. (ed.), (2006). Perspectives on the evolution and future of educational measurement. In R. Brennen (ed.), *Educational Measurement*, 4th edn. Westport, CT: Greenwood Publishing, 1–16.

Brooks, L. and Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBTTM and real-life academic speaking activities. *Language Assessment Quarterly* 11: 353–73. https://doi.org/10.1080/15434303.2014.947532.

Byrnes, H., Maxim, H. H. and Norris, M. J. (2010). Realizing advanced foreign language writing: Development in collegiate education: Curricular design, pedagogy, assessment. *Modern Language Journal* 94: 1–8.

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81–105. https://doi.org/10.1037/h0046016.

Chalhoub-Deville, M., Chapelle, C. A. and Duff, P. A. (2006). *Inference and Generalizability in Applied Linguistics: Multiple Perspectives*. Amsterdam, Netherlands: John Benjamins Publishing Company.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), *Interfaces between Second Language Acquisition and Language Testing Research*. Cambridge, UK: Cambridge University Press, 32–70.

———. (2021). *Argument-based Validation in Testing and Assessment*. Thousand Oaks, CA: Sage Publishing.

Chapelle, C. A. and Douglas, D. (1993). Foundations and directions for a new decade of language testing. In D. Douglas and C. Chapelle (eds.), *A New Decade of Language Testing Research*. Arlington, VA: TESOL Publications, 1–22.

Chapelle, C. A., Enright, M. E. and Jamieson, J. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. London, UK: Routledge.

Chapelle, C. A. and Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (ed.), *The Companion to Language Assessment*. Chichester, UK: Wiley, 1079–97.

Cheng, L. and Sun, Y. (2015). Interpreting the impact of the Ontario secondary school literacy test on second language students within an argument-based validation framework. *Language Assessment Quarterly* 12: 50–66. https://doi.org/10.1080/15434303.2014.981334.

Cherryholmes, C. (1988). *Power and Criticism: Poststructural Investigations in Education*. New York, NY: Teachers College Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer and H. Braun (eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum, 3–17.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52: 281–302. https://doi.org/10.1037/h0040957.

Cumming, A. (1996). Introduction: the concept of validation in language testing. In A. Cumming and R. Berwick (eds.), *Validation in Language Testing*. Clevedon, UK: Multilingual Matters, 1–14.

Davies, A. and Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah, NJ: Lawrence Erlbaum Associates, 795–813.

Douglas, D. (2000). *Assessing Language for Specific Purposes*. Cambridge, UK: Cambridge University Press.

Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing* 27: 317–34. https://doi.org/10.1177/0265532210363144.

Ercikan, K. W. and Pellegrino, J. W. (eds.). (2017). *Validation of Score Meaning in the Next Generation of Assessments. The Use of Response Processes*. New York: Routledge.

Finocchiaro, M. and Sako, S. (1983). *Foreign Language Testing: A Practical Approach*. New York, NY: Regents Publishing Company.

Fulcher, G. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. New York, NY: Routledge.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London, UK: Routledge.

Halliday, M. A. K. and Hasan, R. (1989). *Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective*. Oxford, UK: Oxford University Press.

Harris, D. P. (1969). *Testing English as a Second Language*. New York, NY: McGraw Hill.

Heaton, J. B. (1975). *Writing Language Tests*. London, UK: Longman.

Hughes, A. (1989). *Testing for Language Teachers*. Cambridge, UK: Cambridge University Press.

Johnson, R. C. (2011). *Assessing the Assessments: Using an Argument-Based Validity Framework to Assess the Validity and Use of an English Placement System in a Foreign Language Context*. Unpublished doctoral dissertation, Macquarie University, Sydney, Australia.

Jones, N. and Saville, N. (2016). *Learning Oriented Assessment: A Systemic Approach. Studies in Language Testing*, vol. 45. Cambridge: UCLES/Cambridge University Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin* 112: 527–35. https://doi.org/10.1037/0033-2909.112.3.527.

———. (2001). Current concerns in validity theory. *Journal of Educational Measurement* 38: 319–42. https://doi.org/10.1111/j.1745-3984.2001.tb01130.x.

———. (2006). Validation. In R. Brennen (ed.), *Educational Measurement*, 4th edn. Westport, CT: Greenwood Publishing, 17–64.

———. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement* 50: 1–73. https://doi.org/10.1111/jedm.12000.

Kane, M. T., Crooks, T. and Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice* 18: 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.

Klebanov, B. B., Ramineni, C., Kaufer, D., Yeoh, P. and Ishizaki, S. (2019). Advancing the validity argument for standardized writing tests using quantitative rhetorical analysis. *Language Testing* 36: 125–44. https://doi.org/10.1177/0265532217740752.

Kunnan, A. J. (1998). *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach*. Mahwah, NJ: Lawrence Erlbaum Associates.

———. (2018). *Evaluating Language Assessments*. New York and London: Routledge.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. New York, NY: McGraw-Hill.

Lissitz, R. W. (ed.). (2009). *The Concept of Validity: Revisions New Directions and Applications*. Charlotte, NC: Information Age Publishing, Inc.

Llosa, L. (2005). *Building and Supporting a Validity Argument for a Standards-Based Classroom Assessment of English Proficiency*. Unpublished doctoral dissertation, University of California, Los Angeles.

Llosa, L. and Malone, M. E. (2019). Comparability of students' writing performance on TOEFL iBT and in required university writing courses. *Language Testing* 36: 235–63. https://doi.org/10.1177/0265532218763456.

Markus, K. A. and Borsboom, D. (2013). *Frontiers of Test Validity Theory: Measurement, Causation, and Meaning*. New York, NY: Routledge.

McNamara, T. (1996). *Measuring Second Language Performance*. London, UK: Longman.

———. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly* 3: 31–51. https://doi.org/10.1207/s15434311laq0301_3.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA: Blackwell Publishing.

Messick, S. A. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin* 89: 575–88. https://doi.org/10.1037/0033-2909.89.3.575.

———. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: Macmillan Publishing Co, 13–103.

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. W. Lissitz (ed.), *The Concept of Validity: Revisions New Directions and Applications*. Charlotte, NC: Information Age Publishing, Inc, 83–108.

Newton, P. E. and Shaw, S. D. (2014). *Validity in Educational & Psychological Assessment*. London, UK: Sage Publications.

Norris, J. (2008). *Validity Evaluation in Foreign Language Assessment*. New York, NY: Peter Lang.

Oller, J. (1979). *Language Tests at School*. London, UK: Longman.

Purser, E., Dreyfus, S. and Jones, P. (2020). Big ideas & sharp focus: Researching and developing students' academic writing across the disciplines. *Journal of English for Academic Purposes* 43: 100807. https://doi.org/10.1016/j.jeap.2019.100807.

Sawaki, Y. and Sinharay, S. (2018). Do the TOEFL iBT® section scores provide value-added information to stakeholders? *Language Testing* 35: 529–56. https://doi.org/10.1177/0265532217716731.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language if Language Tests*. Harlow, UK: Pearson Education.

Sireci, S. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (ed.), *The Concept of Validity: Revisions New Directions and Applications*. Charlotte, NC: Information Age Publishing, Inc, 19–38.

Sireci, S. and Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema* 26: 100–7. https://doi.org/10.7334/psicothema2013.256.

Stoynoff, S. and Chapelle, C. A. (2005). *ESOL Tests and Testing: A Resource for Teachers and Program Administrators*. Alexandria, VA: TESOL Publications.

Valette, R. A. (1967). *Modern Language Testing*. New York, NY: Harcourt, Brace and World.

Voss, E. (2012). *A Validity Argument for Score Meaning of a Computer-Based ESL Academic Collocational Ability Test Based on a Corpus-Driven Approach to Test Design*. Unpublished doctoral dissertation, Iowa State University, Ames, Iowa, USA.

Wang, H., Choi, I., Schmidgall, J. and Bachman, L. F. (2012). Review of Pearson Test of English Academic. *Language Testing* 29: 603–19. https://doi.org/10.1177/0265532212448619.

Weir, C. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Hampshire, UK: Palgrave Macmillan.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing* 32: 199–225. https://doi.org/10.1177/0265532214557113.

Zwick, R. (2006). Higher education admissions testing. In *Educational Measurement*, 4th edn. Westport, CT: Greenwood Publishing, 221–56.

2

Articulating a validity argument

Michael T. Kane

As the most recent edition of the Standards for Educational and Psychological Testing define it:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests.
>
> *(AERA, APA, and NCME, 2014: 11)*

That is, validation involves an evaluation of the appropriateness of proposed uses of test scores and of the plausibility of their supporting interpretations, and therefore, it is critical that the proposed interpretations and uses be clearly stated: "Ultimately, the need for validation derives from the scientific and social requirement that public claims and decisions be justified" (Kane, 2006). Without a clear statement of the claims being made, any validation will necessarily be incomplete and fragmentary.

Framework

An argument-based approach to validation imposes two basic requirements: first, state the claims that are being made, and second, evaluate the credibility of these claims. One way to specify the proposed interpretation and uses in some detail is to lay out a chain or network of inferences and supporting assumptions that would get us from the test scores to the proposed interpretations and uses of the scores. Given an explicit statement of the proposed interpretation and use, along with strong evidence supporting the underlying assumptions, it would be reasonable to accept the score-based inferences and the appropriateness of the uses. Kane (2013) refers to the network of inferences and assumptions that specify the proposed interpretation and use as an *interpretation/use argument (IUA)*. Bachman and Palmer (2010) proposed a similar approach with more emphasis on the uses of scores.

Once the IUA is in place, it can be evaluated in a *validity argument*, which examines its coherence and the plausibility of its inferences and assumptions. The validity argument is

 DOI: 10.4324/9781003220756-4

to evaluate the plausibility of the IUA and also to provide an evaluation of alternative possible interpretations, given the intended uses of the scores.

The interpretation/use argument (IUA)

The IUA is to specify the proposed uses of the test scores and their supporting interpretations and, in doing so, to lay out the inferences and assumptions inherent in these uses and interpretations. Inferences generally take the form of "if-then" rules. For example, a scoring rule might indicate that if a performance has met certain requirements (e.g., it responds fully and accurately to the question), it should get a certain grade. A less complete or accurate response would get a lower grade. The scoring rubrics award higher scores to performances that are better in some appropriate sense than performances that get lower scores.

The scores on a sample of such performances might be averaged to get an estimate of the student's expected level of performance over a larger domain of similar tasks. This inference from the score summarizing a sample of performances to the mean over the domain of possible performances rests on statistical theory and assumptions about sampling. It will generally rely on evidence that the sample is large enough and representative enough to warrant the generalization.

An inference from a student's score to a claim that the student has achieved some level of proficiency in some performance domain is based on a rule that if a student's score is above some cutscore (associated with the "proficient" category), the student will be classified as proficient. In this case, another level of interpretation involving categorical descriptions (e.g., proficient, advanced) is superimposed on the test scores.

Each inference involves an extension of the interpretation or a decision. For the inferences to be considered reasonable, they have to be either highly plausible *a priori* or supported by appropriate evidence; that is, the inferences should be warranted in some way.

The inferences rely on a large number of assumptions, but many of these assumptions can ordinarily be taken for granted; for example, on standardized college admissions tests, we typically assume that students understand the instructions and the questions and can record their responses, unless there is some reason (e.g., a physical disability, testing conditions) to think otherwise. For assumptions that are not plausible *a priori*, the validity argument needs to provide evidence (e.g., from empirical studies). It is particularly important to identify the inferences in the IUA that are most questionable and to subject these assumptions to extra scrutiny.

The IUA is intended to make the reasoning inherent in the proposed interpretations and uses explicit so that they can be evaluated by the validity argument. The IUA thereby serves three important functions. First, it provides a framework for test development by indicating the assumptions that need to be met. Second, it provides a framework for the validation of the proposed interpretation and use. Third, it provides a basis for evaluating the adequacy of the validity argument; if the validity argument provides adequate support for the coherence of the IUA, and for the plausibility of its inferences and assumptions, the proposed interpretations and uses can be considered valid.

The validity argument

The *validity argument* provides an overall evaluation of the proposed interpretations and uses, as outlined by the IUA. To claim that the proposed interpretations and uses are valid

is to claim that the IUA is complete and coherent, that its inferences are reasonable given the assumptions that are made, and that the assumptions are either inherently plausible or are supported by adequate evidence for the population of interest. The validity argument "should make clear and, to the extent possible, persuasive the construction of reality and the value weightings implicit in a test and its application" (Cronbach, 1988: 5).

Different kinds of inferences require different kinds of evidence for their evaluation. Scoring rules that take us from observed performances to a score generally rely on expert judgment about the criteria to be used in scoring and on quality control of scoring procedures (and possibly on data regarding rater accuracy and consistency), as well as statistical models for scaling and equating. Generalizations from a sample of observations to expected performance over domains of possible observations rely on evidence that the sampling is representative of the intended performance domain and is large enough to control sampling errors. Inferences from assessment scores to different kinds of performance in different contexts rely on empirical evidence (e.g., from an empirical regression analysis) and/or on analyses of the overlap in the skills required for the different kinds of performances in different contexts. Inferences to some future criterion performances (i.e., predictions) require empirical evidence of the accuracy of the predictions. The justification of decisions based on the scores requires an analysis of the likely consequences of such decisions, positive and negative, intended and unintended.

As noted earlier, some of the assumptions in the IUA may be accepted a *priori* or with minimal support, but other assumptions may require extensive empirical support. In particular, strong claims (e.g., causal inferences or predictions of future performance in different contexts) would typically require strong empirical support. The most questionable assumptions should get the most attention in the validity argument. That is, the structure of the validity argument depends on the structure of the IUA, and more ambitious interpretations and uses will typically require more evidence than more modest interpretations and uses.

The development and appraisal stages of validation

The validation of a proposed interpretation and use of assessment scores typically involves two stages, which can overlap quite a bit but involve somewhat different kinds of evaluative analyses.

In the *development stage*, the goal is to develop (or adopt) an assessment and to develop an IUA that links the assessment outcomes to the proposed interpretations and uses of the scores. Initially, a preliminary IUA is outlined, and an assessment plan is developed. The IUA makes some assumptions about the assessment and about the resulting scores, and the assessment procedures would be designed in ways that make it likely that the assumptions in the IUA will hold.

If some assumptions are found to be violated during test development, the assessment or the IUA, or both, can be modified to remove the discrepancies. This process of development and revision continues until the test developers are satisfied with the fit between the test and IUA. The evidence produced during the development stage tends to support the validity of the proposed interpretations and uses because, if any evidence indicates a weakness in the assessment design or in the IUA, the problem is likely to be fixed in some way. For example, if the generalizability/reliability of the scores is found to be inadequate, the test might be lengthened, thereby improving the generalizability of the scores.

The development stage is expected to deliver a clear statement of what is being claimed in terms of an explicit IUA linking test performances to the proposed interpretations and

uses of the scores. For low-stakes applications involving relatively plausible inferences and assumptions (e.g., a teacher-made test used for formative assessment), the evidence derived from the development phase may be sufficient for the validation of the interpretation and use.

For high-stakes applications, a more extensive and arm's-length evaluation of the proposed interpretation and use is also called for. In the *appraisal stage*, the IUA is subjected to challenge, preferably by a neutral or skeptical evaluator. If the validity of the proposed IUA is to be evaluated by the assessment developers, as is often the case, they should seek to identify and examine the challenges that might be posed by a skeptical critic.

The appraisal stage would begin with a critical review of the IUA, with particular attention to identifying any hidden assumptions and alternative plausible interpretations of the test scores and would provide a critical evaluation of the assumptions built into the IUA. Without the discipline imposed by an explicit statement of the claims being made, it is easy to take much for granted.

The appraisal stage would also focus on empirical investigations of the most questionable assumptions in the IUA. As Cronbach (1980: 103) has suggested:

> The job of validation is not to support an interpretation, but to find out what might be wrong with it. A proposition deserves some degree of trust only when it has survived serious attempts to falsify it.

To the extent that the IUA survives these challenges, confidence in the claims being made increases. If these analyses reveal serious weaknesses, either the assessment or the IUA has to be revised or abandoned.

It is appropriate for test developers to adopt a confirmationist approach during the development stage, in which the goal is to develop a plausible IUA, based on an appropriate assessment. During the appraisal stage, it is appropriate to adopt a more critical stance.

The argument-based approach to validation was developed mainly as a way of facilitating the process of validation. By the 1990s, a general unified model of construct validation had been developed (Cronbach, 1971; Cronbach and Meehl, 1955; Messick, 1989), but the model did not include a clear strategy for conducting validations, and there was a widespread perception that validation was a daunting and potentially endless process. The argument-based approach was designed to retain the generality inherent in the theoretical frameworks while providing a more straightforward approach to the implementation of validation efforts (Chalhoub-Deville and Deville, 2006; Cronbach, 1988; Kane, 2006, 2013; Chapelle, 1999; Chapelle *et al*., 2008; Xi, 2008; Bachman, 2005; Bachman and Palmer, 2010). The process outlined in the argument-based approach is basically quite simple. First, state the claims that are being made (explicitly or implicitly) in a proposed interpretation or use (the IUA), and, second, evaluate these claims (the validity argument).

Developing the IUA and the validity argument

The testing procedure and the IUA are interdependent and are best developed in tandem. It is also possible, of course, to develop IUAs and validity arguments for an existing assessment, but in these cases, the choices to be made in developing the interpretation and implementing applications will be constrained by the properties of the existing assessment.

An effective way to develop an assessment and an associated IUA starts with the conclusions and decisions included in the proposed interpretation or use and works backward to

the kind of assessment and the inferences required to support the proposed conclusions and decisions (Bachman and Palmer, 2010; Kane, 2006). For example, if the goal is to measure students' levels of proficiency in spoken English, we can begin by identifying the range of performances and contexts that are of interest. We could then develop an assessment plan specifying the kinds of tasks to be included in the assessment and the criteria to be used in evaluating responses to these tasks. The most direct approach would involve observing extended performances in which the test taker is required to use the language effectively in one or more real or simulated situations. A less direct approach would involve observing a large number of brief, structured performances in which the test taker is required to use the language to achieve some specific goal. Either of these or some combination of these or other approaches might work well in a particular case.

We start with the interpretation and uses that are of interest and then figure out how to get to these interpretations and decisions. And, as we move backward from the conclusions and decisions to the observations on which they are based, we develop a chain or network of inferences, and we identify the assumptions on which these inferences depend. The conclusions to be drawn and the decisions to be made define the end point of the IUA but do not specify how to get there. There may be several IUAs that could support the same conclusions and decisions, starting from different kinds of assessment procedures. Some of these assessment procedures and IUAs are likely to be more appropriate or plausible than others.

The development of the assessment and the IUA is likely to be iterative. First, we develop a preliminary assessment plan and a general version of the IUA that would lead from the observed performances to the conclusions and decisions to be made. Second, we develop a validity argument by evaluating the plausibility of the IUA in terms of the coherence of the argument and the plausibility of its inferences and assumptions (particularly its more questionable assumptions) in light of all available evidence. If gaps or weak links are identified, it may be necessary to adjust the interpretation and/or the assessment or to conduct additional research. Third, we restate the IUA and the validity argument and repeat the evaluation of the IUA until the validity argument provides adequate support for the IUA, or the IUA is rejected. An IUA that survives all reasonable challenges to its inferences and assumptions can be accepted (with the caveat that new challenges may arise in the future). Once the IUA has been specified, it can be played forward, using the network of inferences to draw conclusions based on each test taker's performances and to make decisions based on the conclusions.

The IUA will generally contain a number of inferences and assumptions, and the studies to be included in the validation effort are those that provide critical checks on the inferences and assumptions in the IUA under consideration (Chapelle et al., 2008, see also Chapelle, this volume). It is the content of the interpretation that determines the kinds of evidence that are most relevant and, therefore, most important in validation.

Some examples

As indicated earlier, the development of an IUA linking the assessment scores to the proposed interpretation and use is a critical component of effective validation. If we don't know what is being claimed, it is hard to evaluate the plausibility of the claims. Although certain types of interpretations or uses (e.g., placement testing systems) have a common structure, each IUA is unique in many ways (e.g., the specific claims made, the contexts, the populations assessed), and, therefore, the associated validity argument is also unique.

In order to make this discussion more concrete, it is useful to consider a few examples of what is involved in developing IUAs, their associated assessments, and the validity arguments.

Observable attributes/performance assessments

Simple inductive inferences can be used to draw general conclusions about people based on samples of their performance. If we observe a child riding a bicycle with ease, we can reasonably conclude that the child can ride a bicycle. If we observe a student solving quadratic equations, we can conclude that the student can solve quadratic equations. This is a very basic and useful way to draw conclusions about competencies, skills, and dispositions.

Observable attributes can be defined in terms of how well persons can perform some kind of task or how they respond to certain situations (e.g., how persistent they are in trying to perform an assigned task). Cronbach and Meehl (1955) refer to this kind of attribute as an "inductive summary" and suggest that such variables can be defined entirely in terms of descriptive dimensions and need not involve any theory.

For standardized measures of observable attributes, the observed performances are scored using some scoring rules, which reflect the intended interpretation of the observable attribute, and the value of the observable attribute can be defined in terms of the average score over some domain of possible performances. The attribute is observable in the sense that its interpretation is specified in terms of a domain of possible standardized performances that can be observed and scored.

Observable attributes are basically simple inductive summaries from observed performances to a larger domain of possible performances. Yet the definition of observable attributes raises many questions about how the target domain should be defined (Bachman and Palmer, 2010; Chalhoub-Deville, 2009).

Target domains

The domain of performances over which the observable attribute is defined will be referred to as a *target domain* for the observable attribute, and a person's expected score over the target domain is the value of the observable attribute for the person (Kane, 2006). The target domain involves some range of possible tasks, performed under some range of possible conditions, and it defines what we mean by the observable attribute, as an observable attribute. The decisions made in defining the IUA can rely on tradition, on experience, on structural similarities in the tasks and contexts, or on assumptions about the processes involved in the performances, but once the domain is specified, it defines the observable attribute.

A major determinant of the target domain is, of course, the intended uses of the scores. In educational contexts, in which specific skills are being taught, a target domain could focus on relatively simple tasks requiring the use of some skill. The skill of forming plural nouns could be defined in terms of a target domain in which the tasks involve taking a singular noun and turning into a plural noun. Note that even this narrowly defined domain contains a very large number of possible tasks (involving all nouns in the English language and perhaps a number of simple task formats).

A much broader target domain could be specified for reading comprehension in English. In the context of college admissions, the reading tasks that are of interest would probably involve the kinds of prose found in textbooks, literature, and perhaps journals and other

documents and would typically not assume any specific prior knowledge. In adult literacy classes for immigrants, most of the attention might be on everyday reading materials like signs, menus, timetables, newspapers, and so on. Note that in defining these target domains, the tasks are not restricted to test items. In drawing conclusions about reading comprehension for college admissions, we want to draw inferences about a student's ability to read a text (e.g., a chapter, a textbook, a novel), and the target domain includes these extended tasks. In a test of reading comprehension, we may restrict ourselves to short passages followed by multiple-choice questions, but the target domain for the observable attribute is not restricted to specific kinds of test tasks.

In some cases, the content of the target domain for language use may be determined by the operational requirements of a particular kind of activity and/or context (e.g., in aviation). The tasks in this domain would involve samples of the activity, and success in these tasks might require a wide range of skills, including linguistic competence, content knowledge, familiarity with certain conventions, and interpersonal skills. In other cases, the domain definition may reflect a set of related skills and abilities that are not tied to any particular setting or any particular kind of goal (e.g., the ability to understand a variety of spoken utterances in a language and to respond appropriately).

The argument-based approach does not suggest that the target domain for an observable attribute be defined in any particular way. It does suggest that the target domain be defined as clearly as possible and that assessments of the observable attribute be evaluated in terms of how well they represent the corresponding target domain. In particular, it is important to avoid the tendency to define the target domain in one way (e.g., narrowly) for validation while implicitly adopting a broader and more ambitious domain as the basis for score use.

Note that observable attributes are not operational definitions. Operational definitions focus on specific measurement operations and tend to be narrowly specified. The target domains for observable attributes are intended to be defined broadly enough to include the full range of performances associated with the attribute label and with the intended use of the scores.

We can add additional meaning to an observable attribute by assuming that the behavioral regularities in the target domain are due to an underlying *trait* of the person. A trait is not observable; its existence is inferred from observed regularities in performance, but the trait, per se, is not observable. General assumptions about the trait and its relationship to other traits and to the performances in the target domain can play a significant role in interpreting test scores and in defining the boundary of the target domain, but it is not necessary to make this kind of assumption in order to define the observable attribute.

The target domains for the observable attributes that are of most interest in language assessment are not restricted to standardized assessment tasks or to assessment contexts, although they include this kind of structured performance as a subset. A person's level of literacy in a language depends on his or her ability to perform a variety of reading tasks in a variety of contexts, ranging from the casual reading of a sign to the careful reading of a textbook or manual. The fact that it might be difficult to include some of these activities (reading the menu and discussing the options with several people in a restaurant) in an assessment does not indicate that they be removed from the target domain for the attribute.

The target domain is likely to be very large and not very precisely defined. Whether a particular kind of document or a particular situation should be included in the target domain may be questionable, but some ambiguity at the edges is not a fatal flaw. The point is not to develop a tidy domain or one that is easy to assess, but to identify the range of performances associated with the attribute of interest.

Performance assessments

An obvious way to assess a person's level of competence in in a domain is to observe the person's performances on a sample of tasks from the domain and use the average score on these performances to estimate the person's expected score over the target domain. For example, we could use observed performance on a sample of reading tasks drawn from a domain of reading tasks to estimate proficiency on this kind of task. If the observed performances can be considered a random or representative sample from the target domain, this generalization from the sample to the target domain would be a straightforward statistical inference; the IUA would involve two main inferences, an evaluation of the observed performances, and a generalization of the observed results to a much larger target domain of possible performances (Kane *et al*., 1999).

The scoring inference takes us from the observed performances to an observed score, which represents an evaluative summary of the observed performances. The scoring rule (including the scoring rubrics and any statistical models used for scaling, equating, etc.) needs to be consistent with the intended interpretation and use of the scores, and the scoring rule has to be implemented accurately and consistently if the proposed interpretation is to be plausible.

The observed score can then be used as an estimate of the expected value over the target domain, the target score. In generalizing from the observed score to the target score, the score does not change, but the interpretation is extended from the observed performance to the expected performance over the target domain.

The evidence supporting the evaluation of the test taker's performances would involve justifications for scoring rubrics and administration procedures. The evidence for the generalization to the mean over the universe of possible performances defining the observable attribute would involve procedural evidence supporting the representativeness of the sampling and empirical estimates of the reliability or generalizability of the statistical inference from the observed score to the target score.

Standardized assessments

Assessments are generally standardized in a number of ways and for a number of reasons, and, therefore, the observed performances are not representative of the full target domain. Some performances from the target domain are likely to be excluded for practical reasons; the assessment has to be completed in a reasonable period of time and at modest cost. Some performances may be ruled out for reasons of fairness; we would generally not want to include tasks or situations that might make some test takers very uncomfortable or that might give some candidates an unfair advantage or subject them to an unfair disadvantage.

The observations are actually sampled from a *universe of generalization* for the measurement procedure, and a person's expected score over the universe of generalization is the person's *universe score* (Brennan, 2001). While the target domain for adult literacy would include a very wide range of written material (e.g., novels, instructional manuals, magazines, memos, signs), responses (answering specific questions, giving an oral or written summary, taking some action based on the manual or sign), and contexts (e.g., at home, in a library, at work), the universe of generalization for a measure of literacy may be limited to responses to a set of questions following short passages while sitting at a desk or computer terminal. The performances involved in standardized reading tasks are legitimate examples of literacy, but they typically constitute a very narrow slice of the target domain for literacy.

That is, the universe of generalization from which the observed performances are drawn is a relatively small and non-random subset of the target domain.

Because the samples of tasks included in measurements are not usually random or representative samples from the target domain, it is not generally legitimate to simply generalize from the observed score to the target score. It is, however, often reasonable to generalize from the observed score to the universe scores for the universe of generalization associated with the measurement procedure.

The IUAs for the standardized assessments of observed attributes require three main inferences. The scoring inference goes from the test results to an observable score, the observed score is generalized to the universe score, and the universe score is extended, or extrapolated, to the target score.

Extrapolation from the universe of generalization to the target domain extends the interpretation from the universe score to the target score. The extrapolation inference assumes that the universe score is related (rationally and/or empirically) to the target score and that the extrapolation is relatively free of systematic and random error (Bachman and Palmer, 2010). Again, the score does not change, but the interpretation of the score is extended from the universe of generalization for the assessment to the target domain that is of interest.

Target domains and universes of generalization for language proficiency

The definition of a target domain can raise a number of serious issues. We tend to think that we know the kinds of performances subsumed under terms like language proficiency and literacy, but identifying the range of performances to include in the target domain and the emphases to be given to different kinds of performance tends to get complicated once we attempt to specify the domain well enough to use it as a guide in developing an assessment plan. These issues get progressively more complex as we move from the mechanics of language use to general conceptions of communicative competence (Bachman and Palmer, 2010; Fulcher and Davidson, 2007).

For example, in assessing competence in a language (first or second language), the target domain can focus on the skills needed for basic interpersonal communication or on the skills needed for academic settings at various levels. The cognitive/academic competency domain is likely to include vocabulary and grammatical structures that are not generally used in everyday conversations (Chalhoub-Deville, 2009).

On a more pragmatic level, we have a range of options in deciding on the kinds of performances to use in the assessment and in the relationship between the universe of generalization for the assessment and the target domain. As noted earlier, one option is to sample randomly or representatively from the target domain. Under this approach, the performances that could be observed are likely to be quite varied and, at least in some cases, quite complicated and lengthy. The observed performances may be considered to be relatively authentic, but they also have at least two potential disadvantages. First, in drawing performances from a broadly defined and diverse target domain, the sampling variability is likely to be quite large, and in evaluating the generalization inference, this variability will be considered to be a major source of errors of measurement. This problem can be addressed by drawing large samples of performance and possibly by adopting efficient sampling plans, but the cost of these options tends to be high.

The second problem associated with sampling directly from the target domain arises because the effectiveness of a person's performance in various tasks will depend in part on

knowledge and skills that go beyond what we usually think of as language skills, and performance can vary substantially from task to task and context to context. To the extent that the goal is to assess facility in using a language, specific content knowledge or lack of such knowledge that might have an effect on performance on specific tasks or in specific situations is said to be construct-irrelevant variance (Messick, 1989), and it can lead to biased results for individuals and groups.

The standard way to address this problem without using very large samples of performance is to select relatively neutral performances, neutral in the sense of not introducing emotional factors in cognitive tests and not requiring any knowledge or skills not explicitly associated with the intended interpretation and use of the test scores.

In some cases, we may be interested in assessing how well individuals can communicate in contexts in which effective communication requires specialized knowledge and vocabulary, specialized skills (e.g., interpreting graphical representations of data), or specialized conventions, and in these cases, employing tasks and contexts from the domain of interest will tend to enhance the accuracy of extrapolation from the assessment to the target domain (Bachman and Palmer, 2010). If the target domain is defined in terms of tasks that require specialized knowledge or skills, it is appropriate that the assessment tasks also require the relevant knowledge or skills. This kind of target domain defines a kind of communicative competence that is different from general linguistic competence, and it is important to be clear about the difference (Widdowson, 2001).

Placement testing

In a sequence of language courses (e.g., foreign language courses in a university), the goal is to develop increasing proficiency in the use of the language in some range of contexts. Each course builds on the previous courses in the sequence. The first course would typically be designed for novices, with little or no skill in the language, and would aim to develop some rudimentary proficiency in the use of the language. Success in the first course allows students to move on to the second course, where the competencies are developed further. Similarly, the third course builds on the first two, and so on.

If all students entered the language program as complete novices, there would be no need for a placement testing program: in this case, all students would start in the first course and continue in succeeding courses until they completed the sequence or stopped for some reason. However, many of the incoming students may have acquired some proficiency in the language (e.g., in high school courses, at home, or by spending time in locations where the language is used), and it would be useful to be able to place each student into a course that would be most appropriate for the student. The intended use of the placement program is to place each student into a course in which she or he will be challenged but will have a good chance of success (Frisbie, 1982).

We already have a placement system for students who have started out in the first course. If these students complete the first course successfully, say with a grade of B or better, they are allowed (or encouraged) to take the second course. Students who get a B or better in the second course are allowed/encouraged to take the third course, and so forth.

This system does not apply to new students who have some competence in the language because they do not have grades in any of the courses. However, if we can develop an assessment that reflects the competencies developed in the courses and can get the students who are currently completing the various courses to take this test, we can use scaling methods to link specific test scores to the level of competence indicated by scores of B in each

of the courses (Frisbie, 1982). The test scores can be used as proxies for grades of B in the courses.

Identifying the test score that corresponds to the achievement level associated with a grade of a B or better in a course can be difficult because the test scores will not be perfectly correlated with course grades (which are not perfect measures of achievement in a course), and some students with a grade of C in a course may have higher test scores than students with a grade of B in the course. One way to get around this difficulty would be to form test score distributions for students with a B and with a C; these distributions can be smoothed to iron out some of their variability, and the cutscore can be set at the point on the test-score scale where the two distributions cross.

In addition to the basic scoring inference (from performances to the observed score), placement testing systems involve two inferences and a decision rule that depends on these inferences. The first inference is from scores on the test to conclusions about level of achievement in the competencies developed in the sequence of courses. The second inference is from competency, as measured by the test, to conclusions about level of achievement relative to performance at the end of courses in the sequence. The third inference is a decision about suggested course placement, based on cutscores on the test-score scale. This is not the only way to think about placement systems, but it is a useful framework.

Inference from test scores to level of achievement

The inference from assessment scores to conclusions about level of achievement in the competencies developed in the sequence of courses depends on at least four assumptions:

1 The courses are designed to develop a cluster of competencies, with a student's overall level of competency expected to increase from one course to the next in the sequence.

This assumption is likely to be supported by the judgment of those involved in running the courses. If the courses do not have this kind of sequential, hierarchical structure, a simple placement test of the kind outlined here would not make much sense.

2 The test assesses the competencies developed in the courses over the range covered by the courses.

This assumption is likely to be supported by an analysis of the competencies evaluated by the assessment relative to those developed in the courses. Assuming that the assessment is developed to reflect the competencies developed in the courses, much of the work done in test development will contribute evidence in support of this assumption. The test may not cover all the competencies developed in the courses, but it should cover most of the competencies.

3 Scores on the test are generalizable across relevant facets (e.g., tasks, scorers, occasions).

The scores should be generalizable enough for dependable conclusions about each student's level of competence. This assumption is likely to be supported empirically in reliability or

generalizability studies, indicating that the scores do not fluctuate too much over potential sources of error.

4 There are no sources of systematic error that would interfere with the intended interpretation.

There are at least two main sources of systematic errors in the assessments: construct-irrelevant variance and construct underrepresentation. Any source of score variability that does not reflect the competencies to be assessed is a source of construct-irrelevant variance. For example, an assessment that employed formats that were not familiar to the students could underestimate student achievement, and students with disabilities may need accommodations.

Language courses generally seek to develop competencies that involve listening and speaking, as well as reading and writing, and at least some of the courses may focus on the ability to communicative effectively in various contexts. To the extent that the assessment does not fully cover all these competencies, it may suffer from construct underrepresentation. Some omissions may not cause problems, especially if the omitted competencies do not play a large role in the courses in which most placements occur and if level of achievement in the tested competencies is highly correlated with level of achievement in the non-tested competencies, but the tradeoffs involved in covering most competencies versus the cost in time and resources deserve serious consideration.

If these four assumptions can be supported in a convincing way, it would be reasonable to accept the claim that scores on the placement test can be interpreted as measures of the competencies developed in the courses. The analyses of this first inference rely on methods traditionally categorized under the heading of content-related validity evidence.

Inference from test score to end-of-course grade

The second inference is from competency as measured by the test to levels of achievement corresponding to a grade of B or better in the courses in the sequence.

5 The levels of achievement needed to get a B or better in the courses increase across the sequence, and the assessment scores are positively related to these levels of achievement.

This assumption would be supported by evidence indicating that students completing higher-level courses generally get higher scores on the assessment and, in particular, that students completing higher-level courses with a B or better get higher scores on the assessment.

Placement decisions

The final inference is a decision about suggested placement of students into courses based on level of achievement, as indicated by test scores, relative to cutscores on the test-score scale. The courses in the sequence require increasing levels of competence as prerequisites, and students whose assessment scores indicate that they have mastered the competencies

developed in a course well enough to get a B or better in that course are generally prepared to do well in the next course.

6 Students are placed in the highest-level course for which they seem to be prepared because they have a level of achievement comparable to the achievement of students getting a B or better in the previous course in the sequence.

As indicated earlier, the cutscores can be determined empirically by administering the assessment to students completing the courses and identifying points on the assessment score scale corresponding to the boundary between students who get a B or better and students who get a grade lower than B. The procedures used to develop the assessment and to set the cutscores can provide strong support for this inference.

In addition, the decisions being made can be evaluated by investigating how well the students who are placed in various courses do in these courses, student self-evaluations of how well prepared they were for the course, and faculty evaluations of whether the placement seemed appropriate.

The point of this exercise is not to develop a definitive outline for placement testing, but rather to illustrate an argument-based framework in which the intended interpretation and use of the assessment scores are outlined in some detail (i.e., as an IUA), with an emphasis on the inferences being made and the assumptions required for these inferences to be plausible. The assessment is developed to satisfy the requirements inherent in the proposed interpretation and use (i.e., the development phase), and the effectiveness of the assessment system in use is evaluated (the appraisal phase).

Concluding remarks

The argument-based model provides a relatively pragmatic approach to validation. The goal is to develop a measurement procedure that supports the intended interpretations and uses of test scores and an IUA that is plausible, given the measurement procedure and the proposed interpretations and uses.

An explicitly stated IUA serves three functions. First, it provides a framework for test development by indicating the assumptions that need to be met. A measure of competence in the use of a language for everyday activities would be different in many ways from the development of a language test for a particular academic or work setting, in which specialized concepts and terms are especially important.

Second, the IUA provides a framework for the validation of the proposed interpretation and use, as specified in the IUA. The evidence called for in the validity argument is that needed to support the inferences and assumptions in the IUA.

Third, the IUA provides a basis for evaluating the adequacy of the validity argument. If the validity argument provides adequate support for the coherence of the IUA and for the plausibility of its inferences and assumptions, the proposed interpretations and uses are warranted. If the IUA is incomplete or some of its inferences or assumptions are shaky, the proposed interpretations and uses are not fully justified.

The theory of validity has evolved quite a bit over the last century, especially since the publication of Cronbach and Meehl's (1955) seminal article on construct validity, but arguably, the practice of validation has not changed that much. Validity theorists have tried to attend to the range of issues that arise in practice, but in doing so, they tended to make

validity theory progressively more general and abstract, and as a result, the relevance of the theories to practice became less obvious.

The argument-based approach to validation was intended to reverse this trend by making the theoretical framework for validation more responsive to the details of proposed interpretations and uses of assessment scores in particular contexts and with particular populations. However, in some ways, the argument-based approach to validation is even more general than other validity frameworks. In particular, it does not propose any particular kind of interpretation or use. Rather it suggests that the proposed interpretations and uses and the intended contexts and target populations be specified in the form of an IUA and that the appropriateness of the claims being made in this IUA be evaluated. However, implementing these suggestions is, in most cases, no easy task, and expositions of the argument-based framework do not generally specify who is to do this work: the assessment developer or user or some third party.

The analysis of practical arguments plays out in very different ways in different areas and for different kinds of arguments (Bachman and Palmer, 2010; Kane, 2006; Toulmin, 1958). So, in addition to a general argument-based theory of validity, more specific frameworks for the evaluation of IUAs for different kinds of interpretations and uses (placement, certification, diagnosis), in different areas (language learning, science), and in different contexts are needed to facilitate applied work in assessment development and validation. That is, we need validity frameworks at various levels of generality between the abstract, general models and specific applications. In language testing, much conceptual work has been done on how to support the development and validation of different kinds of assessments for different purposes (Bachman and Palmer, 2010; Chapelle *et al*., 2008), but there is always a lot more to do.

Further reading

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2: 1–34.

Bachman (2005) provides a very clear and concise introduction to an argument-based framework for validation, with an emphasis on the evaluation of test uses in terms of the consequences of such use. He examines the kinds of warrants and evidence needed for a utilization argument over and above those typically needed for the validations of a score interpretation per se.

Chapelle, C. A., Enright, M. and Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle, M. Enright and J. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY: Routledge, 1–25.

Most of the examples of IUAs and validity arguments in the literature are quite simple and quite general and have been developed for illustrative purposes. Chapelle, Enright, and Jamieson (2004) developed and documented an IUA and a validity argument for a large and complex testing program and, in doing so, have pushed the argument-based methodology forward in a dramatic way by applying the approach to an existing assessment program.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52: 281–302.

It took a while for the sophisticated framework introduced by Cronbach and Meehl to gain traction, but it became the base for most work on the theory of validity from the 1980s to the present. It introduced a conceptualization of score interpretation and validation that was much broader and more sophisticated than those that had been in use up to that point. This article is over 50 years old and does not provide specific guidance for validation, but it is essential reading for anyone who wants to follow current debates on validity and validation.

Kane, M. T. (2006). Validation. In R. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: American Council on Education and Praeger, 17–64.
Kane provides an overview of an argument-based approach to validation and indicates how some common interpretations and uses of test scores can be validated.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50: 1–73.
Kane (2013) provides an updated version of the argument-based validation described in Kane (2006), with greater emphasis on validating score uses.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (AERA, APA, and NCME). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2: 1–34. https://doi.org/10.1207/s15434311laq0201_1.

Bachman, L. F. and Palmer, A. (2010). *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford, UK: Oxford University Press.

Brennan, R. (2001). *Generalizability Theory*. New York, NY: Springer-Verlag.

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. Lissitz (ed.), *The Concept of Validity*. Charlotte NC: Information Age Publishing, 241–63.

Chalhoub-Deville, M. and Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: Praeger Publishers, 517–30.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics* 19: 254–72. https://doi.org/10.1017/S0267190599190135.

Chapelle, C. A., Enright, M. and Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle, M. Enright and J. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY: Routledge, 1–25.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (ed.), *Educational Measurement*, 2nd edn. Washington, DC: American Council on Education, 443–507.

———. (1980). Validity on parole: How can we go straight? In *New Directions for Testing and Measurement: Measuring Achievement Over a Decade*, *5*. San Francisco, CA: Jossey-Bass, 99–108.

———. (1988). Five perspectives on validity argument. In H. Wainer and H. Braun (eds.), *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum, 3–17.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52: 281–302. https://doi.org/10.1037/h0040957.

Frisbie, D. A. (1982). Methods of evaluating course placement systems. *Educational Evaluation and Policy Analyses* 4: 133–40. https://doi.org/10.3102/01623737004002133.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London, UK: Routledge.

Kane, M. T. (2006). Validation. In R. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: American Council on Education and Praeger, 17–64.

———. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50: 1–73. https://doi.org/10.1111/jedm.12000.

Kane, M. T., Crooks, T. J. and Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice* 18: 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: American Council on Education and Macmillan, 13–103.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge, UK: Cambridge University Press.

Widdowson, H. G. (2001). Communicative language testing: The art of the possible. In C. Elder, A. Brown, E. Grove *et al*. (eds.), *Experimenting with Uncertainty, Essays in Honor of Alan Davies*. Cambridge, UK: Cambridge University Press, 12–21.

Xi, X. (2008). Methods of test validation. In E. Shohamy and N. Hornberger (eds.), *Encyclopedia of Language and Education: Language Testing and Assessment*, 2nd edn, Vol. 7. New York, NY: Springer, 177–96.

3

Inference and prediction in language testing

Steven J. Ross

Approaches to inference

Validity has long been the central concern of language assessment practice and research. Historically, validity has been conceptualized in various ways and in two main strands. The objective approach to testing validity has relied on concurrent, criterion, discriminant, and construct validation argumentation supported by empirical evidence, primarily with the use of correlation, regression, and covariance structure analysis methods. Validity in this approach is made evident through representative sampling of genuine test candidates, who provide quantitative test records ranging from a single predictor with a criterion test to the use of a large battery of tests devised to differentiate putative test constructs from the testing methods used (Sang et al., 1986; Llosa, 2007). The subjective approach to validation, in contrast, requires interpretation of construct representation through sufficiently authentic test content coverage (Lissitz and Samuelson, 2007); an ecological match between test design and norming with actual test uses; a sufficiently strong degree of face validity of the test to the various stakeholders; and most recently, an interpretation of the immediate and long-term consequences of test use (Chalhoub-Deville, 2016). The subjectively interpretive approaches to validation depend on expertise and consensus about the test outcomes and the desirability of the processes by which they have come about.

In the last two decades, much effort has been expended to subsume the varied approaches under a single framework (see Chapelle, this volume). Bachman (2005), Chapelle *et al*. (2008), and Kane (2006) have sought to situate the different facets of validity evidence under a more coherent test use argument. By subsuming the empirical and interpretive evidence into a single test use argument, inferences about test validity are more readily restructured to assertions about validity to be compared to counter-interpretations of the evidence. While the test use argument approach has done much to provide a coherent framework, lacunae remain, especially in the subjective interpretation of what constitutes adverse societal consequence of test uses (Chalhoub-Deville, 2016).

Inferences about test validity use both empirical and interpretive approaches. In the case of the former, inferential statistics, primarily those predicated on the frequentist null hypothesis testing approach, assess observed validity coefficients relative to either a random

 DOI: 10.4324/9781003220756-5

variation criterion or, in the form of tests of how well a model fits observed data, as in covariance structure analysis (Hancock and Mueller, 2006). The validity coefficients might be as simple as a zero-order correlation coefficient (Bernstein et al., 2010; Ginther and Yan, 2017) or regression coefficients in models devised to control for covariates or nuisance variables (Bridgeman and Cho, 2012). The interpretation of observed correlations or regression coefficients as larger than what would be observed through random variation is, however, generally recognized as an insufficient criterion for making strong validity inferences. Correlation coefficients are sample dependent and, when used for test validation purposes, can change substantively when non-random samples are intentionally stratified or when subsets of test candidates are removed from the aggregate (Bridgeman et al., 2016). The evidence of the validity of tests correlated with criterion outcomes has also long been known to vary when data sets are aggregated (Blyth, 1972). Validity studies with large samples generally show that even miniscule correlations between a test and the criterion can still be "significant," requiring a secondary interpretation of how large a validity coefficient would have to be in order to be substantively generalizable and meaningful. In such cases, the validity coefficient is often converted to an effect size, which then requires a substantive interpretation of the effect size magnitude. This approach has proven useful in experimental research (Klein, 2013), aiding in the interpretation of the meaningfulness of the effect size and whether there is any practical significance beyond rejection of the null hypothesis. For validity coefficients based on correlations and regression models, effect sizes such as the percent of shared or accounted-for variance still require interpretation of the test's practical significance and utility.

Inferences about criterion outcomes

Test validity research involves the use of criterion outcomes, which are themselves often variable and fuzzy. Ginther and Han (2018), Cho and Bridgeman (2012), and Bridgeman *et al*. (2016), for instance, have used first-year college grade-point averages as the criterion variable in their validity research on TOEFL iBT. Two issues emerge with the use of a criterion such as a grade-point average. One is the possibility that the GPA as an indicator of educational success may vary dramatically across instructors, courses, majors, or institutions. The low validity coefficients observed in such research can lead to many interpretive challenges as the criterion itself may be unstable or, in some cases, be of questionable validity.

The second and more onerous problem in the correlational approach to validation is the truncation of data. A language test used in the admissions process usually has a locally defined cut-off score, below which the predicting measure is not observable. This circumstance produces a missing counterfactual condition, which, in hypothetical-deductive experiments, is conceptualized as the value of the criterion measure at the case level that would have been observed if the members of the control condition had received the intervention of interest (Morgan and Winship, 2007). In language test validation, the analogous counterfactual condition refers to the unobserved scores on the criterion measure for the candidates who were not selected for instruction or admission. Test validation researchers could conceivably work around this inferential dilemma experimentally: for instance, by conducting the validity research at institutions that offer open or non-selective admissions. Assuming that the admitted students would have a sufficiently strong incentive to attempt the language test not actually used for admissions screening, the resulting measures could provide

conceivably larger validity coefficients as the grade-point averages, graduation rates, and other relevant outcomes would be available, as well as a full range of language test scores.

The more commonly used solution to the data truncation problem is to correct the observed validity coefficients for restriction of range (Ginther and Yan, 2017; Linck *et al*., 2013). In this approach, the unrestricted validity coefficient is considered the upper bound of the validity of the predictor against the criterion. This approach is contingent on the availability of a stable estimate of the standard deviation of the population of test takers. For large-scale standardized tests, such estimates are generally available. Ultimately, corrections for restriction of range and attenuation provide, at best, the theoretical upper bound for the validity of a test used for candidate selection.

Interpretive inference

Recent efforts to specify the relations among constructs, tasks devised to measure such constructs, and performance characteristics have been applied to a variety of language assessments. The evidence-centered design approach (Mislevy *et al*., 2002; Mislevy (2017), Chapelle *et al*. (2008), building on Kane *et al*. (1999), uses a set of inferential procedures derived from Toulmin's (2003) argument structure for the integration of models of examinee performance, construct validation, and task design specifications, which collectively strengthen the linkage of examinee performance characteristics to the constructs the performances are meant to instantiate. These schematic model formats require a multi-step interlinking chain of arguments from the domain description to the utilization of assessment results. Between each step, evidence-based justifications are required for the ultimate validity claims to be made. For performance assessments, the evidential basis for test score interpretation crucially depends on the validity of arguments linking the body of information or prior experience (Mislevy, 2017) to the test data that justifies interpretive claims. The test use argument approach has proven practical in that it can integrate different forms of empirical evidence and can also include interpretive data, such as that gathered in proficiency interviews.

The core inferential task centers on the comparison of the warrant and its backing against the potential rebuttal, with whatever evidence that supports it. The positioning of the evidence for a claim against the counter-evidence sets up a form of temporary inference rather than one that supports an absolute or irrefutable claim of validity. In the integrated framework now familiar in language testing validity research, the warrant and backing for a validity claim are compared to the evidence brought up as a potential rebuttal to those claims. The inference drawn from the weighing of the evidence and counter-evidence is taken to support the validity claim or challenge it. The warrant plus backing versus rebuttal plus its backing can be seen as a comparison requiring inferences about which side provides the best explanation.

In a validity argument framework, there is a possible inferential step at each decision point for which there is data amenable to interpretation. As Purpura *et al*., 2015 argue, there are possible inferences about scoring process, the generalizability of the sample of items and persons, the explanation that links the test construct to the particular sample of items and tasks devised to represent it, and an inference about how the test scores relate to criterion measures or outcomes that a language test would be expected to predict. Chapelle *et al*. (2008) and Chapelle (2020) asserted, in the context of TOEFL test design, that the three inferential steps of evaluation, generalization, and extrapolation outlined by Kane *et al*. (1999) would not adequately link test performance to a theoretical construct such as

language proficiency. Accordingly, the inferential chain was extended to include an explanatory inference to link observed performances to a latent construct, a decision-making or utilization inference to rationalize institutional policy decisions to mandate the use of particular tests, and a domain inference to specify the content of test tasks.

At all steps of the inferential process, evidence-based comparisons are essential. Lipton (2004) has referred to the head-to-head comparison of evidence as requiring abductive inference, or "inference to the best explanation," a process of reasoning that pits evidence in support of a validity claim against the available counter-evidence. The evidence that best explains the observed phenomenon is taken as the most plausible. In language test validation, claims of validity are weighed against plausible counter-claims at each point at which an inference is to be made (Kane, 2006). Abductive inference is likely to be variably applicable to each of the six inferential steps outlined by Chappelle *et al*. (2008), as evidence favoring backing or refutation might not in all cases be empirically grounded.

As Chalhoub-Deville (2016) has noted, the test use validity argument approach may still fall short when the test's societal impacts or consequences are brought into the discussion. Consequences may be deduced from descriptive statistics when test data are disaggregated and, to some degree, be a reflection on the ideological stance taken by stakeholders interpreting the descriptive statistics. In such instances, the "best" explanation may ultimately be subjective and remain open to alternative interpretations, especially when it cannot be firmly anchored in unambiguous evidence that serves to unravel the complex of factors leading to the controversial consequences.

Prediction

Many language tests are utilized for candidate selection, placement, and certification. For these purposes, language testing research and practice generally involve inferences about predictive validity. Predictive validity is considerably more difficult to investigate because the outcomes of interest evolve distally relative to the preceding interpretation of test scores and the decisions made about them. Prediction inherently entails a set of assumptions relating the antecedent test to the consequential outcomes of interest. Predictive accuracy is also subject to several moderating factors, such as intervening variables, selective attrition, and the reliability of the criterion measure. The core rationale for the use of a language test such as TOEFL or IELTS and for aptitude tests such as MLAT, DLAB, and HILAB is that the constructs the tests measure are antecedent to the person-level variation observed later in the criterion outcomes. Selection tests thus always precede the candidate selection process, and their predictive validity presumes that the predictor and criterion do not dynamically co-evolve. A close examination of many assessment and individual difference constructs appearing in the language testing and applied linguistics research literature suggests that the assumption of antecedent-to-consequent relation between the test and the outcome is often ambiguous and without firm empirical support.

As outlined in Davies (1985) and Asher (1983), a core assumption in a predictive argument is the pre-existence and anterior measurement of the predictors relative to the criterion outcomes. A review of the research literature on phenomena such as proficiency, aptitude, and motivation, however, provides ample cases in which the time of measurement ambiguously overlaps with the measurement of the outcome. Regression models and structural equation models based on concurrent measures via surveys or one-shot testing sessions, for instance, often do not satisfy the empirical assumption that the predicting variables preexist the outcomes they are thought to predict. Correlation data based on data collected

concurrently in the same setting are most vulnerable to this weakness. A survey of learners' motivation, willingness to communicate, self-efficacy, or test anxiety given just after a language test, or even after a program of instruction, could undermine a researcher's presumption that such affective variables are causal factors accounting for individual differences in language proficiency (cf. Gardner, 2000). The inference that learner motivation, as a case in point, is an antecedent of successful learning would assume that the motivational state of the language learner pre-dates the time of the outcome testing and that the individuals tested were ignorant of all their prior learning histories. The latter condition is hardly plausible, leading to much ambiguity because relatively motivated language learners typically are aware of their language learning success.

Time-mediated regression models such as in path analysis models (e.g., Noels, 2001) make the assumption that anterior measures, usually described as predictors, have evolved prior to the downstream measures or dependent variables. For grounded claims of causal inference in such observational (non-experimental) studies, the anterior measures are assumed to be measured well before, and logically evolve prior to, the measured outcome variables. This condition generally requires a longitudinal research design during which the language acquisition outcomes have ample time to evolve through instruction or immersion and for individual differences to emerge. Elaborate path or covariance structure models essentially face the same time-ordering challenges that simple correlations do. Researchers generally have to make the assumption that the predictors pre-exist the outcomes, even though they are often measured at the same time.

One step removed from the antecedent to consequent approach to prediction is one that samples groups whose members represent levels of an outcome, such as second language proficiency. The usual goal in this kind of research is not obviously predictive but assumes some of the conditions of prediction and, for some research questions, serves as a proxy for prediction (Linck *et al*., 2013). The goal in group-based analysis is to identify the optimal set of measurements that can be used to ultimately predict group membership. This approach often uses logistic regression or multinomial regression models. More frequently used in language testing research is discriminant analysis.

Discriminant analysis and binary or multinomial logistic modeling approaches have subtle assumptions that in language testing validity research are easy to overlook. One essential assumption for discriminant analysis is that the groups to be differentiated by a set of language test measures were formed through "natural" processes (Tabachnick and Fidell, 2007). This assumption implies that the groups are not synthetically constructed out of measures that were originally synonymous with, or proxies for, the set of predictive measures to be used to discriminate between or among them. If this assumption is violated, test validation research may slip into a circularity of reference that can plague attempts to test the coherence of frameworks such as the Common European Framework of Reference (Hulstijn and Schoonen, 2011) and the Interagency Language Roundtable (Long *et al*., 2012). Essentially, "natural" groups of proficient foreign or second language speakers are not readily identifiable independently of some form of prior language assessment.

Discriminant function analysis normally requires two data sets and analyses – a retrodictive analysis that finds the optimal subset of measures serving to discriminate among the natural groups, followed by a predictive phase in which the discriminant function weights from the retrodictive phase are used to predict future and still unknown group membership. Validation, then, is a matter of correspondence between predicted group membership and the observed outcomes. If the original sample is large enough, a common strategy is to randomly split the sample of cases into the retrodictive set and save the other half for the

prediction phase. Since the actual group membership is known in the focal subset of cases, predictive validity is assessed thorough a comparison of matches between the predicted and actual group memberships. This approach is at best a proxy for future predictive validity. More convincing are predictions of group memberships that evolve in the future. A potential complication to the future group discrimination approach occurs, however, when the predicted membership in a yet-to-be-formed group is itself used diagnostically as the basis for a pre-emptive intervention. A discriminant model used, for instance, to predict dropping out or higher risk of instructional failure, when pre-emptively used to assess risk and then to prevent the unwanted outcome can result in a missing or truncated counterfactual – individuals who were at risk of failure received the intervention based on the discriminant function analysis predictions and thus were prevented from failing and dropping out. Validation in such cases generally requires historical data to support the inference that the prediction model and subsequent interventions reduced the frequency of incidences of the problem that otherwise would have occurred.

Aptitude and prediction

Another area of recent second language acquisition research involving predictive validity has focused on affective and cognitive factors hypothesized to account for individual differences in second language learning outcomes. These include constructs like willingness to communicate, motivation, self-efficacy, and cognitive aptitude for language learning. In this chapter, the focus will be on cognitive aptitude for language learning (DeKeyser and Koeth, 2011; Skehan, 2012; Linck *et al*., 2013; Doughty, 2018). The premise of cognitive aptitude, particularly that hypothesized to predict individual differences in second language acquisition success, is that it is largely influenced by working memory and executive functioning through a componential set of stable traits that do not change substantively with experience. Cognitive aptitude is held to be distinct from earlier models of language aptitude, such as the Modern Language Aptitude Test (Carroll and Sapon, 1959) and the Defense Language Aptitude Battery (Petersen and Al-Haik, 1976), which focus primarily on language analysis skills. Cognitive aptitude for language learning, in contrast, is conceptualized as a set of innate individual differences in executive functioning (Doughty, 2018).

Aptitude tests are for the most part used for candidate selection. Two genres of language learning aptitude tests have been used for candidate selection. The linguistic analysis type primarily requires test takers to inductively decode novel language data. The long-standing model for this kind of aptitude test is the MLAT (Carroll and Sapon, 1959). The usual operational policy or decision-making inference is to select only candidates who reach a pre-set threshold of language learning aptitude – eventually compared with criterion scores such as cumulative grade-point averages from intensive immersion training. Figure 3.1 illustrates such a case in which the grade-point averages on the vertical axis are correlated with the language learning aptitude measure on the horizontal axis. The scatterplot in Figure 3.1 indicates there is a moderately positive correlation between the predictor (aptitude) and the outcome (language learning success) – albeit one that is depressed by data truncation, owing to the fact that only candidates scoring above 100 on the standardized aptitude measure were selected for training. The scatterplot also reveals that there are outliers sitting well off the smoothed regression line in both the average-aptitude-but-high-GPA and low-GPA-but-high aptitude patterns.

As noted earlier, in many real-world, non-experimental applications, evidence to support predictive validity is affected by the absence of the counterfactual, or evidence of *what*

would have happened otherwise. In some relatively rare instances, candidates who would have been excluded from foreign language immersion training because of insufficient aptitude are nevertheless selected. A frequently used reason for such selection is that, as is sometimes the case for military linguists, low language learning aptitude trainees possess non-linguistic technical skill sets that are in demand in operational theaters. The exceptional cases provide a limited view of what outcomes would be observed in spite of the trainees' relative low aptitude for language learning and provide a partial sample of the counterfactual condition. Figure 3.2 shows a scatterplot of intensive immersion foreign language training outcomes for a training cohort that included a number of trainees selected in spite of their low aptitude for language learning.

Noteworthy in Figure 3.2 is the negative relation between grade point average and aptitude for the trainees who are below the usual cutoff for selection. While the sparse sampling makes generalization difficult, the overall non-linear relation between aptitude and the cumulative grade point average would suggest that factors other than aptitude may be influential for the subset of trainees selected for their non-linguistic attributes.

For the more typical trainees, those selected for their language learning aptitude above 100, the scatterplot reveals a moderate positive correlation similar to that seen in Figure 3.1. Also noteworthy in Figure 3.2 is how the smoothed regression line shows clusters of cases well above and below the best-fitting line. Higher aptitude scores here only inconsistently lead to higher grade-point average outcomes.

More recent innovations in aptitude test development have integrated executive functioning tasks in the more conventional language analysis battery. Doughty (2018) summarized the current state of research on the hybridization of cognitive aptitude tests with language analysis–type aptitude measures. For both types of language aptitude tests, a core assumption is that aptitude is stable and unaffected by language learning experiences. The scarcity

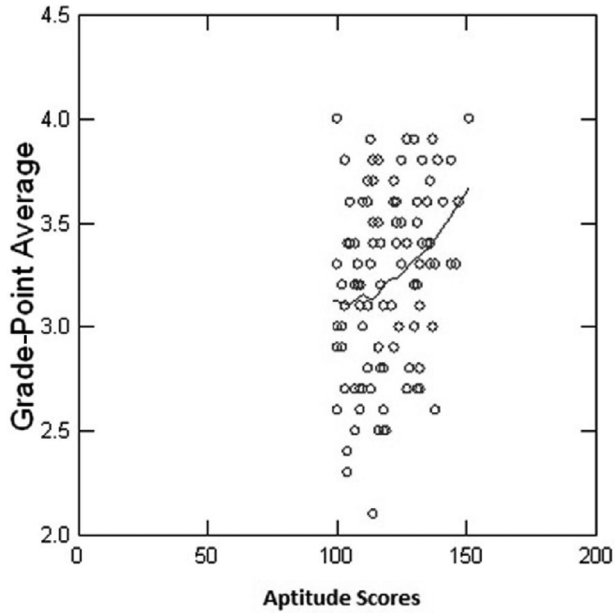


Figure 3.1 Scatterplot of grade-point average with truncated aptitude

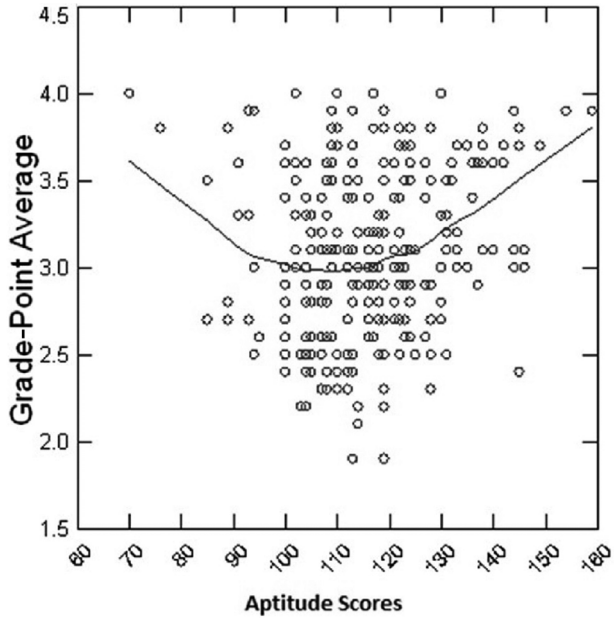


Figure 3.2 Scatterplot of grade-point average with low-aptitude trainees

of empirical research on this core assumption that cognitive aptitude, in particular, is immutable provides the opportunity for longitudinal validation research.

A longitudinal approach to prediction

The thesis that cognitive aptitude for language learning is an immutable individual difference phenomenon remains an unresolved issue in the second language acquisition research literature. Baum and Titone (2014) note that studies of the "bilingual advantage" in executive functioning of bilingual adults relative to monolinguals have been almost exclusively cross-sectional, leaving open the influence of the bilinguals' life histories, varying as they might across an array of experiential factors. The question of the etiology of individual differences in the main cognitive aptitude thought to affect success in adult second language acquisition, executive functioning, remains open to further exploration. The view that such cognitive aptitudes are innate (Granena, 2013; Doughty, 2018) would be supported if individual differences in cognitive aptitude were observed prior to extensive immersion training in language and were the dominant long-term predictor of relative success. Some of the inferences about the influences executive functioning has on adult language learning have come from research on the bilingual advantage (Luk *et al*., 2010). As Morton (2014) notes, if executive functioning is an inheritable trait unaffected by training or practice (Owen *et al*., 2010; Friedman *et al*., 2008), no other contextual or socio-cognitive factors (Mislevy, 2017) or programs of intervention (Redick *et al*., 2013) would influence learning outcomes dependent primarily on executive functioning. Noting the abundance of uncontrolled factors in cross-sectional comparisons of bilinguals and monolinguals, Morton (2014) asserts that there are a plethora of possible alternative explanations. One question that Baum and Titone (2014) pose concerns whether executive control is stable over the lifespan of the bilingual.

If it is conceivable that some aspects of executive control are affected by long-term second language acquisition, variation in the success of adults acquiring a second language could co-vary with repeated measures of executive functioning or control. It is noteworthy that studies of the malleability of executive functions have tended to be based on short-term interventions quite unlike adult second language acquisition in extended immersion settings. For a predictive validation of cognitive language learning aptitude, a longitudinal study in a realistically long immersion setting would be needed to provide the empirical evidence supporting the immutability of cognitive aptitude hypothesis.

One approach to testing a simultaneous change hypothesis involves repeated measures of parallel processes (Singer and Willett, 2003; Kieffer and Lasaux, 2012, Ross, 2005). For an empirical test of concurrent change, dual processes, such as repeated measures of language learning outcomes capturing individual differences in relative success, and repeated measures of cognitive aptitude collected in tandem with periodic retesting of the same language learners would be required. Figure 3.3 provides a model of such a parallel change process in a covariance structure analysis framework (Singer and Willett, 2003; Hancock and Mueller, 2006).

The top portion of Figure 3.3 shows three repeated measures of proficiency (Prof 1 to Prof 3). These observed measures would be separated by equal time periods during which intensive immersion in a foreign language would be experienced by a large and stable cohort of language learners. The repeated measures of proficiency capture individual differences at the start of instruction, yielding a latent intercept (Prof. Level). Individual differences in language learning success, as measured by equated or parallel repeated measures, would be indicated in the latent growth factor (Prof. Growth). The variation in the proficiency growth factor would be the primary criterion variable in the parallel growth analysis, as it would measure the variation in proficiency growth across individuals over a period of language learning like that typically observed in immersion programs.

The lower portion of Figure 3.3 shows the repeated measures of cognitive aptitude for language learning. Like the repeated measures of proficiency, those for cognitive

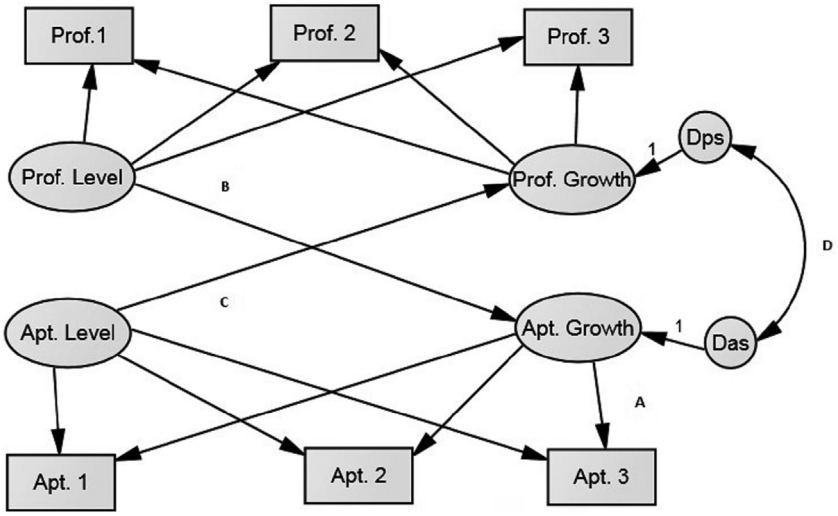


Figure 3.3 Parallel growth model for proficiency and aptitude

aptitude – that is, measures of executive functioning and control – would ideally be equated or parallel forms.

Four hypotheses would be testable in the parallel growth model. Individual differences in cognitive aptitude, if immutable, would not be expected to vary across time. Factor loading A in Figure 3.3 would thus be expected not to differ from the loadings observed from the aptitude growth latent variable to the first and second tests of aptitude measured in parallel with language learning. The second hypothesis would predict that individual differences in language proficiency measured at the outset of immersive language learning would not predict individual differences in change in cognitive aptitude over time. Path B would be predicted to be small and only randomly varying.

Central to the cognitive aptitude as predictor of variation in proficiency hypothesis would be Path C. Here, the individual differences observed at the outset of cognitive aptitude testing would be expected to have a strong and non-random influence on the variation on the individual differences in proficiency growth. Moreover, the magnitude of Path C would be predicted to be substantively larger than the co-evolution hypothesis, represented in Path D. The co-evolution hypothesis tests whether language learning success itself co-varies with individual differences in changes in cognitive aptitude over time. If cognitive aptitude is immutable and unaffected by variation observed in the proficiency measures over time, only the individual differences observed at the outset of aptitude testing will predict individual differences in language learning success. That is to say, Path C will be substantively larger than Covariance D when both are standardized to a common metric. Conversely, if cognitive aptitude, as measured by executive control tasks, is itself mutable and co-evolves with language learning experiences, Path D will be equal to or larger than Path C. The advantage of a parallel growth model such as that in Figure 3.3 is that the predictive validity of cognitive aptitude can be tested simultaneously with an alternative hypothesis.

Summary

Inferences about predictive validity are among the most challenging in the realm of language testing. Evidence derived from zero-order or bivariate correlations are not likely to resolve interpretive ambiguities because prediction inherently assumes an antecedent to consequent relation, which imposes the need for longitudinal designs. Cross-sectional studies in general, those that rely on comparative correlations, or group-based mean difference comparisons usually fall short of establishing a strong foundation for validity inferences when questions of predictive validity are posed.

While relatively costly and susceptive to factors such as non-random attrition and the difficulty of developing equated or parallel measures of proficiency, aptitude, or motivation, longitudinal studies of predictive validity are more likely to provide the bases for robust inferences about the relative magnitude of factors hypothesized to influence variation in language learning success. Longitudinal studies of language learning have been relatively rare to date, and indeed, studies that are designed to rigorously test predictive claims are rarer still. Future research on phenomena such as language learning aptitude, self-efficacy, and motivation stand to benefit the most from research based on longitudinal designs that provide probative empirical support for predictive validity claims.

Further reading

Lipton, P. (2004). *Inference to the Best Explanation*, 2nd edn. London: Routledge.

Peter Lipton summarizes the history of causal inference from a philosophy of science perspective. Lipton also provides an overview of how hypotheses evolve and how evidence is considered in light of plausible alternatives using observational, experimental, and Bayesian methods of interpretation.

Mislevy, R. J. (2017). *Sociocognitive Foundations of Educational Measurement*. New York: Routledge.
In this comprehensive text, Robert J. Mislevy lays out an approach to integrating measurement theory, learner experiences, and contextual factors into the interpretation of measurement outcomes in an evidence-centered design framework.

Morgan, S. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.
Morgan and Winship outline the quantitative methodology needed to untangle the complexities of direct and indirect causal arguments, especially in non-experimental and ex-post-facto designs.

Singer, J. and Willett, J. (2003). *Applied Longitudinal Data Analysis*. New York: Oxford University Press.
Singer and Willett provide an accessible and comprehensive overview of quantitative methods used in longitudinal data analysis, including multi-level modeling, latent growth curve models, and event history analysis.

References

Asher, H. B. (1983). *Causal Modeling*. Sage University Paper series on Quantitative Methods in the Social Sciences. Beverly Hills and London: Sage Press.

Bachman, L. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2: 1–34. https://doi.org/10.1207/s15434311laq0201_1.

Baum, S. and Titone, D. (2014). Moving toward a neuroplasticity view of bilingualism, executive control, and aging. *Applied Psycholinguistics* 35: 857–94. https://doi.org/10.1017/S0142716414000174.

Bernstein, J., van Moere, A. and Cheng, J. (2010). Validating automated speaking tests. *Language Testing* 27: 355–77. https://doi.org/10.1177/0265532210364404.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association* 67: 364–66. https://doi.org/10.1080/01621459.1972.10482387.

Bridgeman, B., Cho, Y. and Depietro, S. (2016). Predicting grades from English language assessment: The importance of peeling the onion. *Language Testing* 33: 307–18. https://doi.org/10.1177/0265532215583066.

Carroll, J. and Sapon, S. (1959). *Modern Language Aptitude Test*. New York: American Psychological Corporation.

Chapelle, C., Enright, M. and Jamieson, J. (2008). Test score interpretation and use. In C. Chapelle, M. Enright and J. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language*. New York: Routledge.

Chapelle, C. (2020). *Argument-based Validation in Testing and Assessment*. London: Sage.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing* 33: 453–72. https://doi.org/10.1177/0265532215593312.

Cho, Y. and Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing* 29: 421–42. https://doi.org/10.1177/0265532211430368.

Davies, J. A. (1985). *The Logic of Causal Order*. Sage University Paper series on Quantitative Methods in the Social Sciences. Beverly Hills and London: Sage Press.

DeKeyser, R. and Koeth, J. (2011). Cognitive aptitudes for second language learning. In E. Hinkle (ed.) *Handbook of Research in Second Language Teaching and Learning*. New York: Routledge.

Doughty, C. J. (2018). Cognitive language aptitude. *Language Learning* 69: 101–26. https://doi.org/10.1111/lang.12322.

Friedman, N., Miyake, A., Young, S. DeFries, J., Corley, R. and Hewitt, J. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General* 137: 201–25. https://doi.org/10.1037/0096-3445.137.2.201.

Gardner, R. C. (2000). Correlation, causation, motivation, and second language acquisition. *Canadian Psychology* 41: 1–23. https://doi.org/10.1037/h0086854.

Ginther, A. and Yan, X. (2017). Interpreting the relationships between TOEFL iBT scores and GPA: Language proficiency, policy, and profiles. *Language Testing* 35: 271–95. https://doi.org/10.1177/0265532217704010.

Granena, G. (2013). Cognitive aptitudes for second language learning and the LLAMA language aptitude test. In M. Long and G. Granena (eds.), *Sensitive Periods, Language Aptitude, and Ultimate L2 Attainment*. Amsterdam: John Benjamins, 105–30.

Hancock, G. R. and Mueller, R. D. (eds.). (2006). *Structural Equation Modeling: A Second Course*. Greenwich, CT: Information Age Publishing.

Hulstijn, J. H. and Schoonen, R. (2011). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the common European framework of reference (CEFR). *Language Testing* 29: 203–21. https://doi.org/10.1177/0265532211419826.

Kane, M. T. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: American Council of Education and Praeger Series on Higher Education, 17–64.

Kane, M., Crookes, T. and Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice* 18: 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x.

Kieffer, M. J. and Lasaux, N. K. (2012). Development of morphological awareness and vocabulary knowledge in Spanish-speaking language minority learners: A parallel process latent growth curve model. *Applied Psycholinguistics* 33: 23–54. https://doi.org/10.1017/S0142716411000099.

Klein, R. (2013). *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*, 2nd edn. Washington, DC: American Psychological Association.

Linck, J. A., Hughes, M. M., Campbell, S. G., Silbert, N. H., Tare, M., Jackson, S. R., Smith, B. K., Bunting, M. F. and Doughty, C. J. (2013). Hi-LAB: A new measure of aptitude for high level language proficiency. *Language Learning* 63: 530–66. https://doi.org/10.1111/lang.12011.

Lipton, P. (2004). *Inference to the Best Explanation*, 2nd edn. London: Routledge.

Lissitz, R. and Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher* 36: 437–48. https://doi.org/10.3102/0013189X07311286.

Llosa, L. (2007). Validating a standards-based classroom-based assessment of English proficiency: A multitrait-multimethod approach. *Language Testing* 24: 489–515. https://doi.org/10.1177/0265532207080770.

Long, M., Gor, K. and Jackson, S. (2012). Linguistic correlates of second language proficiency: Proof of concept with ILR 2–3 in Russian. *Studies in Second Language Acquisition* 34: 99–126. https://doi.org/10.1017/S0272263111000519.

Luk, G., Bialystok, E., Craik, F. I. M. and Grady, C. L. (2010). Lifelong bilingualism maintains white matter integrity in older adults. *Journal of Neuroscience* 31(46): 16808–13. https://doi.org/10.1523/JNEUROSCI.4563-11.2011.

Mislevy, R., Steinberg, L. S. and Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing* 19: 447–96. https://doi.org/10.1191/0265532202lt241oa.

Mislevy, R. (2017). *Sociocognitive Foundations of Educational Measurement*. New York: Routledge.

Morgan, S. and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York: Cambridge University Press.

Morton, J. (2014). Sunny review casts a foreboding shadow over status quo bilingual advantage research. *Applied Psycholinguistics* 35: 929–31. https://10.1017/S0142716414000277.

Noels, K. A. (2001). Learning Spanish as a second language: Learners' orientations and perceptions of their teachers' communication style. *Language Learning* 51: 107–44. https://doi.org/10.1111/0023-8333.00149.

Owen, A., Hampshire, A., Grahn, J., Stenton, R., Dajani, S. and Burns, A. (2010). Putting brain training to the test. *Nature* 465: 775–76. https://doi.org/10.1038/nature09042.

Petersen, C. and Al-Haik, A. (1976). The development of the defense language aptitude battery (DLAB). *Educational and Psychological Measurement* 36: 369–80. https://doi.org/10.1177/001316447603600216.

Purpura, J. E., Brown, J. D. and Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research1. *Language Learning* 65: 37–75. https://doi.org/10.1111/lang.12112.

Redick, T. S, Shipstead, Z., Harrision, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., Kane, M. J. and Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: A randomized, placebo-controlled study. *Journal of Experimental Psychology, General* 142: 359–79. https://doi.org/10.1037/a0029082.

Ross, S. J. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics* 26: 317–42. https://doi.org/10.1093/applin/ami011.

Sang, F., Schmidt, B., Vollmer, J., Baumert, J. and Roeder, P. (1986). Models of second language competence: A structural equation approach. *Language Testing* 3: 54–79. https://doi.org/10.1177/026553228600300103.

Singer, J. and Willett, J. (2003). *Applied Longitudinal Data Analysis*. New York: Oxford University Press.

Skehan, P. (2012). Language aptitude. In S. Gass and A. Mackey (eds.), *The Routledge Handbook of Second Language Acquisition*. London: Routledge, 381–95.

Tabachnick, B. and Fidell, L. (2007). *Using Multivatriate Statistics*, 5th edn. Boston: Allyn and Bacon.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge: Cambridge University Press.

Section 2

The uses of language testing


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

4

Social dimensions of language testing

Richard F. Young

> Toward the end of the academic year, a man with a clipboard turned up at one of the capital city's best high schools. He sauntered from classroom to classroom, ignoring the students and instead engaging in seemingly trivial chitchat with the teachers, 20 minutes at a time.
>
> "Tell me, what subjects are your specialties? How long have you worked here? Can you explain to me a little about how you prepare your lessons?" The inspector didn't seem to be particularly interested in what the teachers said. He only cared about how they said it.
>
> Olga Muravyova teaches biology and geography. She laughed nervously as she recalled her meeting with the inspector. "He wrote a report saying that I understood all the questions, that I answered all the questions, but that I made some errors. That is actually what he claimed," Ms. Muravyova said. "Of course, that is hard to hear."
>
> After the inspector told her that she had failed the test, he told her to attend Estonian classes, which she has tried to do. But she is 57, an age when it is not easy to pick up a new language.

This vignette, based on a story in the *New York Times* (Levy, 2010), describes part of a test of spoken Estonian required since 2008 of teachers and other civil servants. In this small former Soviet republic on the Baltic Sea, the government has been mounting a determined campaign to elevate the status of its native language and to marginalize Russian, the tongue of its former colonizer. Public schools, where students have long been taught in Russian, are now linguistic battlegrounds, and the test itself is a skirmish between Estonian and Russian. The test and Ms. Muravyova's experience illustrate two social dimensions of language testing. The first is the construct of language knowledge on which the test is built and on which test results are interpreted; the second is what happens to individuals, societies, and institutions when the test is used – the social consequences of assessment. In this chapter, I describe both of these dimensions and argue that, until quite recently, language tests have been built on incomplete knowledge of the social ground of language in interaction and the social consequences of language assessment.

DOI: 10.4324/9781003220756-7

The social life of language

Language testing is a branch of applied linguistics and applied linguistics grew out of earlier work by linguists. Linguists trace the history of their field to ancient India and usually to Pāṇini, who flourished in the fourth century BCE. One of Pāṇini's main concerns was the correct pronunciation of the body of oral chants of ancient poems known as the *Veda*, composed in an early form of Sanskrit that was already archaic by the fourth century. About a thousand years earlier in Shang Dynasty China, we have the earliest written records of a language that today we call Chinese. Ancient Chinese writing was found inscribed on animal bones and turtle shells used for divination. In those days, when a ruler wanted advice, he would ask the spirits of his ancestors and other supernatural beings a question, which was written by court officials on an oracle bone or turtle shell. This was then heated, and court officials interpreted the pattern of cracks in the bone as an answer to the question.

In these two ancient examples, interpreting language written on bones and correcting the pronunciation of sacred oral chants was done by people whom today we would call linguists. Linguists today do much the same thing; that is, they take a record in some form and consider (and oftentimes correct) its oral form just like Pāṇini in ancient India, or they interpret the meaning of a written record that has been responded to in some way just like the court officials in Shang Dynasty China. Throughout the long history of their field, linguists have been bound by the physical form of the records that they consider data. When a society transitions from one physical form of language to another – from an oral culture to restricted forms of literacy, for instance – people often complain that the written record does not do justice to the hurly-burly of oral social interaction. In the oral culture of Anglo-Saxon England, when only a select few could read or write, there were many metaphors contrasting the written record with the spoken word. Written words were called "mouthless speakers," "dead lifegivers," and "dumb knowledge-bearers." To the Anglo-Saxons, the written record of interaction was dead, dumb, and limited – a thing. In contrast, spoken interaction itself is not a thing but a live discursive practice, which the technology of writing took and alienated from the world in which it was originally created. As the Anglo-Saxon scholar O'Brien O'Keeffe (1990: 54) wrote, "The technology which preserves also kills."

Today, we can go beyond the remembrances of speech and written records that constrained ancient linguists and offended ancient readers. In the twenty-first century, we have still pictures, sound recordings, and movies, and thanks to these new records, our understanding of the forms and functions of language in human interaction – the social life of language – extends far beyond what we know about the forms that language takes in speech and writing. To revivify language, to put language back in its social context, a theory is needed that goes beyond records of disembodied and decontextualized records. The response is practice theory, developed by anthropologists (Bourdieu, 1977, 1990; Sahlins, 1981, 1985), sociologists (de Certeau, 1984; Giddens, 1984), and applied linguists (Erickson, 2004; Young, 2009), whose aim was to explicate the nature of social interaction in context. Practice theorists study discursive practices like language tests; this involves understanding not only the production of meanings by participants as they employ the communicative resources at their command but also how employment of such resources reflects and creates the processes and meanings of the community in which the test occurs. As Erickson (2004) wrote, although the conduct of talk in local social interaction is unique and crafted by local social actors for the specific situation of its use at the moment of its uttering, it is at the same time profoundly influenced by processes that occur beyond the temporal and spatial horizon of the immediate occasion of interaction. The goal of practice theory is to describe both the

global context of action and the communicative resources that participants employ in local action. When the context of a practice is known and the configuration of communicative resources is described, the ultimate aim of practice theory is to explain the ways in which the global context and the local employment of resources are mutually constituted. Features of global context are described later in this chapter, but how does the global context influence the nature of the test? In other words, how does the global context impact social constructs in language tests?

Social constructs in language tests

The format of Ms. Muravyova's Estonian test could be called a conversation, one in which one party asks the questions, and the other party responds. The aim of the test is to evaluate one party's knowledge of language. But what is knowledge of language? When language testers and score users interpret people's scores on a test, they do so by implicit or explicit reference to the construct on which the test is based. Almost all the constructs that underlie high-stakes language tests are theories of individual cognition that can be measured in one context (the test) and are stable enough to be ported to other non-testing contexts where the language is used.

This view of the independence of linguistic knowledge from context underlies the history of linguistics as far back as ancient India and China. In the early twentieth century, Saussure (1983) stated the distinction between internal linguistics and external linguistics, which was recently summarized by Lantolf (2006: 74). Lantolf wrote, "In essence Saussure drew a circle around language (Agar, 1994: 41) and proposed that inside-the-circle language, the proper and exclusive domain of linguistic science, was restricted to the study of grammar and dictionary." If the construct of cognitive ability underlying a language test involves linguistic knowledge that is independent of context – internalized language or I-language in Chomsky's (1986) formulation – then the testing context in which it is elicited is only important to the extent that it helps draw out underlying cognitive abilities (Chalhoub-Deville, 2003: 371). According to this construct, the knowledge of Estonian grammar and vocabulary that Ms. Muravyova demonstrated in her conversation with the inspector is independent of the manner in which it was elicited, and her ability and knowledge can be generalized to another context in which she might be called upon again to speak Estonian.

By its emphasis on the mutual constitution of local resources and global context in interaction, practice theory provides a very different interpretation of the construct of language knowledge, an interpretation that is best approached by considering how communicative resources are employed in different contexts. In place of knowledge and ability, in practice theory, the performance of a person in context is construed as a configuration of communicative *resources* – resources that a person employs together with others in a particular configuration. Assessment of a person in context is challenging, however, because from a person's performance on a test, not only do we wish to infer specific resources employed in the discursive practice of the test, but we also wish to know how the same person will perform in other practices. The challenge was eloquently expressed by Chalhoub-Deville and Deville (2005: 826):

> Evaluating test-takers' performance according to this model offers a conundrum. Generally speaking, we administer tests to, assign scores to, and make decisions about individuals for purposes such as selection, placement, assignment of grades/marks, and the like. If we view language as co-constructed, how can we disentangle an individual's

contribution to a communicative exchange in order to provide a score or assess a candidate's merit for a potential position?

The conundrum can be solved by considering the relationship between test performance and the construct underlying a test proposed by Messick (1989, 1996) and revisited by Chapelle (1998) and Norris (2008). Chapelle distinguished between three perspectives on construct definition: a construct may be defined as a *trait*, as *behavior*, or as *some combination of trait and behavior*. In a trait definition of a construct, consistent performance of a person on a test is related in a principled way to the person's knowledge and speech production processes. That is to say, a person's consistent performance on a test is taken to index a fairly stable configuration of knowledge and skills that the person carries around with them – and which that person can apply in all contexts. In contrast, in the definition of a construct as behavior, the consistent performance of a person on a test is related in a principled way to the context in which behavior is observed. That is to say, test performance is assumed to say something about a person's performance on a specific task or in a specific context, but *not* on other tasks or in other contexts – unless they can be shown to be related to the task or context that was tested.

Clearly, neither definition of a construct – as trait or as behavior – is satisfactory for theories of language in use because, as Bachman (1990: 84) emphasized, communicative language ability consists of both knowledge and "the capacity for implementing, or executing that competence" in different contexts of use. For this reason, it is desirable to consider the third of Messick's and Chapelle's definitions of a construct, which they refer to as the *interactionalist definition*. In an interactionalist validation of a test, a person's performance on a test is taken to indicate an underlying trait characteristic of that person, and, at the same time, the performance is also taken to indicate the influence of the context in which the performance occurs. The interactionalist definition is, in other words, a way to have your cake and eat it: to infer from test performance something about both practice-specific behavior and a practice-independent, person-specific trait.

If an interactionalist definition is to allow test users to generalize from performance in one context to another – that is, from the discursive practice of a test to other practices – then what is needed is a theory that relates one discursive practice to another in a principled way. We need to know whether testees like Ms. Muravyova have the skill to mindfully and efficiently recognize contexts in which resources are employed and to use them when participating in different practices. If the purpose of her conversation with the inspector is to discover whether Ms. Muravyova is able to use Estonian when teaching, we really need to know whether the communicative resources that she displayed in her conversation are portable to another discursive practice: teaching high school biology and geography, practices she currently does in Russian. Although the contexts of use are very different, it is an empirical question whether the same resources are required in the same configuration in the two contexts. To answer such an empirical question, McNamara (1997: 457) argued that what is needed is "a close analysis of naturally occurring discourse and social interaction [to] reveal the standards that apply in reality in particular settings." Such an analysis of discourse and social interaction is one objective of practice theory.

Language testing as discursive practice

An analysis of social constructs in discursive practice is characterized by three features (Young, 2011). First, analysis of language in social interaction is concerned with

communicative resources employed by persons in specific discursive practices rather than language ability independent of context. Second, it is characterized by attention to the co-construction of a discursive practice by *all* participants involved rather than a narrow focus on a single individual. Mehan (1982: 65) stressed the interactional nature of discursive practice when he wrote, "'Competence' becomes interactional in two senses of the term. One it is the competence necessary for effective interaction. Two, it is the competence that is available in the interaction between people." Mehan's focus on interaction was taken up later by Kramsch (1986: 367), who wrote:

> Whether it is a face-to-face interaction between two or several speakers, or the interaction between a reader and a written text, successful interaction presupposes not only a shared knowledge of the world, the reference to a common external context of communication, but also the construction of a shared internal context or "sphere of intersubjectivity" that is built through the collaborative efforts of the interactional partners.

Intersubjectivity is the conscious attribution of intentional acts to others and involves putting oneself in the shoes of an interlocutor. Originating in the phenomenology of Husserl (Beyer, 2007), intersubjectivity was first inferred empirically from studies of infant development by Trevarthen (1977, 1979). Examples include an infant's following the direction of an adult's direction of gaze when she points and recognition by an infant of transition-relevance moments in interaction with others.

Third, analysis of social interaction identifies a set of verbal, interactional, and nonverbal resources that participants employ in specific ways in order to co-construct a discursive practice. The resources employed by participants in social interaction are not the mouthless, dead, and dumb records that the Anglo-Saxons complained about. Communicative resources, like language, are embodied and include a participant's whole body; the physical presence and movement of the body; the muscles of the face, arms, and upper body – in particular gaze and gesture – and yes, of course, a participant's speech and writing.

When the communicative resources employed by all participants in a testing practice are specified, the problem of generalizability is resolved by identifying the particular configuration of resources that participants employ in a particular practice. Then, by comparing the configuration of resources in that practice with others, it is possible to discover what resources are local to that practice and to what extent those same resources are employed in different practices and how common resources are configured.

Definitions of verbal, interactional, and nonverbal resources and examples of how participants configure resources in various interactive practices were provided in Young (2008, 2009, 2011) and are revisited here. *Verbal resources* include the register of the practice, defined as a recognizable repertoire of pronunciation and lexicogrammar that occurs with high frequency in certain practices, the combination of which is associated with a specific activity, place, participants, or purpose. In addition, verbal resources that participants employ create certain kinds of meaning in a practice. In systemic functional grammar, Halliday (1994) identified verbal resources as instantiating *ideational, interpersonal*, and *textual* metafunctions. The ideational metafunction is how language mediates participants' construction of their experiences of the external physical and biological world, their own internal thoughts and feelings, and the logical relations among them. The interpersonal metafunction, as its name implies, is how language mediates speakers' and writers' personal and social relationships with other participants, such as their interlocutors and readers. The textual metafunction is how participants build sequences of discourse, organize the

discursive flow, and create cohesion and continuity within a linear and time-dependent flow. In every communicative exchange, participants employ all three metafunctions, although they may be configured in different ways in different practices. As Schleppegrell (2004) showed for the language of schooling, these three metafunctions are realized by various aspects of linguistic substance, and their configuration defines communicative registers and genres.

Interactional and nonverbal resources that participants use to construct a discursive practice include the selection and sequential organization of actions, the turn-taking system that participants use to manage transitions from one speaker to another, and the ways in which participants repair interactional trouble. Sequences of actions realized in speech or nonverbally have been investigated most rigorously within the framework of conversation analysis developed in the 1960s. Conversation analysts have observed that certain acts occur in ordered pairs and that the production of the first act of a pair gives rise to participants' expectations of the production of the corresponding second act of the pair in an adjacent turn by a different speaker. Schegloff and Sacks (1973) showed how this expectation works with question-and-answer in American English, but it applies equally to other sequences of two adjacent utterances produced by different speakers, including greeting-greeting, offer-acceptance/refusal, and call-response. Act sequences longer than adjacency pairs are not generally recognized by conversation analysts, but the study of discursive practices of the same type does, in fact show, that more than two actions occur quite regularly and that they usually occur in a sequence.

The system of turn taking in conversation was first described in detail by Sacks *et al*. (1974), who answered two basic questions about turn taking: How is the next speaker selected? And how do participants know when to end one turn and when to begin another? The question of next speaker selection was answered by Sacks *et al*. by means of the following algorithm.

1 If the current speaker selects the next speaker, then that party has the right and obligation to speak.
2 If no next speaker is selected, then self-selection may (but need not) occur. The first starter acquires rights to the turn.
3 If no next speaker is selected, then current speaker may (but need not) continue.

Sacks *et al*. answered the second question about transitions between turns by invoking the notion of the turn-constructional unit or TCU. Such a unit may be a unit of the lexicogrammar, of intonation, or a pragmatic unit (a complete idea), and, as Ford and Thompson (1996) pointed out, these units often coincide to make a complex TCU. In any given conversation, the moment at which a transition between speakers occurs is not necessarily at the boundary of a TCU, but speakers are able to predict when a boundary is forthcoming and are therefore able to project the completion of the TCU. Transitions between speakers occur at places when participants project the completion of the TCU, projecting not only the form of the next word but also the completion of larger lexicogrammatical, intonational and pragmatic units. Prediction is thus an important part of what recipients do when listening to talk in progress, and the place in an ongoing turn when participants are able to project the completion of the TCU is called a transition-relevance place or TRP. That is to say, participants do not necessarily take a turn *at* a TRP, but if they do, then they are more likely to do so at a TRP than elsewhere.

Repair is the treatment of trouble in talk in interaction. Trouble can be anything in talk to which participants in interaction orient as problematic. One participant may use a word that is misunderstood or misheard by another participant; one participant may realize that a phrase that they have just used is less preferable than another phrase. Although the source of trouble is often a word or phrase, it may be anything to which participants orient as repairable. Thus, the absence of the second pair part of an adjacency pair may elicit an apology, or, when a listener projects a TCU and takes a turn while the current speaker wishes to continue, this may be oriented to as an interruption. In many cases, however, the source of trouble in a repair is a choice of words or phrasing, and in understanding repair, conversation analysts have focused on two questions about the participants in the repair: In whose turn did the trouble occur? And who initiated the repair sequence? Beyond the participants, the analysis focuses on the sequence of actions in the repair. For example, a repair is called an *other-initiated self-repair* if the repair is initiated by a different participant from the one in whose turn the trouble source occurred, and the repair is completed by the same participant in whose turn the trouble source occurred. Excerpt 1 is an example of an other-initiated self-repair in Estonian between a government official and a client who needs information.

Excerpt 1

Pardon? (Gerassimenko *et al*., 2004)		
1	Client:	*Aga kallis see tööluba on?* (0.5)
		How much does this work permit cost?
2	Official:	*Kuidas?*
		Pardon?
3	Client:	*Allis tööluba on ?*
		How much does the work permit cost?
4	Official:	*Ei, tö-tööluba ei ole vaja*.
		No work permit is needed.

In Excerpt 1, Client's question in line 1 is indexed as a source of trouble by the following 0.5-second pause and by Official's repair initiation in line 2. Client concludes the repair by repeating the question in line 3. Client is the other who initiated the repair that is completed by Official (the self). Three other kinds of repair are classified in a similar way according to the participant who initiates the repair and the participant who completes the repair as *other-initiated other repair*, *self-initiated self-repair* and *self-initiated other repair*.

In addition to these, one further resource identified by Levinson (1992) is the way in which participants construct boundaries of a practice. In order for participants to establish mutual orientation to how what they say is creating a context in which the meaning of what they say can be interpreted (what Gumperz [1982, 1992, 1995] called *contextualization cues* in a discursive practice), participants must distinguish it from other practices in which contextualization cues are functioning differently. This is done by means of locating the boundaries of the practice – the opening and closing verbal or nonverbal actions in the sequence of a practice. Not all practices begin and end abruptly, and, in fact, boundaries of a practice may be vague, may be negotiated, or may be resisted by one or more participants; nonetheless, boundaries are essential for participants to establish mutual orientation to meaning.

In summary, then, we can describe a discursive practice by specifying the ways in which participants avail themselves of the verbal resources of register and modes of meaning, together with the interactional and nonverbal resources of action selection and sequencing,

turn taking, repair, and boundary construction. Taken together, these six resources are the fundamental building blocks of intersubjectivity in conversational interaction and the means by which participants craft their local identities as social actors in oral and literate practices. In order to describe local construction of identity, Goffman (1979, 1981) developed the concepts of *participation framework* and *footing*. For Goffman, the identity of a participant in interaction can be animator, author, or principal: animator is an individual engaged in the role of utterance production; author has selected the sentiments being expressed and the words in which they are encoded; and principal's position is established by the words being spoken, whose beliefs have been told and who is committed to what the words say. Three important corollaries of Goffman's theory are: (1) not all participants are necessarily physically present in the interaction – as McNamara (1997: 459) wrote, there are others "whose behavior and interpretation shape the perceived significance of the candidate's efforts but are themselves removed from focus"; (2) an individual's identity may change from moment to moment throughout the interaction – recognized by Goffman as changes of footing and much expanded by Bucholtz and Hall (2004) as tactics of intersubjectivity; and (3) the participation structure of the practice is the configuration of the identities of all participants, present or not, official or unofficial, ratified or unratified and their footing in the practice.

If we follow McNamara's call for a close analysis of naturally occurring discourse in order to compare it with the discourse of a testing practice, we will see to what extent the communicative resources employed and their configurations are similar or different. Such a comparison was carried out by Young and He (1998) in their collection of studies analyzing the spoken discourse of language proficiency interviews. The studies collected by Young and He addressed a simple comparative question: Is a language proficiency interview an instance of natural conversation? Most studies answered the question in the negative, pointing out that the system of turn taking and goal orientation of language proficiency interviews reflects the institutional context in which they are embedded, while in ordinary conversation, topics and turns are neither prescribed nor proscribed by a specific speech activity, and none of the participants has a predefined role in managing the conversation. Such a comparison may be extended to compare the test of spoken Estonian that Ms. Muravyova took and the discourse of classroom interaction in an Estonian high school. A corpus analysis will identify the features of Estonian vocabulary, grammatical structures, and pronunciations that occur with frequency in both practices. A systemic functional analysis will identify how participants in the test conversation and the classroom make ideational, interpersonal, and textual meanings and which participants make which meanings. A conversation analysis of both practices will reveal how participants select and sequence social actions; how they manage the turn-taking system, including transitions from one speaker to another; and the ways in which participants repair interactional trouble. A similar analysis will describe how participants orient to boundaries and transitions in the classroom and in the oral test.

An examination of these social constructs in tests and in the classroom also reveals how the physically present participants – Ms. Muravyova, her students, and the inspector – construct their identities in interaction: in particular, how they co-construct relative power. Power is not only the institutional control of people by a powerful group, nor is it just a mode of thought control, nor does knowledge imply liberation. For Foucault (1978; Foucault and Gordon, 1980), power is exercised in every social interaction, and its insidiousness lies in its very ordinariness. At his most explicit, Foucault wrote, "Power is everywhere; not because it embraces everything, but because it comes from everywhere" (1978: 93).

In discourse, power often involves controlling and constraining the contributions of non-powerful participants, and the system of allocation of turns in conversation is a particularly effective means of doing so. However, powerful participants are not alone in the exercise of power. Power in discourse is co-constructed by all participants – both the powerful and the non-powerful. For instance, the use of a technical register by one participant constructs power if other participants do not challenge that power by expressing their lack of comprehension. Another example is in Ms. Muravyova's language-proficiency interview. She understood and answered all the questions; in other words, the inspector controlled talk by means of allocating the next turn to her, thus constraining her right to speak. Non-powerful participants co-construct power by accepting the constraints imposed upon them. The student who doesn't understand a word may decide to search for the meaning in a dictionary or to ask somebody to explain it, and testees in a language proficiency interview may simply accept the fact that their discursive contributions will be limited.

But the deployment of communicative resources by persons in local interaction does not create these results alone. Power is also created by the system – organized social situations and political institutions that create enduring identities for individuals as testers and testees, as teachers and students, as officials and clients, and expectations for their roles in society. As I argue in the following section, language – more specifically, language testing – is the construction and reflection of these social expectations through actions that invoke identity, ideology, belief, and power.

The social consequences of language tests

After reviewing a variety of nationally mandated tests of language proficiency, McNamara and Roever (2006) concluded that through language tests, political goals affect language learning and the lives of testees at every level. Such influences of the global on the local are to be expected in practice theory, the ultimate aim of which is to explain the ways in which the global context and the local employment of resources are mutually constituted.

Mutual constitution means that the actions of a local practice are influenced by the global context *and* that the global context is influenced by local actions. In Ms. Muravyova's experience, the influence of the global context is painfully clear: the government of Estonia required her to take Estonian language lessons. It was the government that mandated that she be tested, it was the government that established a National Examinations and Qualifications Centre to administer the tests, it was this body that designed the Estonian language proficiency examinations, it was by this body that the language inspector was trained and paid, and it was this inspector who examined Ms. Muravyova. There is a long political trail before their 20-minute conversation, and their actions have a long history, a history that goes back long before the 1995 Estonian Law on Citizenship that decreed the test. Modern Estonian is a descendant of one or possibly two of the original Baltic-Finnic dialects. As long ago as the first century CE, the Roman historian Tacitus mentioned a language he called *aestii* (the Estonians' name for their language is *eesti*), but the language has had a long and difficult history with strong early influences from German and Finnish and, in the twentieth century, from Russian. When Estonia was invaded and occupied by the Soviet Union in World War II, the status of the Estonian language changed to the first of two official languages (Russian being the other one). In the second half of the 1970s, the pressure of bilingualism intensified, resulting in widespread knowledge of Russian throughout the country. The Russian language was termed the language of friendship of nations and was taught to Estonian children as early as kindergarten. Although teaching Estonian to non-Estonians in schools was

compulsory, in practice, learning the language was often considered unnecessary. The collapse of the Soviet Union led to the restoration of the Republic of Estonia's independence in 1991, and Estonian went back to being the only official language in Estonia.

Given the recent history of Estonia, the government's desire to strengthen and disseminate the Estonian language by requiring teachers and other civil servants to be proficient in it is understandable, but it is clearly an example of political goals affecting language testing. In other societies, too, political goals of ruling elites have had significant influence on examinations. Miyazaki (1976) described the long history of the imperial Chinese examination system, which allowed people from all walks of life to enter the prestigious and powerful imperial civil service. At the time it was instituted over 1,400 years ago, the system was designed to weaken the power of the hereditary aristocracy at court by requiring aristocrats and commoners to compete on equal terms for positions in imperial service. From a twenty-first-century perspective, the system may appear equitable, but it was not designed to be so, for the amount of preparation time required for candidates to study the classical texts on which examinations were based required a degree of economic support simply not available to poor people. Fulcher (2004, 2009) provided many other examples, in societies both ancient and modern, of attempts by political elites to gain control over the education system through testing and to establish norms. Tests are powerful means of political control, and tests are effective largely because, as Foucault (1995: 184) wrote, they are a generally accepted "normalizing gaze, a surveillance that makes it possible to qualify, to classify and to punish. It establishes over individuals a visibility through which one differentiates them and judges them."

The normalizing gaze of tests affects the lives of individuals like Ms. Muravyova, and at the same time, it has been effective in changing the status of languages in multilingual societies and, consequently, the power and prestige of their speakers. Shohamy (2006: 95–98) listed three ways in which language policy objectives are achieved by language tests. Tests are instrumental in (1) determining the prestige and status of languages (and thus maintaining the power of speakers of prestigious language varieties), (2) standardizing and perpetuating language correctness (and thus maintaining the subordinate status of speakers of non-standard varieties), and (3) suppressing language diversity (in favor of speakers of the prestigious standard variety).

Language tests preserve the prestige of the national language

In societies that Fulcher (2009) termed collectivist, the identity and value of individuals is equated with their membership in a collective unit such as a state, a nation, or an institution. One way in which membership is maintained is through use of a common language, and it follows that it is in the interest of the collective that a common language be preserved. Just as when the Estonian Soviet Socialist Republic formed part of the Soviet Union, it was in the interest of the collective to uphold Russian as the national language, in the independent Republic of Estonia today, it is in the interest of the collective to develop Estonian as the national language. Such a collectivist dynamic is in tension with a different political philosophy, termed individualism by Fulcher, which was described by Locke (1690: 95) as radically different: 'Men being . . . by nature all free, equal, and independent, no one can be put out of this estate, and subjected to the political power of another, without his own consent.' Tensions between collectivism and individualism abound in the plurilingual, multicultural, highly mobile societies of the early twenty-first century. According to Shohamy (2006), a

collectivist ideology drives the requirement in American public schools for students to be tested in English in order to graduate and to be tested, again in English, for admission to higher education. Taking an individualist stance, Canagarajah (2009) argued that in India, communities had developed local varieties of English to the extent that the language had now become "plurilingual English," and these varieties were the ones most appropriate for local schooling and testing.

Language tests help maintain standards

In language tests, norms of lexicogrammar and style are enforced by the evaluation of a response to a norm as either correct or incorrect. In the case of languages such as English, which are used in and among many different communities around the world, there are many varieties, both regional and international and both nativized and non-nativized. The role of language tests has generally been to maintain the international standard variety at the expense of regional standards or nativized varieties. As Lowenberg (1993) demonstrated in discussing the Test of English for International Communication (TOEIC), a widely used test of Standard English for international communication developed by the US-based Educational Testing Service, lexicogrammatical or stylistic variants prevalent in regional varieties of English are considered errors on TOEIC. Examples of TOEIC "errors," which Lowenberg argued are acceptable to educated speakers of non-native varieties of English, are the following italicized elements taken from TOEIC tests:

> His proposal met with a lot of *resistances*.
> Chemicals in the home *they* should be stored out of the reach of children.
> We *discussed about* the problem until a solution was found,
>
> *(Lowenberg, 1993: 102)*

The collectivist norms used to evaluate test performance on TOEIC are those of the group that uses standard American, British, or Australian English, no matter whether different groups adopt different norms. The normative gaze of the collective is internalized by speakers of nondominant varieties of the standard language as a subject position that requires them to accept personal responsibility for the communication problems that they encounter. In prescriptive grammars of English, examples abound of non-standard English "errors" such as *irregardless*, *you could of got one*, and *I could care less*. Those are all "errors" that were supposedly committed by native speakers of English, but the problematic status of the norms of standard English becomes, *like, ginormous* in plurilingual contexts or when English functions as a lingua franca, as Seidlhofer *et al*. (2006) have described.

Language tests suppress diversity

Perhaps most problematic for language assessment in multicultural settings are differences in discourse pragmatic norms between a socially dominant group and less dominant groups. Such differences are most often found in the contexts in which directness and volubility are evaluated positively and those in which the same degree of directness and volubility are evaluated negatively. Second language pragmatics is often assessed by discourse completion items such as Excerpt 2, in which verbal action is required.

Excerpt 2

Apology (Rover, 2005: 130)
Ella borrowed a recent copy of *Time* magazine from her friend Sean, but she accidentally spilled a cup of coffee all over it. She is returning the magazine to Sean.
Ella: _____
Sean: "No, don't worry about replacing it. I read it already."

The pragmatic ideology that Excerpt 2 promotes is that Ella should say something related to her action, and she should promise to replace the damaged magazine because of the rejoinder from Sean. In other words, the response of the party who has damaged the possession of another is entirely satisfied *verbally*. No action is required except verbal action, and yet there are occasions when a physical action may be more appropriate and more welcomed by the injured party than any words, although such a test item promotes the pragmatic ideology that verbal action is sufficient.

A critical analysis of language testing practices thus brings to the forefront the social dimension of language testing including speaker subject positions, lexicogrammatical norms, transcultural pragmatic conventions, and many other aspects of societal ideology. Practice theory proposes that the practices of language testing occur in contexts that are much broader than the testing practice itself, including not only the designers and takers of a particular test, but also the purposes for which the test is designed, the purposes for which people take the test, and the ends to which the results of the test are put. McNamara and Roever (2006) have stressed the importance of these broader political questions because, they argue, the requirement to distinguish between *them* and *us* has increased in intercultural societies and in a world of cross-border migration. Distinguishing between them and us is famously recorded in tests such as the password used by American defenders of the Bataan Peninsula against the Japanese in World War II. Stimpson (1946: 51) recounted an Associated Press dispatch from the Bataan front. The Americans discovered an infallible way to distinguish friendly troops from Japanese who attempted to pass the sentries at night dressed in American or Filipino uniforms:

> They simply pick a password with numerous l's, such as *lollapalooza*. Sentries challenge approaching figures and if the first two syllables of *lollapalooza*, for instance, should come back as *rorra* they open fire without waiting to hear the remainder.

Many more recent and less fatal ways of distinguishing friend from foe are recorded in the language assessment of immigrants, asylum seekers, and those who wish to become citizens. The political context of language testing is just as pertinent in widespread language testing enterprises resulting from the No Child Left Behind (NCLB) Act of 2001 in the United States (US Congress, 2002. Jan. 8) and the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) (Council of Europe, 2001). Both of these frameworks are designed to achieve collectivist policy goals and do so, as Foucault (1995: 184) predicted, by combining "the ceremony of power . . ., the deployment of force and the establishment of truth."

In the case of NCLB, the policy was designed to improve education for all by allowing communities to distinguish between schools where students do well on tests and schools where students perform poorly and to direct financial resources to those schools with good testing results and, over the long term, to sanction those with consistently poor results. In

the early years of the twenty-first century, NCLB came to be seen as the poster child for assessment-based intervention in language education by the Republican Party; in fact, central political control over education in the United States is neither recent nor partisan. NCLB was actually a reauthorization of the Elementary and Secondary Education Act (ESEA), which had its roots in the 1960s during the administration of Lyndon Johnson, a Democrat, and his Great Society. The first mention of language education was discussion of bilingualism in a reauthorization of the ESEA in the 1970s. Harbingers of NCLB were seen in the reauthorization debates in the early and mid-1990s. The movement to control language education began during a Democratic administration, and legislators on both sides of the aisle have long seen that being strong on education helps get votes. NCLB is part of an evolutionary trajectory toward greater state control over education, a movement that continued under the Obama administration, which continued along the path with "Race to the Top," designed to spur reforms in state and local public education.

McNamara and Roever (2006) summarized the assessment procedures involved in NCLB as statewide tests of reading/language arts and mathematics in grades 3 through 8 and at least once during high school. The subjects tested and the grades tested are mandated by NCLB, but states develop their own testing instruments. Aggregate results are reported at the school, district, and state level for the entire student populations at the different grade levels. Speakers of English as a second language, however, are one of the four groups of students whose scores are disaggregated from the population. Each school is required to make *adequate yearly progress* in all parts of the assessment: scores of its entire student body and all its disaggregated subgroups on both reading/language arts and mathematics. Evans and Hornberger (2005) argued that the consequences of NCLB for English as a second language (ESL) learners have been mixed. On the one hand, added attention has been paid by school districts to their ESL students, and small increases in funding of ESL programs have resulted because of the recognition in NCLB of ESL learners as one of the disaggregated groups, who must also show adequate yearly progress. On the other hand, because students' achievement in foreign languages is not assessed in the NCLB framework, bilingual education programs are disappearing in the push to quickly develop students' proficiency in English. In addition, Rosenbusch (2005) reported that NCLB has resulted in a decrease in instructional time for foreign languages, especially in schools with high minority populations. In effect, NCLB has contributed to the hegemony of English in schools and in US society.

The CEFR was initially designed to facilitate the recognition of language credentials across national boundaries, and the framework promulgated one particular theory of language knowledge and the establishment of a progressive set of standards. It has rapidly become institutionalized throughout Europe but, as Fulcher (2004: 260) has argued, the impact of the CEFR and the adequacy of its underlying construct are in need of debate because

> For teachers, the main danger is that they are beginning to believe that the scales in the CEFR represent an acquisitional hierarchy, rather than a common perception. They begin to believe the language of the descriptors actually relates to the sequence of how and what learners learn.

In both the American and European cases, the establishment of a particular assessment framework has had a very significant effect on teaching and learning (see Deygers, this volume).

Because language assessments like these serve the purpose of distributing scarce resources such as jobs, higher education, and financial support to those who desire them, the question of how to distribute those resources fairly is by no means academic. In recent years, the critical theory of Foucault as applied to language testing by Shohamy and Fulcher has set out quite clearly the goals both overt and covert of policy makers and the role of language tests in achieving those goals. No matter how policy is implicated in language tests, no matter whether power is exercised in the service of ends that we admire or ends that we abhor, knowledge of power and how power is exercised is liberating for testees and test designers. It is only by taking this critical perspective on language testing and by implementing proposals such as those by Shohamy (2001, 2004) for democratic assessment and by Fulcher and Davidson (2007) for effect-driven testing that those involved in language testing can assume full responsibility for tests and their uses.

Chapter summary and a look ahead

In recent years, the social dimensions of language testing have received the attention that they deserve largely thanks to the new technologies of recording human interaction that erased the circle around language, which had characterized the linguistic theories of Saussure and his predecessors. Once language is seen as greater than vocabulary and grammar, and human meaning making is seen as embodied, theories of language can be expanded to embrace how persons employ communicative resources in social interaction. A theory of the mutual constitution of language and social context permits us to revivify language and to reject a construct of language knowledge that interprets test performance in a disembodied, decontextualized context.

In practice theory, social constructs in language tests are seen as a configuration of communicative resources employed by all participants in a test, not just the testee. It is the configuration of communicative resources rather than language knowledge or strategic abilities that characterizes a test as a discursive practice, and the configuration of communicative resources is a means by which participants in a test create local identities and discursive power. Through close analysis of discourse and social interaction, a person's performance in a testing practice can be compared with practices outside the testing room. Testing practice can, however, never be removed from the global context in which powerful elites design and administer language tests and interpret test results. The political context of a newly independent Estonia and the desire of the Estonian government to establish the country's independence from its colonial ruler by language planning and testing have an effect on the lives of teachers like Ms. Muravyova. In communities around the world, the same linguistic battles are fought between local languages and the languages of colonial rulers, between standardized languages and local varieties, and between national languages and the languages of immigrants. The role of language testing in these battles is to maintain the power of speakers of prestigious language varieties and to maintain the subordinate status of speakers of non-standard varieties. It is only recently that critical language testing has made clear the relationship between language and social context, the social consequences of assessment, and the power of language tests.

What are the consequences of this newfound understanding of the social context of language testing? What debates will still be raging, and which issues will be unresolved, say, five or ten years from now? As I gaze into the crystal ball of the future of language testing, I observe three events. One is an image of two psychometricians, experts in the field of

educational measurement, sitting in front of a computer monitor scratching their heads as a waterfall of data pours down the screen. I interpret this image to mean that the attempt to measure the rich social context of language will produce so much quantitative data that new means must be developed to analyze and understand it. I also hear audio from a satellite that picks up sounds from the places over which it flies. The sounds are familiar, almost like English, but while I can understand the audio from the satellite as it flies over some countries, the audio coming from other parts of the world I can't understand, although some words sound like English. I interpret this to mean that world languages such as English will continue to spread as they are doing today in China and Korea, but local varieties will diverge further and further from what a native speaker of standard English can understand today. Finally, I see an image of a band of brown-clad brothers and sisters, guardians of a sacred code of ethics by which they wish all language testers should live. Their code is displayed in a temple of shining marble, and the brown-clad band recites it every day. They are, however, alone in their recitations, and their temple is empty.

Further reading

Davies, A. (ed.). (1997). Ethics in language testing. *Special Issue of Language Testing* 14: 3.
The articles in this special issue were presented in a symposium on the ethics of language testing held at the World Congress of Applied Linguistics in 1996. In ten articles, well-known scholars of language testing address the role of ethics (and the limits of that role) in professional activities such as language testing. The authors discuss language testing as a means of political control, the definition of the test construct, the effects of language tests on the various stakeholders who are involved, and criteria for promoting ethicality in language testing.

McNamara, T. F. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA: Blackwell.
This book focuses on the social aspects of language testing, including assessment of socially situated language use and societal consequences of language tests. The authors argue that traditional approaches to ensuring fairness in tests go some way to addressing social concerns, but a broader perspective is necessary to understand the functions of tests on a societal scale. They consider these issues in relation to language assessment in oral proficiency interviews and to the assessment of second language pragmatics. They argue that traditional approaches to ensuring social fairness in tests go some way to addressing social concerns, but a broader perspective is necessary to fully understand the social dimensions of language testing.

Shohamy, E. (2006). *Language Policy: Hidden Agendas and New Approaches*. London, UK: Routledge.
Shohamy illuminates the decisions surrounding language policy and tests and emphasizes the effects of these decisions on different groups within society. Drawing on examples from the United States, Israel, and the UK, Shohamy demonstrates different categories of language policy, from explicit use by government bodies and the media to implicit use where no active decisions are made. She also reveals and examines the mechanisms used to introduce language policy, such as propaganda and even educational material. Her critical exploration of language policy concludes with arguments for a more democratic and open approach to language policy and testing, suggesting strategies for resistance and ways to protect the linguistic rights of individuals and groups.

Young, R. F. (2009). *Discursive Practice in Language Learning and Teaching*. Malden, MA: Wiley-Blackwell.
Young sets out to explain practice theory and its implications for language learning, teaching, and testing. He examines the consequences of considering language in interaction as discursive practice and of discourse as social action. Discursive practice is the construction and reflection of social realities through language and actions that invoke identity, ideology, belief, and power. The ultimate aim

of practice theory is to explain the ways in which the global context affects the local employment of communicative resources and vice versa. In Chapters 5 and 6, Young uses practice theory to take a new look at how the employment of communicative resources in a specific discursive practice may be learned, taught, and assessed.

References

Agar, M. (1994). *Language Shock: Understanding the Culture of Conversation*. New York, NY: Morrow.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Beyer, C. (2007). Edmund Husserl. In *Stanford Encyclopedia of Philosophy*. Retrieved September 18, 2010, from plato.stanford.edu/entries/husserl.

Bourdieu, P. (1977). *Outline of a Theory of Practice* (R. Nice, trans.). Cambridge, UK: Cambridge University Press.

———. (1990). *The Logic of Practice* (R. Nice, trans.). Stanford, CA: Stanford University Press.

Bucholtz, M. and Hall, K. (2004). Language and identity. In A. Duranti (ed.), *A Companion to Linguistic Anthropology*. Malden, MA: Blackwell.

Canagarajah, S. (2009). The plurilingual tradition and the English language in South Asia. *AILA Review* 22: 5–22. https://doi.org/10.1075/aila.22.02can.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing* 20: 369–83. https://doi.org/10.1191/0265532203lt264oa.

Chalhoub-Deville, M. and Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah, NJ: Erlbaum.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge, UK: Cambridge University Press.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. New York, NY: Praeger.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.

de Certeau, M. (1984). *The Practice of Everyday Life* (S. Rendall, trans.). Berkeley, CA: University of California Press.

Erickson, F. (2004). *Talk and Social Theory: Ecologies of Speaking and Listening in Everyday Life*. Cambridge, UK: Polity.

Evans, B. A. and Hornberger, N. H. (2005). No child left behind: Repealing and unpeeling federal language education policy in the United States. *Language Policy* 4: 87–106. https://doi.org/10.1007/s10993-004-6566-2.

Ford, C. E. and Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff and S. A. Thompson (eds.), *Interaction and Grammar*. Cambridge, UK Cambridge University Press.

Foucault, M. (1978). *The History of Sexuality*, vol. 1 (R. Hurley, trans.). New York, NY: Pantheon.

———. (1995). *Discipline and Punish: The Birth of the Prison*, 2nd edn. (A. Sheridan, trans.). New York, NY: Vintage.

Foucault, M. and Gordon, C. (1980). *Power/Knowledge: Selected Interviews and Other Writings, 1972–1977* (C. Gordon *et al*., trans.). New York, NY: Pantheon.

Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly* 1: 253–66. https://doi.org/10.1207/s15434311laq0104_4.

———. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics* 29: 3–20. https://doi.org/10.1017/S0267190509090023.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment*. London, UK: Routledge.

Gerassimenko, O., Hennoste, T., Koit, M. and Raabis, A. (2004). *Other-Initiated Self-Repairs in Estonian Information Dialogues: Solving Communication Problems in Cooperation*. Paper presented at the Association for Computational Linguistics Special Interest Group Workshop on Discourse and Dialogue, Boston, MA, April 30–May 1.

Giddens, A. (1984). *The Constitution of Society: Outline of the Theory of Structuration*. Berkeley, CA: University of California Press.

Goffman, E. (1979). Footing. *Semiotica* 25: 1–29. https://doi.org/10.1515/semi.1979.25.1-2.1.

———. (1981). *Forms of Talk*. Philadelphia, PA: University of Pennsylvania Press.

Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge, UK: Cambridge University Press.

———. (1992). Contextualization and understanding. In A. Duranti and C. Goodwin (eds.), *Rethinking Context: Language as an In1teractive Phenomenon*. Cambridge, UK: Cambridge University Press.

———. (1995). Mutual inferencing in conversation. In I. Markova, C. Graumann and K. Foppa (eds.), Mutualities *in Dialogue*. Cambridge, UK: Cambridge University Press.

Halliday, M. A. K. (1994). Systemic theory. In R. E. Asher and J. M. Y. Simpson (eds.), *The Encyclopedia of Language and Linguistics*, vol. 8. Oxford, UK: Pergamon.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal* 70: 366–72. https://doi.org/10.2307/326815.

Lantolf, J. P. (2006). Re(de)fining language proficiency in light of the concept of languagculture. In H. Byrnes (ed.), *Advanced Language Learning: The contribution of Halliday and Vygotsky*. London, UK: Continuum.

Levinson, S. C. (1992). Activity types and language. In P. Drew and J. Heritage (eds.), *Talk at Work: Interaction in Institutional Settings*. Cambridge, UK: Cambridge University Press.

Levy, C. J. (2010, June 8). Estonia raises its pencils to help erase Russian. *New York Times*, A6.

Locke, J. (1690). *An Essay Concerning the True Original, Extent, and End of Civil Government*. Retrieved September 18, 2010, from http://jim.com/2ndtreat.htm.

Lowenberg, P. H. (1993). Issues of validity in tests of English as a world language: Whose standards? *World Englishes* 12: 95–106. https://doi.org/10.1111/j.1467-971X.1993.tb00011.x.

McNamara, T. F. (1997). Interaction in second language performance assessment: Whose performance? *Applied Linguistics* 18: 446–66. https://doi.org/10.1093/applin/18.4.446.

McNamara, T. F. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA: Blackwell.

Mehan, H. (1982). The structure of classroom events and their consequences for student performance. In P. Gilmore and A. A. Glatthorn (eds.), *Children in and Out of School: Ethnography and Education*. Washington, DC: Center for Applied Linguistics.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: American Council on Education and Macmillan.

———. (1996). Validity of performance assessments. In G. W. Phillips (ed.), *Technical Issues in Large-Scale Performance Assessment*. Washington, DC: US Department of Education, Office of Educational Research and Improvement.

Miyazaki, I. (1976). *China's Examination Hell: The Civil Service Examinations of Imperial China* (C. Schirokauer, trans.). New York, NY: Weatherhill.

Norris, J. M. (2008). *Validity Evaluation in Language Assessment*. New York, NY: Peter Lang.

O'Brien O'Keeffe, K. (1990). *Visible Song: Transitional Literacy in Old English Verse*. Cambridge, UK: Cambridge University Press.

Rosenbusch, M. H. (2005). The no child left behind act and teaching and learning languages in U.S. schools. *The Modern Language Journal* 89: 250–61. Retrieved from www.jstor.org/stable/3588685.

Röver, C. (2005). *Testing ESL Pragmatics: Development and Validation of a Web-Based Assessment Battery*. Frankfurt, Germany: Peter Lang.

Sacks, H., Schegloff, E. A. and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language* 50: 696–735. https://doi.org/10.1016/B978-0-12-623550-0.50008-2.

Sahlins, M. D. (1981). *Historical Metaphors and Mythical Realities: Structure in the Early History of the Sandwich Islands Kingdom*. Ann Arbor, MI: University of Michigan Press.

———. (1985). *Islands of History*. Chicago, IL: University of Chicago Press.
Saussure, F. de (1983[1916]). *Course in General Linguistics* (R. Harris, trans.). London, UK: Duckworth.
Schegloff, E. A. and Sacks, H. (1973). Opening up closings. *Semiotica* 8: 289–327. https://doi.org/10.1515/semi.1973.8.4.289.
Schleppegrell, M. J. (2004). *The Language of Schooling: A Functional Linguistics Perspective*. Mahwah, NJ: Erlbaum.
Seidlhofer, B., Breiteneder, A. and Pitzl, M.-L. (2006). English as a lingua franca in Europe: Challenges for applied linguists. *Annual Review of Applied Linguistics* 26: 3–34. https://doi.org/10.1017/S026719050600002X.
Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing* 18: 373–91. https://doi.org/10.1177/026553220101800404.
———. (2004). Assessment in multicultural societies: Applying democratic principles and practices to language testing. In B. Norton and K. Toohey (eds.), *Critical Pedagogies and Language Learning*. Cambridge, UK: Cambridge University Press.
———. (2006). *Language Policy: Hidden Agendas and New Approaches*. London, UK: Routledge.
Stimpson, G. (1946). *A Book About a Thousand Things*. New York, NY: Harper.
Trevarthen, C. (1977). Descriptive analyses of infant communicative behaviour. In H. R. Schaffer (ed.), *Studies in Mother-Infant Interaction: Proceedings of the Loch Lomond Symposium, Ross Priory, University of Strathclyde, September 1975*. London, UK: Academic Press.
———. (1979). Communication and cooperation in early infancy: A description of primary intersubjectivity. In M. Bullowa (ed.), *Before Speech*. Cambridge, UK: Cambridge University Press.
US Congress. (2002, January 8). *No Child Left Behind Act of 2001*. Public Law 107–10. United States Congress.
Young, R. F. (2008). *Language and Interaction: An Advanced Resource Book*. London, UK: Routledge.
———. (2009). *Discursive Practice in Language Learning and Teaching*. Malden, MA: Wiley-Blackwell.
———. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*, vol. 2. London, UK: Routledge.
Young, R. F. and He, A. W. (eds.). (1998). *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Amsterdam, The Netherlands: Benjamins.

5

Designing language tests for specific purposes

Carol Lynn Moder and Gene B. Halleck

Introduction

In language for specific purposes (LSP) testing, the task of the test designer is to mirror as accurately as possible the language, tasks, and contexts of the target language situation in order to accurately predict workplace language performance (McNamara, 1996; Douglas, 2000; O'Sullivan, 2012; Elder and McNamara, 2016).

For work-related LSP tests, the target language use situation is most often framed within a communicative competence model (Hymes, 1972; Canale and Swain, 1980). In this approach, appropriate language testing requires test takers to engage in extended acts of communication in which they assume social roles and perform authentic tasks common in appropriate real-world settings. In order to determine what roles and tasks must be included in the LSP test, a careful analysis of the work domain is essential. The needs analysis requires a focus on interaction in the work setting to determine what kind of exchanges typically occur, who the participants are, what the physical setting of the interaction is, what the purpose of the interaction is, and what communicative functions and linguistic forms are used. However, language testers should also be cognizant of the social and political ramifications of test design and consequent decisions. What is the larger context in which the testing is framed? Who is to be tested and why? What societal pressures have established the need for the test? What cultural norms and values are realized in the formulation of the standards? In high-stakes situations, where testing decisions have major consequences for the test takers and for the general public, understanding the socio-political underpinnings of a test situation is essential to ethical and fair test design (McNamara and Roever, 2006).

This chapter traces assessment design practices in a socio-political context for language tests that are employed in work-related contexts in which the language competency has direct effects on a large segment of society. We will outline the procedures for determining the target language to be assessed: gathering information from stakeholders, consulting work-related documents and manuals, examining discourse-based research, observing workplace task and language use, corpus generation and analysis, and piloting sample items with various stakeholders. Throughout our discussion, we will consider the ways in which

DOI: 10.4324/9781003220756-8

social and political concerns affect testing decisions. The domain in which we will illustrate these practices is aviation English. Aviation English is of particular relevance for our purposes because of the socio-political context in which the language proficiency standards emerge and the distinctiveness of the language required in the domain (see also Deygers, this volume).

The use of English in aviation is governed by the International Civil Aviation Organization (ICAO), a specialized agency of the United Nations, which creates standards and recommendations governing civil aviation "by and for" its 190 contracting member states (ICAO, n.d.). The inherently political nature of a test situation in which 190 member nations with varying interests and resources must debate and agree on a common policy that balances workplace needs with public safety makes the testing of aviation English an excellent example of the social uses of language tests.

Critical issues

Any effective approach to LSP testing must begin with a consideration of the social or political issues that have prompted the need for the test, including the ways in which language use interacts with key domain-specific goals.

In the aviation domain, English is used as a *lingua franca* – a language of communication between those who might not share the same native language. The main international use of English is for radio communication between pilots and air traffic controllers in contexts in which one of the parties does not speak the local language of air traffic control. A Spanish-speaking South American pilot flying into Spain may interact with the air traffic controllers in Madrid in Spanish, but if he flies into Paris, he will communicate with the French-speaking controllers in English. Because of the wide use of aviation English by non-native speakers, the professional use of English in this context has a prescribed phraseology with reduced syntax and vocabulary for routine actions. The highly specialized nature of the content and the prescribed syntax make the target language domain especially challenging for language testers with no aviation training. In such restricted LSP contexts, it is essential for the language tester to conduct a thorough analysis, not only of the prescribed language use, but also of the actual use in context. In language testing situations in which standards are prescribed by local, national, or international policies, the language tester must be thoroughly conversant with all requirements and with the socio-political imperatives that motivated them.

A first step is to consult relevant information provided by government agencies and by professional groups charged with the language training of the professionals in the target domain. For the aviation domain, the ICAO website and their publication, the *ICAO Journal*, provide extensive information on the standards and the motivations behind them. Of particular interest to the aviation language tester are recent efforts to improve aviation safety worldwide by raising English proficiency standards. A full discussion of the new standards appears in ICAO Document 9835, *Manual on the Implementation of ICAO Language Proficiency Requirements* (ICAO, 2004).

As these documents reveal, in 1998, in response to a number of aviation accidents in which investigators found that insufficient English language proficiency on the part of the flight crew or an air traffic controller had played a contributing role, ICAO passed an international resolution that would put in place new requirements for aviation English (Mathews, 2004: 4). ICAO member nations were charged with reviewing and strengthening the English language testing and training of air traffic control personnel and flight crews

working in airspaces where the use of English is required. In the review, public concerns about air safety had to be balanced against practical concerns about whether and how soon international aviation professionals could meet the new standards. ICAO imposed the new standards in 2003, and member nations were to implement them by 2008. The way the new standards were formulated speaks of the way in which the political process can have significant consequences for language testing (see Moder and Halleck, 2009). In order to preserve the autonomy of the member nations, ICAO did not prescribe any particular test or test design; rather, the policy defined the uniform minimum English proficiency level, Operational Level 4, in terms of a cutoff on a prescribed rating scale.

The policy engendered extensive international efforts in aviation English testing and teaching to try to meet the standards within the five-year period. The enormity of the tasks of establishing aviation English tests to implement the new policy and training tens of thousands of air traffic controllers worldwide to meet the new standard in only five years sent countries and commercial businesses scrambling. Many member nations turned to English teaching professionals, most of whom had little or no familiarity with the aviation domain. Most countries had no established aviation English programs or teachers, and the few trained in teaching aviation English had little background in workplace LSP testing. Initially, few countries engaged established LSP testers or consulted professional testing organizations like the International Language Testing Association. Some countries asked the aviation professionals charged with personnel licensing and standards to address the new English proficiency requirement. These professionals were intimately familiar with the aviation domain but generally had no background in LSP testing or teaching. At the start of the five-year policy implementation phase, very few professionals had the requisite combination of knowledge of the workplace domain and LSP testing expertise.

As Alderson (2010) has detailed, during this period, member nations adopted tests of widely varying validity and reliability, and many states did not meet the new standard by the deadline. ICAO then extended the compliance deadline to March 2011, provided that the member states not in compliance post language implementation plans to the ICAO website (ICAO, 2007). The most recent ICAO resolution on language proficiency, Resolution A38/8, approved in October 2013, suggests that some member nations continue to struggle with compliance. Alderson (2011) provides detailed information about problems with the inadequacy of tests meant to implement the ICAO standards internationally. In order to ameliorate such problems with tests implemented by various member nations, in 2011, ICAO created the Aviation English Language Test Service (AELTS). AELTS is meant to provide evaluative information on test quality to member nations by measuring proposed tests against the ICAO proficiency guidelines.

The ICAO proficiency standards detailed in Document 9835 specify that an appropriate language test must test listening and speaking in a context appropriate to aviation based on the ICAO rating scale and holistic descriptors, and it must test language use in a broader context than in the use of ICAO phraseologies alone (ICAO, 2004: 6–5). The ICAO Proficiency Rating Scale – a product of collaborative compromise between aviation professionals, applied linguists, and politicians – specifies descriptors for six levels of proficiency in each of the areas of pronunciation, structure, vocabulary, fluency, comprehension, and interaction.

The descriptors clearly reference a communicative approach to language use, highlighting the importance of interaction and the negotiation of meaning, as illustrated by descriptor statements that include phrases like “Initiates and maintains exchanges” and “Deals adequately with apparent misunderstandings by checking, confirming, or clarifying” (ICAO,

2004: A-8–9). However, the descriptors make little mention of the aviation English domain, other than requiring the ability to communicate about or comprehend "common, concrete, or work-related topics" (ICAO, 2004: A-8–9). Another noteworthy aspect of the descriptors that emerges from the socio-political context is the repeated reference to unusual or unexpected contexts. These requirements reference the societal concerns that motivated the increased standards – the need to ensure that the language of aviation professionals will be adequate to attenuate future emergency situations in which standard phraseology would not be sufficient.

By studying these documents, the aviation English tester can glean some of the essential aspects of the domain, as specified by ICAO. All aviation professionals must be able to speak and understand English, to negotiate meaning through clarification and confirmation, to employ standard phraseology effectively, and to go beyond phraseology as required in emergencies.

In order to delineate the communicative functions and language structures that define the testing domain, further study of available resources is necessary. In the aviation context, professional manuals provide needed information about the actions and events that are part of the work situation. ICAO Document 4444, *Air Traffic Manual*, establishes the training and licensing standards for air traffic controllers and pilots. It includes extensive information about the work domain and prescribes the formulaic language to accompany particular actions and requests. The manual states that established phraseologies should be used whenever possible in order to avoid confusion. If circumstances differ from the routine, aviation personnel may use "plain language," but such language should be as "clear and concise as possible" (ICAO, 2001: 12–13). An example of the phraseology prescribed in the manual is given in Example 1.

Example 1

CLIMB (*or* DESCEND)
TO (level)
REPORT REACHING (significant point)

(ICAO, 2001: 12–13)

In Example 1 we see the schematic nature of the language – action verbs in bare imperative form and key prepositions, supplemented by lexical items designating flight levels or specified locations on a route. The assumption is that an aviation professional will use the formulae, filling the slots as appropriate in all routine contexts. The manual lists over 30 pages of phraseologies, which all aviation professionals should be able to fluently adapt to the appropriate contexts. Familiarity with the appropriate use of such prescribed language is absolutely essential to the appropriate design and evaluation of language used in work-related contexts.

Recommendations for practice

Once the LSP tester has gained a thorough understanding of the domain-specific issues surrounding language, he or she must go beyond published materials to examine the actual language use situation. One key step is for the tester to consult stakeholders about key communicative functions. As Elder and McNamara (2016) have argued, investigating such "indigenous criteria" can enrich our understanding of the specific purpose situation and

help language testers move beyond more restrictive proficiency-focused criteria that seek to exclude non-linguistic cognitive factors that are essential to performance. Furthermore, in highly political contexts, such measures can provide important insights into the priorities of the various stakeholders. Aviation test designers have used varying approaches to gain information from stakeholders. Estival and Molesworth (2009) conducted a survey with pilots to determine the types of problems that they had encountered in using English with air traffic controllers. Kim and Elder (2015) used focus groups, interviews, and questionnaires to identify specific problems that arise during radiotelephony exchanges between pilots and controllers. Following a large sampling of pilots and air traffic controllers using the questionnaire, a smaller number of interview informants listened to recordings and commented on the perceived difficulties that they heard. Such issues as limited English proficiency, pronunciation, verbosity, and word choice were cited as having contributed to the misunderstandings that took place. Kim (2018) followed up this work by exploring what aviation professionals considered to be important in radiotelephony communications. Kim elicited in-depth commentary from three pilots and three air traffic controllers on an extended authentic pilot–air traffic controller interaction. The six aviation professionals noted several relevant features in the interaction, including failure to comply with radiotelephony conventions, inappropriate use of general English, and failure to accommodate to the interlocutor.

Next, we outline the procedures we used in designing and evaluating the aviation curricula and test items described in Moder and Halleck (2009). After our examination of the situations in the ICAO documents, we consulted with personnel who train air traffic controllers, asking them to list the key communicative functions that pilots and air traffic controllers must perform in various stages of flight. Based on their responses, we designed a written survey that listed the functions and asked large numbers of aviation professionals to rate their importance on a five-point scale. Compiling these ratings allowed us to construct a list of the communicative functions that the pilots and air traffic controllers considered to be most important. Given the social context for the increase in proficiency standards, we asked them to rate those functions that would be most critical in emergency situations as well.

We also interviewed aviation professionals about the language used in radiotelephony. These interviews revealed strong beliefs about the language use situation that directly affected testing practices. For example, although more proficient pilots and controllers tolerated the use of plain language instead of the more direct standard phraseologies, it was typical for those with lower general English proficiency to vehemently support the requirement that all aviation professionals use phraseology, except in highly unusual situations. Furthermore, many international air traffic controllers expressed their disdain for the way that native English-speaking pilots disregarded established phraseology, using plain language even for routine activities. In their view, this unnecessary use of plain language placed an unreasonable burden on non-native-speaking aviation professionals. They further asserted that the comprehensibility of these pilots was exacerbated by their rapid and imprecise delivery and by the use of non-standard accents.

The question of the need for native speakers of particular varieties of English to comply with established international norms is part of the broader socio-political context of the ICAO implementation. The ICAO standards apply to all aviation professionals, even native speakers of English, but most member nations where English is a standard language have not imposed the same testing requirements on native speakers. In part this is because of the focus on plain language, rather than its combination with phraseology in the workplace context. Language testers concerned with the ethical uses of tests need to be cognizant of the relation of such attitudes and practices to the fairness of the test implementation procedures.

The interview discussions of phraseology and plain language revealed another significant belief held by aviation professionals. Many aviation professionals, especially those charged with licensing standards, were adamant that non-licensed personnel were completely unqualified to rate any language sample that included the use of phraseology. They were less likely to believe that aviation professionals were not fully qualified to rate language testing samples. These strongly held views about the use of phraseology in tests that would be rated by language specialists informed extensive debates at ICAO and elsewhere about the appropriateness of work-related tasks that used phraseology in English proficiency tests. Following best practices in LSP test design, the tasks on the language test should reflect as closely as possible the actual target language situation. However, when influential aviation professionals refuse to sanction tests in which language testers rate English in a task that uses phraseology, the quality of the test can be severely compromised. The role of the tester is to make a cogent and persistent case for the need for authentic work-related language and tasks in a valid LSP test, but the concerns of powerful stakeholders must also be addressed if the test is to be successfully implemented.

Examining published manuals and other documentation can establish a clearer understanding of the scope of the target language domain. Surveys and interviews can reflect the values and beliefs of the stakeholders about the proper use of language in context. However, as with all language use in context, there is often a great variation between what is prescribed and what is actually used. In order to get a complete view of the target language situation, especially in interactional contexts, an examination of actual discourse usage is essential.

In many domains, the language tester may rely on previously conducted discourse-based research to help delineate the features of the language used in the target context. In the aviation domain, this research is limited but still revealing. A number of discourse-based studies have highlighted the socio-technical setting of aviation communication in which the co-presence of equipment and instrumentation plays a major role in the cognition and task-oriented communication of aviation professionals. They also discuss the large amount of information the participants share and the way that their discourse roles are coded in the use of specific language features (Goodwin, 1996; Linde, 1988; Nevile, 2004). Concerning the use of phraseology or plain language, studies of authentic pilot–air traffic controller communication have consistently found that plain language is not only very common, but, more importantly, they find that it also contributes significantly to the local coherence and effectiveness of the interactions (Linde, 1988; Sänne, 1999; Nevile, 2004; Wyss-Bühlmann, 2005; Hinrich, 2008). For example, one aspect of plain language use that aviation professionals often assert is not part of the domain is the use of questions, but Hinrich (2008) found them to occur frequently in pilot-controller communications (28 per hour). The questions were often used to clarify, check, or repair information – key abilities for the new ICAO standards. Such information about actual use should guide task design and may also be useful in discussions with doubtful stakeholders.

In addition to reviewing existing research, the language tester may conduct more specific targeted analyses of samples of authentic language from the target context. An example of this kind of analysis is Mell (2004), in which he uses an aviation corpus in order to delineate communicative functions and target language structures that must be included in authentic aviation English tests.

Ideally, the analysis of LSP communicative functions and tasks should also be based on observations and data recorded in the workplace. Recorded samples may be available from the stakeholders, or the tester may need to collect them. In either case, the samples should be

transcribed in order to allow a detailed examination of task and language features. Studies that have collected data from actual controller-pilot communications to use as the basis for item design include Teasdale (1996), Kim and Elder (2009), Moder and Halleck (2009), and Van Moere *et al*. (2009).

To illustrate the information that may be gained from transcripts of authentic language, we discuss two transcribed examples of work-related communication from Moder and Halleck (2009).

Example 2

CC: Alfa Juliet Tango one four two, radar service terminated, maintain flight level three-niner-zero, contact Peachtree Controller one-two-five decimal five, So long.
AJT: one-two-five decimal five, Alfa Juliet Tango one four two.

The exchange in Example 2 is a routine interaction communicated following standard phraseology. This single transcript provides extensive evidence of the communicative functions and language of aviation English. The example is a routine radiotelephony exchange between an area or en route air traffic controller and the pilot of a small plane with the call sign AJT 142 (Alfa Juliet Tango one four two). The plane is progressing along its established flight path and has reached the position where it will be turned over to the controller for the next area. The controller first identifies the aircraft that is the recipient of the communication. The use of speaker and recipient identifiers is an essential aspect of the communicative situation since a controller is in communication with multiple pilots at any given time, and all use the same frequency. Indeed, pilots must continually monitor the controller's communication with the other planes in order to ensure safety by maintaining awareness of other planes and events in the airspace. The controller informs the pilot that he is terminating his radar contact with her and provides her with the identifier (Peachtree Controller) and the radio frequency (125.5) for the next area controller. The key information in this exchange is the radio frequency, which the pilot reads back to the controller to verify that she received the information correctly. Since the routine readback is accurate, the controller does not respond further.

We see in Example 2 the use of distinctive turn-taking identifiers for speakers (the plane's call sign AJT 142), prescribed number and letter readout format ("Alfa Juliet Tango" and "three-niner-zero"), limited syntax (imperatives, participles, no articles or auxiliaries), and restricted vocabulary ("maintain," "contact," "radar service," "flight level"). The exception to standard phraseology is the signoff greeting "so long." This expression is an instance of a plain language expression used to regulate the interaction, indicating that the controller is ending the exchange with this plane. Such greetings are common in pilot-controller communication as part of the process of opening and closing interactions. The typical abbreviated length of the exchanges is also evident. The brevity of the communication is made possible by the large amount of given information that the pilot and the controller share. Each is already aware of the call sign of the plane and its intended flight path, which were previously specified in a flight plan. The radio frequencies and the navigation points along the path are also known to the pilot and controller in advance. These discourse-based observations provide a foundation for framing the interactional format of the LSP test tasks.

In Example 2, both parties hear the expected information, and there is no need for any negotiation of meaning. As in this example, most exchanges proceed as expected, but it is critical to safety that both the pilot and the controller maintain focus on assuring the

information is accurately delivered and acknowledged and on correcting any discrepancies. Indeed, this ability is specifically referenced in the ICAO guidelines for comprehension and interaction. Example 3 illustrates a situation in which more negotiation of meaning is required.

Example 3

A2220:	Airline two-two-two-zero at three-five-zero and about forty miles ahead we'll need to deviate for weather
CC:	Airline two-two-two-zero, roger contact, flight level three-five-zero, confirm what side, left or right?
A2220:	uh deviation uh left for weather in about uh probably start in about thirty miles
CC:	two-two-two-zero, Center Control, roger, left deviation approved, after deviation fly direct to Casa
A2220:	Left deviation approved afterwards cleared direct to Casa, Airline two-two-two-zero
CC:	two-two-two-zero, affirm

The pilot of Airline 2220 informs the controller that he intends to deviate from the prescribed flight path because of bad weather. Although standard phraseology exists for such a request ("request weather deviation"), the native English–speaking pilot frames the request in plain language: "we'll need to deviate for weather." He also uses plain English numbers ("forty" rather than the prescribed single digit delivery "four-zero"). These are the kinds of instances of a native speaker's use of plain language to which some aviation professionals object. In the actual exchange, the non-native-English-speaking air traffic controller acknowledges the plane and its position and requests a clarification of the intended direction of the deviation, using a combination of phraseology and a plain language question: "confirm what side, left or right?" Although weather deviations are not entirely unexpected, their occurrence is not predictable in advance. The controller must be certain that the plane's new path will not interfere with other planes in the airspace. Accordingly, we see greater negotiation between the controller and the pilot to ensure that both parties understand the proposed shift. After getting the pilot's reply, the controller identifies the plane ("2-2-2-0"), identifies himself ("Center Control"), indicates his receipt of the communication ("roger"), explicitly states his approval of the direction of the deviation ("left deviation approved"), and specifies the subsequent flight path ("after deviation fly direct to Casa"). The pilot reads back the entire clearance and ends the readback with his identifying call sign. In this case, even though the readback is correct, the controller specifically confirms it by giving the plane's call sign and using the standard phrase "affirm."

This brief examination of transcribed examples illustrates the wealth of essential information that the language tester can glean from the analysis of transcripts of authentic discourse. Such information about the communicative context, the purposes of the communication, the roles of the participants, the length and type of exchanges, the turn-taking rules, and the specific language used are all essential to the development of authentic, communicative LSP test tasks.

More extensive, targeted analyses are possible if the authentic workplace transcripts can be collected into an LSP corpus and analyzed using a concordance program (see Sinclair, 1991 for an early overview of the use of corpora in applied linguistics; see also Cushing, this volume). Moder and Halleck (2010) report on a corpus analysis of authentic pilot-controller

communication used to identify common structures necessary for work-related aviation contexts. Van Moere *et al*. (2009) also report on the use of the Oklahoma State University Aviation Corpus to construct the Versant Aviation English Test (VAET).

The Oklahoma State University Aviation Corpus includes 200,000 words of recordings of authentic pilot-controller communication. Because of the emphasis of ICAO policy on the ability of aviation professionals to use language effectively to deal with "unexpected situations," a significant portion of the corpus included transcriptions of cockpit recordings from accidents that were available on the Aviation Safety Net website (https://aviation-safety.net). These were supplemented with recordings of the actual workplace performance of practicing international air traffic controllers, as well as transcripts we obtained from other aviation researchers.

One of the ways in which such a corpus can be used to define the workplace domain is to generate the overall frequency of key lexical items and structures. Such information can supplement interviews and surveys with stakeholders in order to establish the language features that should be sampled within a specific workplace domain. For example, in the aviation corpus, the 20 most frequent words, in order starting with the most frequent, were *the*, *to*, *one*, *two*, *zero*, *you*, *three*, *five*, *and*, *of*, *four*, *seven*, *is*, *on*, *six*, *we*, *at*, *it*, *eight*, and *right*. This list differs dramatically from what one would find in a corpus of general English. For example, in the Corpus of Contemporary American English (corpus.byu.edu/coca), the most frequent 20 words are: *the, is, and*, *of*, *a*, *in*, *to*, *have*, *it*, *I*, *that*, *for*, *you*, *he*, *with*, *on*, *do*, *say*, *this*, and *they*. Notable distinctions in the top 20 words in the aviation corpus when compared to the general English list are the high frequency of numbers, the occurrence of the directional term *right*, the appearance of only three prepositions (*of*, *on*, and *at*), the appearance of only three pronouns (*you*, *it* and *we*), the absence of the article *a*, the absence of verbs, and the absence of the demonstratives *this* and *that*. These distinctions help highlight how the specialized nature of LSP contexts requires that test tasks be designed in concert with domain-specific analyses of lexicon and grammar.

Going beyond the list of the most frequent words overall, we can use the corpus to analyze the content of the domain. The most common technique for doing so is to compile a list of frequent nouns and verbs in the corpus.

The most frequent nouns in the corpus are *sound, heading, flight, captain, approach, air, tower, time, act, feet, miles, pilot, level, speed, controller, radio, officer, wind, aircraft, call, voice, number, gear, engine*, *flaps*. The frequency of these nouns gives the LSP tester an indication of the nature of the domain and the concepts most mentioned. The corpus analysis can be used to formulate the target proportions of content items in the LSP test. Such frequency data, in combination with the information from documents and stakeholders, should inform the design of the task items.

An examination of the most frequent verbs in the corpus is illustrative not only of typical actions in the workplace setting, but also of the structures commonly used in workplace language. The most frequent verbs are *is*, *are*, *have*, *turn*, *cleared*, *can*, *maintain*, *get*, *got*, *hold*, *going*, *contact*, *land*, *see*. In order to preserve structural variation, we list these words in the forms that actually appeared. The frequent actions include those typically associated with various stages of a flight: *hold*, *turn*, *maintain*, *contact*, *land*, *cleared*, and *going*. In addition, there are the verbs of possession (*have*, *get*, and *got*) and the perception verb *see*. With respect to structure, the list indicates that verbs most frequently occur in bare imperatives, with some instances of *-ed* and *-ing* participles. An analysis of the fuller context shows that these participles typically occurred without any accompanying auxiliaries. Given the prescribed absence of auxiliaries in phraseology, the high frequency of *is*, *are*, and *can* is of

interest, suggesting some deviance from the standard. Further examination of the utterances in the corpus confirms that the most common verb forms are imperatives ("Turn left whiskey, Hold short"), followed by *-ed* forms without auxiliaries ("Cleared Amsterdam via 5-1 November 0-8 Whiskey upper blue 4-0-0") and *-ing* forms without auxiliaries ("Proceeding direct to Selek, maintaining level 3-6-0"). In the corpus, auxiliaries occurred in statements with participles ("AI-182 is cleared to London) only about 20 percent of the time, and their use in these contexts did not appreciably improve the effectiveness or the flow of communication. Such corpus-based verb frequency analysis can be used as a guide to the writing and selection of lexical and structural items appropriate to the target LSP context.

In the aviation corpus, more critical occurrences of auxiliaries appear in questions, which are significant because of their frequency of use in emergency situations ("Are you dumping fuel?"). The same frequency in emergencies occurs with the modal "can" ("We're maintaining altitude with difficulty uh but uh we can maintain altitude we think"; "Airline two forty three, can you give me your souls on board and your fuel on board?"). If we return to the corpus and search for question words, we find that only three words appear (*what*, *when*, and *how*). As with the other questions, these are used most in critical situations ("What altitude are you descending to at the moment?").

Though these uses of questions and the modal "can" are considerably less frequent than other forms in routine contexts, they appear with much greater frequency in the critical emergency situations that the ICAO proficiency standards are meant to mitigate. The importance of questions in these contexts belies the assertion of many aviation professionals that "There are no questions in aviation English." The analysis of the word frequency and the specific utterance content in the corpus supports the view that in order to maintain public safety standards, aviation tests absolutely must sample these plain language structures in work-related contexts.

Based on all sources of data, a set of domains and item types should be devised that sample representative communicative functions and language in the context of authentic work-related tasks.

In most LSP tests, the design would also include the formulation of rating criteria (see McNamara, 1996 for a description of this process for an LSP context). In the aviation context, the rating scale and descriptors have been established by ICAO policy. However, it is imperative that the test designer consider the importance of training raters to use the scale. Such training is particularly important in LSP contexts when the descriptors appear to reference general proficiency criteria that actually require different standards in the target workplace. For example, Moder and Halleck (2009) discuss the problem of evaluating structure using the ICAO descriptor that references the control of "basic grammatical structures." As we have seen from the needs analysis, these basic structures must be interpreted in the context of workplace tasks; general proficiency structures, such as tense sequencing and subject-verb agreement, are irrelevant. Training raters to apply the ratings effectively requires that they become familiar with the distinctive use of language in the aviation domain.

In all LSP contexts, the tester should pilot items, but the need for such piloting is more imperative in socio-political contexts in order to establish the validity of the test constructs to stakeholders. In Moder and Halleck (2009), we reported on pilot trials of test tasks that we devised to engage the socio-political issues raised in our needs analysis.

The reluctance of aviation licensing professionals to allow LSP professionals to rate samples including phraseology combined with the discomfort that many LSP teachers felt with the aviation domain led some member nations to consider using general English proficiency

test tasks with aviation content. We wished to examine the potential consequences of testing aviation English using only plain English tasks. We administered a set of work-related radiotelephony-based tasks and general English tasks on aviation topics to 14 international air traffic controllers and compared their performance using the ICAO scale.

A sample radiotelephony question, based directly on the transcripts analyzed in the needs assessment, appears in Example 4.

Example 4: radiotelephony question

Below is an exchange between an air traffic controller at Center Area Control and the pilot of a small plane with call sign WJA 4124. The center controller hands the plane off and advises the pilot to contact Peachtree Control at the frequency 125.5. The pilot then calls in to the center controller again. Listen to the pilot and respond appropriately.

CC:	WJA 4124 radar service terminated, maintain flight level 390, contact Peachtree controller 125.5. So long.
WJA:	125.5 WJA 4124
WJA:	Center control, WJA 4124
CC:	
WJA:	Confirm the frequency is 125.5?
CC:	
WJA:	135.5 will monitor and try them again over WPP WJA 4124
CC:	
WJA:	Thanks for your help.

The test was administered by audiotape. In order to simulate the interactional context of the workplace, test takers wore headphones with microphones, similar to those used in the work setting. The goal of this item was to test the ability to listen to the information in the readback and to provide an appropriate correction as needed. Such items were included in the test because this ability was highlighted in the ICAO proficiency scales and ranked as highly important by all stakeholders. In order to avoid inappropriate memory requirements that are not part of the normal workplace context, the item included the relevant background information about the aviation situation before the listening task began. The item also provided in writing the first three utterances of the exchange to further establish the appropriate context for the interaction. The remaining communications of the pilot were only presented orally.

The general English tasks required students to listen to instructions and describe an aviation situation. The tasks included both common and less expected occurrences.

In the sample item, the test takers saw a picture of a commercial plane partially off the runway in a ditch. They heard the prompt in Example 5 through their headphones.

Example 5: general English in aviation context question

The picture below shows a plane in an unusual position. Describe where it seems to be, how you think it may have gotten there, and the condition of any passengers who may have been aboard at the time of the accident.

Because such general description tasks are not a normal part of the controller's work domain, one would expect that the air traffic controllers would perform better on the familiar radiotelephony task. The overall group results indicated that this was the case. The majority of the practicing air traffic controllers that we tested did perform better on the radiotelephony task. Thus, for most controllers, the use of general English tasks in an aviation test would be likely to underestimate their ability to adequately perform in the workplace, resulting in the potential loss of employment. Of greater concern for public safety is that some controllers performed at Operational Level 4 on one of the general description tasks but failed to demonstrate minimum proficiency on the radiotelephony task. For such test takers, the general aviation English task would have inaccurately predicted the controller's performance level on a critical workplace task.

As the needs analysis demonstrated, the communication requirements of pilot-controller communication overlap little with those found in general English, so it is not surprising that the ability to narrate or describe events within the aviation content domain does not show a direct relationship to workplace communication. However, the empirical demonstration of this discrepancy served to alert some stakeholders of the dangers that might result if some stakeholder beliefs were to unduly influence the test design.

For our study, we had the advantage of having access to willing stakeholders with whom we could pilot our test tasks. For the Versant Aviation Test (Van Moere *et al*., 2009), items were vetted by a paid advisory board of international aviation professionals, who commented on the appropriateness of language and task type. Although the trial items were carefully devised to sample a representative distribution of communicative functions and language based on the Oklahoma State University Aviation English Corpus, the vetting process required extensive negotiation. Some advisory board members commented negatively on any item that did not strictly follow standard phraseology. The applied linguists who designed the items argued strongly for maintaining the authenticity of the language represented in the sample. The resulting item bank was a compromise between these positions, including plain language in more unexpected contexts and adhering more closely to phraseology in routine exchanges.

Other socio-political complications in engaging subjects to trial test items in the development stage are reported by Huhta (2009) and Read and Knoch (2009). These researchers also caution that once the test developer's initial job is finished, the agencies that commissioned their work may not continue to oversee how the test items or the tests themselves are used. The member states or the commercial enterprises that implement the tests will be the ones that determine how ratings are done and whether or not acceptable standards will be maintained. The long-term integrity of the test design remains a significant concern for test designers working in high-stakes socio-political contexts.

Conclusions and future directions

McNamara and Roever (2006: 18) have asserted that the validity of socially mandated LSP testing depends on establishing a balance between the responsibility of the test designer to protect the test taker from unfair exclusion and his or her responsibility to the institution that seeks to use the test to assess the test taker's proficiency for a particular social purpose. Given the increasing use of language tests as instruments of social policy and the likelihood that such uses will continue to affect more societies and more test domains in the coming years, language testers must be better prepared to engage with policy makers and stakeholders in order to assure the ethical and fair use of the tests with which they are involved.

Language testers will have to develop more transparent ways of communicating the key components of test validity and fairness to stakeholders at various levels of sophistication. A greater use of discourse-based research methods is one means of making testing concepts more intelligible to non-specialists, but these will still need to be supported by appropriate psychometric analyses to meet policy makers' expectations of scientific rigor. The language tester will also need to much more effectively deploy social, psychological, and discourse considerations in constructing language tests. Two key issues that have been increasingly important in LSP testing are issues of indigenous criteria and the situationally based performance of native and non-native speakers (McNamara, 2014). Both issues are critical in aviation English testing.

In establishing the balance between responsibility to social institutions and responsibilities to test takers, the LSP tester must effectively combine knowledge of the socio-political context, dialogue with stakeholders, and best practices in workplace needs analysis and task design. He or she must also be prepared to exercise considerable patience and persistence in articulating the potential consequences of specific test design choices. When the political stakes are high, the possibilities for both social harm and social benefit are great. For all specific- purpose uses of tests, the more informed the tester, the more likely the positive outcome.

Further reading

Bowles, H. (2012). Analyzing languages for specific purposes discourse. *The Modern Language Journal* 96: 43–58.

This article provides an overview of current methods of discourse-based research in LSP.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge, UK: Cambridge University Press.

This book provides a comprehensive discussion of the issues related to LSP assessment. The author evaluates high-stakes specific-purpose tests by investigating target language use domains and combines insights from applied linguistics with issues related to relevant social/psychological constructs. Specific-purpose tests of listening and speaking, such as the Occupational English Test, the Japanese Language Test for Tour Guides, IELTS, and the Professional Test in English Language for Air Traffic Controllers (PELA) are discussed.

ICAO www.icao.int/icao/en/jr/2004.

This link provides online access to English, French, and Spanish versions of the *ICAO Journal*, an official ICAO publication that includes articles by aviation professionals on issues of current interest. The 2004 Volume 59 Number 1 issue provides background on the ICAO proficiency requirements. The 2008 Volume 63 Number 1 issue gives an update on policy related to member nations' efforts to meet targeted levels.

Knoch, U. and Macqueen, S. (2020). *Assessing English for Professional Purposes*. Oxon, UK: Routledge.

This recent book provides an overview of language assessment for professional purposes (LAPP), touching on healthcare, aviation, and legal contexts, among others.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA: Wiley-Blackwell.

This volume provides a broad-ranging discussion of the need to integrate socio-political considerations into approaches to language testing, with specific reference to test validity, design, and use. Tests of oral proficiency and pragmatics and those used in educational contexts are discussed in detail.

O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. *The Modern Language Journal* 96: 71–88.

This article, part of a focus issue on LSP, gives a thorough history of LSP testing, highlighting the benefits of communication between scholars assessing a variety of languages.

www.aero-lingo.com.

This website contains extensive information on the language of air traffic control. It includes a list of 50 airplane crashes connected to language difficulties and related articles. In addition, it provides information about air traffic control phraseology and a useful link to ICAO Doc 9835 (*Manual on the Implementation of ICAO Language Proficiency Requirements*), as well as links to audio files of speech samples of pilots and controllers who have been rated at ICAO Levels 3, 4, and 5.

References

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing* 27: 51–72. https://doi.org/10.1177/0265532209347196.

———. (2011). The politics of aviation English testing. *Language Assessment Quarterly* 8: 386–403. https://doi.org/10.1080/15434303.2011.622017.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1: 1–47.

Corpus of Contemporary American English (COCA). (n.d.). Retrieved from www. americancorpus.org.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge, UK: Cambridge University Press.

Elder, C. and McNamara, T. (2016). The hunt for "indigenous criteria" in assessing communication in the physiotherapy workplace. *Language Testing* 33: 153–74. https://doi.org/10.1177/0265532215607398.

Estival, D. and Molesworth, B. (2009). A study of EL2 pilots radio communication in the general aviation environment. *Australian Review of Applied Linguistics* 32: 24.1–24.16. https://doi.org/10.2104/aral0924.

Goodwin, M. (1996). Informings and announcements in their environment: Prosody within a multiactivity work setting. In E. Cooper-Kuhlen and M. Selting (eds.), *Prosody in Conversation: Interactional Studies*. Cambridge, UK: Cambridge University Press, 436–61.

Hinrich, S. W. (2008). *The Use of Questions in International Pilot and Air Traffic Controller Communication*. Unpublished doctoral dissertation, Oklahoma State University, Stillwater, OK, USA.

Huhta, A. (2009). An analysis of the quality of English testing for aviation purposes in Finland. *Australian Review of Applied Linguistics* 32: 26.1–26.14. https://doi.org/10.2104/aral0926.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride and J. Holmes (eds.), *Sociolinguistics*. Harmondsworth, UK: Penguin, 269–93.

International Civil Aviation Organization (ICAO). (n.d.). *Strategic Objectives of ICAO*. Retrieved September 6, 2009, from www.icao.int/icao/ en/strategic_objectives.htm.

———. (2001). *Air Traffic Management Document 4444. ATM/501*. Montreal, Canada: ICAO.

———. (2004). *Manual on the Implementation of ICAO Language Proficiency Requirements Document 9835AN453*. Montreal, Canada: ICAO.

———. (2007). *Working Paper Assembly 36th Session A36-WP/151; TE/36.*

———. (2013). Retrieved from www.icao.int/safety/lpr/Pages/Language-Proficiency-Requirements.aspx.

Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? *Language Testing* 35: 405–26. https://doi.org/10.1177/0265532218758127.

Kim, H. and Elder, C. (2009). Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel. *Australian Review of Applied Linguistics* 32: 23.1–17. https://doi.org/10.2104/aral0923.

———. (2015). Interrogating the construct of aviation English: Feedback from test-takers in Korea. *Language Testing* 32: 129–49. https://doi.org/10.1177/0265532214544394.

Linde, C. (1988). The quantitative study of communicative success: Politeness and accidents in aviation discourse. *Language in Society* 17: 375–99. Retrieved from www.jstor.org/stable/4167952.

Mathews, E. (2004). New provisions for English language proficiency are expected to improve aviation safety. *ICAO Journal* 59: 4–6.

McNamara, T. (1996). *Measuring Second Language Performance*. London, UK: Longman.

———. (2014). Epilogue: Thirty years on: Evolution or revolution? *Language Assessment Quarterly* 11: 226–32. https://doi.org/10.1080/15434303.2014.895830.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA: Wiley-Blackwell.

Mell, J. (2004). Language training and testing in aviation needs to focus on job-specific competencies. *ICAO Journal* 59: 12–14.

Moder, C. L. and Halleck, G. B. (2009). Planes, politics and oral proficiency: Testing international air traffic controllers. *Australian Review of Applied Linguistics* 32: 25.1–25.16. https://doi.org/10.2104/aral0925.

———. (2010). *Can We Get a Little Higher? Proficiency Levels in Aviation English*. Presentation at the Language Testing Research Colloquium, Cambridge, UK.

Nevile, M. (2004). *Beyond the Black Box: Talk-in Interaction in the Airline Cockpit*. Aldershot, UK: Ashgate.

O'Sullivan, B. (2012). Assessment issues in Languages for specific purposes. *The Modern Language Journal* 96: 71–88. https://doi.org/10.1111/j.1540-4781.2012.01298.x.

Read, J. and Knoch, U. (2009). Clearing the air: Applied linguistic perspectives on aviation communication. *Australian Review of Applied Linguistics* 32: 21.1–21.11. https://doi.org/10.2104/aral0921.

Sänne, J. M. (1999). *Creating Safety in Air Traffic Control*. Lund, Sweden: Arkiv Forlag.

Sinclair, J. (1991). *Corpus Concordance Collocation*. Oxford, UK: Oxford University Press.

Teasdale, A. (1996). Content validity in tests for well-defined LSP domains: An approach to defining what is to be tested. In M. Milanovich and N. Saville (eds.), *Performance Testing Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium Cambridge and Arnhem*. Cambridge, UK: Cambridge University Press, 211–30.

Van Moere, A., Suzuki, M., Downey, R. and Cheng, J. (2009). Implementing ICAO language proficiency requirements in the versant aviation English test. *Australian Review of Applied Linguistics* 32: 27.1–27.17. https://doi.org/10.2104/aral0927.

Wyss-Bühlmann, E. (2005). *Variation and Co-operative Communication Strategies in Air Traffic Control English*. Bern, Switzerland: Peter Lang.

6

Revisiting language assessment for immigration and citizenship

The case of US citizenship and the Naturalization Test

Antony John Kunnan

Introduction

In the twenty-first century, migration around the world has continued unabated, much like in the previous century. Typically, there are two types of immigrants: *voluntary immigrants* and *involuntary immigrants*. The former group are people who voluntarily move from their home country to another in search of better work opportunities, for temporary residence based on work or study, and for family reunification. These immigrants may feel little or none of the wrenching separation from their home country, family, and culture mainly due to the digital age we live in today (with email, the internet, global news, and international video communication). In addition, the ease of international travel has made it possible for immigrants to be less isolated from their homeland and family. In this chapter, I will focus on voluntary immigrants and, specifically, on American immigration and immigration and citizenship policy and practice.

Today's voluntary immigrants are different even from the migrants of three decades ago. Immigrants of the past may have felt that their citizenship and identity were intertwined and that they had to develop a new identity in their new country. Many immigrants today need not give up their home country traditions, languages, and culture in the way they might have decades ago; in fact, the languages, cultures, traditions, and cuisines of their homeland are often available in their receiving countries. They may have dual or multiple identities and dual or multiple nationalities or citizenships. Thus, we can say that these immigrants (who will later become citizenship applicants) are *transnational*, maintaining familial, social, political, economic, linguistic, and cultural connections across international borders. This change in the traditional concept of citizenship mandates a few key questions. What kind of citizenship or naturalization process would be ideal for these transnational immigrants? Should there be a rethink in immigrant-receiving countries regarding the identity and aspirations of newcomers and residents and potential citizenship applicants?

Voluntary immigrants could be further classified into two types: those who plan to stay in the receiving country for the *short term* (typically, tourists, students, etc.) and those who

 DOI: 10.4324/9781003220756-9

plan to stay or are staying for the *long term* (typically, those on work permits, those joining their spouses or families, etc.). Except for tourists in the former group, most immigrants will need to demonstrate relevant abilities (which may include language abilities) related to their study or employment before they are issued visas to enter the new country (among other requirements). In some receiving countries, there are requirements such as a minimum level of language ability in the standard variety of the dominant language that must be met even for short-term stays. Then there are immigrants who plan to stay for the long term and who, after years in the receiving country (typically three to five years), may consider applying for citizenship. It is these applicants who are required to take language tests, history and civics tests, and/or social integration courses before they can be granted citizenship. This process in the US is called naturalization.

Language policy

Modern nation-states, including liberal democracies, have debated both the state-mandated national language requirements as part of political and social integration policies and the role of minority or indigenous languages. Most countries embrace the idea of a public monolingualism (in terms of a single national language: for example, the US, France, China, and Japan, although there are other languages spoken in each of these countries) while a few countries promote a limited public multilingualism (at least in terms of national languages: for example, two official languages in India and Canada and three or more in Belgium, Singapore, and South Africa).

The evangelical desire for public homogeneity, which includes public monolingualism and social cohesion, was best argued in the US by assimilationist writer Schlesinger (1992). He wrote in *The Disuniting of America* that "the cult of ethnicity"

> reverses the historic theory of America as one people – the theory that has thus far managed to keep American society whole. . . . The national ideal had once been *e pluribus unum* [out of many, one]. Are we now to belittle *unum* and glorify *pluribus*? Will the Centre hold? Or will the melting pot give way to the Tower of Babel.
>
> *(pp. 16–18)*

Many political thinkers have argued for a common language or public monolingualism (see Huntington, 2005; Schlesinger, 1992) because they believe a bilingual or multilingual country could result in ghettoization and social immobility (Pogge, 2003). On the other hand, others have stressed the importance of multiculturalism and individual language rights (Kymlicka, 1995) and tolerance-oriented language rights (May, 2005, 2008). In fact, multilingualism is a common practice in the world. This practice can be seen in a few countries with multilingual populations such as India (with two official languages), Singapore (with four), South Africa (with eleven), and Switzerland (with three).

Public multiculturalism has its ideological supporters in Kymlicka (1995) and May (2005). Kymlicka argued for individual rights as "group-differentiated rights": "Granting special representation rights, land claims, or language rights to a minority . . . can be seen as putting the various groups on a more equal footing, by reducing the extent to which the smaller group is vulnerable to the larger" (p. 24, as cited in May, 2008). Applying Kymlicka's formulation, May (2008) put it succinctly, "the preoccupation of modern nation-state organization with a single language and culture, and an allied public monolingualism, is

both unnecessarily unjust to and exclusive of minority language groups" (p. 24). He further argued along the lines of Kymlicka, by giving examples for this idea from the Welsh in Britain, the Catalans and Basques in Spain, the Bretons in France, the Québécois in Canada, and some Latino groups (e.g., Puerto Ricans) in the USA. Despite these strong arguments for multilingualism, however, language tests for immigration and citizenship have continued to support ostensibly monolingual language policies. Furthermore, such tests have proliferated in the late twentieth and early twenty-first centuries. In a survey we conducted on language requirements for immigration, citizenship, and asylum in 62 countries, we found that 12 countries required language proficiency for immigration and 33 for citizenship, and no countries had requirements for asylum seekers (Kunnan *et al*., 2019).

Monolingual language policy in the US

The monolingual bent in the United States resulted in the English First movement in the 1930s. It was an Americanization effort that included English language programs for foreign-born adults. These efforts were also promoted by American employers such as Ford, US Steel, and Pennsylvania Railroad. According to Crawford (1992), employers believed that low proficiency in English led workers to socialist propaganda and prevented them from believing in free enterprise. Pavlenko (2002) observed that

> these efforts had solidified the link between English and patriotism in the public consciousness so well that twenty years later Philadelphia's *Evening News* still argued that all aliens are to be taught the minimum of English necessary to guarantee a belief in democracy.
>
> *(p. 180)*

Alongside these efforts came the positioning of English with superiority and patriotism. In the words of Pavlenko (2002), English was supposed to have high moral and intellectual values, and the lack of English proficiency was equated with inferior intelligence and low moral standards. A decade before the English First movement, US presidents Wilson and Roosevelt clearly associated knowledge of English with loyalty to country:

> *Wilson:* You cannot become thorough Americans if you think of yourselves in groups. America does not consist of groups. A man who thinks of himself as belonging to a particular national group in America has not yet become an American.
>
> *Roosevelt:* We have room for but one language here, and that is the English language, for we intend to see that the crucible turns our people out as Americans, of American nationality, and not as dwellers in a polyglot boarding house and we have room for but one sole loyalty, and that is the loyalty to the American people.
>
> *(1919; cited in Pavlenko, 2002: 183–84)*

The views of the presidents reflected long historical antecedents. The early decades of the republic were dominated by assimilationist thinking with ideas such as making English the sole official language, limiting access to voting and civil rights to non-English speakers, and opposing any form of bilingual education. Ricento and Wright (2008) reported that during this time, indigenous Native American languages and cultures were stigmatized. In the mid-nineteenth century, states also passed anti-literacy and compulsory ignorance laws that made it a crime to teach slaves to read or write until they were considered unjust and

illegal. In terms of language education, by 1923, 22 states had laws prohibiting the teaching of foreign languages in primary schools in a country that was made of large numbers of first-generation immigrants. Courts intervened and found these laws unconstitutional: the two most important cases were *Meyer v. Nebraska* (1923), in which the US Supreme Court ruled that a previous Nebraska law that forbade teaching in any language other than English to be unconstitutional, and *Farrington v. Tokushige* (1927), in which the US Supreme Court ruled that Hawai'i's effort to abolish private Japanese, Korean, and Chinese language schools was unconstitutional. Later in the 1950s, many states imposed English as the sole language of instruction in schools and passed laws banning the teaching of other languages. Thus, there was coercion to learn English through legislative actions, court rulings and political and societal attitudes in the US (see Kunnan, 2009a). In *Lau v. Nichols* (1974), the US Supreme Court ruled unanimously that the lack of supplemental language instruction in public school for students with limited English proficiency violated the Civil Rights Act of 1964.

More recent examples of the English-only movement in general include the following: in 1981, English was declared the official language in Virginia, and in 2000, Arizona's Proposition 203, also known as English for the Children, was passed by 63 percent of voters. The law limited the type of instruction available to English language learner students. By 1991, 18 states had enacted laws designating English as the official state language. Today, 30 states have established English as the only official language, while Hawai'i recognizes both English and Hawaiian as official, and Alaska has made some 20 native languages official, along with English.

These examples show that US lawmakers have fought hard to embrace language and cultural homogeneity, and federal courts have had to hold them back. This is despite the country's population becoming more multilingual. According to the US Census Bureau (2012), 85 percent of the foreign-born population spoke a language other than English at home. Gambino *et al*. (2014) reported that this group increased "from 14.1 million in 1980 to 40.0 million in 2010. In 2012, the foreign born numbered 40.8 million, including 40.6 million who were five years and older" (p. 2).

US policies toward immigration and citizenship

In the last 245 years in the US, the concept of naturalization and citizenship in terms of ideology, legislation, and court rulings has swung like a pendulum, sometimes favoring inclusion and tolerance of immigrants and citizenship, sometimes exclusion, racism, and self-righteousness. Such attitudes were encapsulated in the words of Emma Lazarus on the pedestal of the Statue of Liberty: "Give me your tired, your poor, your huddled masses yearning to breathe free" (cited by Kennedy, 1964: 45) to the other extreme of xenophobia – indefensible racial intolerance, bigotry, fear, and hatred against mainly the "inferior races" of Europe and all non-white immigrants.[1]

Permitting and restricting citizenship

After the American revolutionaries gave up their allegiance to the British monarch in 1776, the idea of citizenship emerged "grounded in contract, consent, and volition . . . [but] not everyone in the native-born population was deemed a citizen or granted the rights of one. Native Americans and free Black people could be classified as 'aliens' or mere 'inhabitants'" (Parker, 2015: 80). George Washington, the first president of the US (1789–1797), emphatically declared in 1783 that "the bosom of America was open to receive not only

the opulent and respectable stranger, but the oppressed and persecuted of all nations and religions."

The first federal regulation that followed was the Naturalization Act of 1790, which stated that

> any alien, who was a free white person and who had resided within the limits and under the jurisdiction of the United States for the term of two years may be admitted to become a citizen on application to any common law court of record, in any one of the States where he had resided for the term of one year at least, and making proof to the satisfaction of such court that he is a person of good character, and taking the oath or affirmation prescribed by law to support the Constitution of the United States.
>
> *(U.S. Statutes at Large, vol. 1, 1st Congress/2nd Session/Chapter 3, 1790)*

Although Washington's statement came a few years before the Naturalization Act of 1790, the founders were untroubled by the presence of non-white persons as citizenship was restricted to "free white persons." This original white supremacist law was used to exclude free white women, Black people, and Native Americans, in contrast to the much-quoted line from the Declaration of Independence: "We hold these truths to be self-evident, that all men are created equal."

Several new acts were passed and repealed regarding citizenship over the centuries, particularly regarding non-white immigrants. In 1795, Congress repealed the 1790 Act, raised the residence requirement for citizenship to 5 years, and required a declaration of intention to seek citizenship at least three years before naturalization. In 1798, the residency requirement for citizenship was raised to 14 years; in 1802, it was repealed. During the American antebellum period (1812–1861), there was westward expansion, mass migration, and extremely piecemeal and limited rights for non-white people.

In 1855, Congress passed a law that provided alien women (i.e., women not born in the US) to become naturalized citizens upon marriage to a husband who was a US citizen. While free white men and women were granted citizenship rights, Black people, Native Americans, Mexicans, and Asians were still considered alien, foreign, or inhabitants but not citizens. Two years later, the US Supreme Court infamously ruled in the *Dred Scott v. Sandford* case that the Constitution was not meant to include citizenship for Black people, whether they were slaves or free. Thus, the rights and privileges that the Constitution conferred upon American citizens could not apply to them.[2] Thus, they had to live through the daily insults of slavery, the hurts of segregation, the savagery of lynching, and the lack of belonging.

With the passing of the 13th Amendment to the US Constitution, slavery was abolished, but Black people were recognized as citizens only in the 14th Amendment in 1868. The 14th Amendment granted citizenship to all persons "born or naturalized in the United States," including formerly enslaved people, and provided all citizens with "equal protection under the laws." Yet a little more than a decade after expounding the 14th Amendment's guarantees of equal protection, due process, and consent, Congress reverted to a more restrictive immigration policy. Responding to a national xenophobic clamor for the exclusion of Chinese immigration, Congress passed the Chinese Exclusion Act of 1882. This act included the suspension of immigration of Chinese laborers (merchants, teachers, students, and tourists were exempt) to the US for ten years and prohibition of Chinese immigrants becoming citizens through the naturalization process.[3] In the next few years, vigilantes drove out

Chinese residents mainly from California, Oregon, and Washington through harassment, intimidation, arson, bombing, assault, and murder (see Lew-Williams, 2018, for details).

Finally, in 1924, almost as an afterthought, Congress enacted the Indian Citizenship Act, which granted citizenship to all Native Americans born in the US after decades of caricaturing them as savages needing to be civilized. The Immigration Act of 1924 also brought new restrictions on immigration through a numerical limit of 155,000 per year. According to Lee (2019), the quota for western Europeans was 87 percent and for Eastern and Southern Europe, 11 percent, hardening the lines between whites and non-whites.

In the West, during the economic stresses of the 1930s, Mexican and Mexican Americans bore the brunt of mass deportation and repatriation. According to Lee-Williams (2019), "families that included naturalized US citizens and US-born citizens faced just as much pressure to repatriate as noncitizens families" (p. 176). "[B]y the end of the Great Depression, the US had turned into a 'deportation nation' that made no distinction between foreign born and the undocumented . . . and lawful resident and naturalized citizens, and American born citizens of Mexican descent" (p. 180).

In February 1942, three months after the Japanese military attacked Pearl Harbor in Hawai'i in December 1941, President Roosevelt signed Executive Order 6066, authorizing the secretary of war to prescribe certain areas as military zones. This cleared the way for the incarceration of Japanese Americans, German Americans, and Italian Americans in US concentration camps. But far more Americans of Asian descent were forcefully interned than Americans of European descent, both in total and as a share of relative population. About 120,000 Japanese Americans were interned in ten relocation centers in California, Arizona, Colorado, Wyoming, Idaho, Utah, and Arkansas. In December 1944, the US Supreme Court ruled that while in general, the incarceration of Japanese Americans was constitutional, that loyal citizens of the US, regardless of cultural descent, could not be detained without cause. This ruling helped the release of all those interned.

Despite these actions, the US also welcomed Jewish immigrants from 1880s onward, and in the twentieth century, it received immigrants from South Korea in the 1960s, Vietnamese refugees in the 1970s, and the Mariel boatlift refugees from Cuba and Central American immigrants in the 1980s. In addition, the US government granted amnesty to all immigrants who had overstayed in 1986 and the offer of legal permanent residency to all Chinese residents in the US after the post–Tiananmen Square protests in 1989. Thus, the pendulum swings have continued.

The Immigration and Nationality Act (1952)

In 1952, the Immigration and Nationality Act (INA) enshrined both the English language and the history and government requirements for citizenship. The Act itself was controversial in many aspects: it was too conservative to some and more restrictive to others. The Act was vetoed by President Truman. His veto was overridden by a vote of 278 to 113 in the House and 57 to 26 in the Senate. In the words of del Valle (2003), "in passing the English literacy provision, Congressmen clearly linked the inability to speak or understand English to political suspicion" (p. 93). The Act also deemed that aliens (as all non-residents are referred to in the US) who were anarchists or members of or affiliated to the Communist party or other totalitarian organizations were inadmissible or deportable.[4] Such aliens who advocated, taught, wrote, or published in support of communism, a totalitarian dictatorship, or the overthrow of the US were also deemed deportable aliens.

The US government enforced these requirements for naturalization in different ways over time. Starting in 1952, the requirements were informally enforced by immigration examiners. The idea of a standardized test was only raised when aliens seeking citizenship through the Immigration Reform and Control Act of 1986 were required to meet educational requirements.[5]

Educational requirements

The educational requirements stated in the INA of 1952 are as follows:

> No person except as otherwise provided in this title shall hereafter be naturalized as a citizen of the United States upon his own petition who cannot demonstrate:
>
> (1) an understanding of the English language, including an ability to read, write, and speak words in ordinary usage in the English language: Provided, That this requirement shall not apply to any person physically unable to comply with therewith, if otherwise qualified to be naturalized, or to any person who, on the effective date of this Act, is over fifty years of age and has been living in the United States for periods totally at least twenty years: Provided further, That the requirements of this section relating to ability to read and write shall be met if the applicant can read or write simple words and phrases to the end that a reasonable test of his literacy shall be made and that no extraordinary or unreasonable condition shall be imposed upon the applicant; and
> (2) a knowledge and understanding of the fundamentals of the history, and of the principles and form of government, of the U.S. (INA, §312, 8 U.S.C. 1423).[6]

In summary, from the 1790 Act to the 1952 Act, Congress granted citizenship while at the same time permitting immigration as well as restricting immigrants on selective racial and ethnic bases, depending on the decade. But one of the key requirements that remained steady in the last century was the citizenship applicants' English language proficiency and knowledge of history and government. This requirement started with the literacy test in the nineteenth century and then developed into a more standardized test in late twentieth century.

The US Naturalization Test

The literacy test

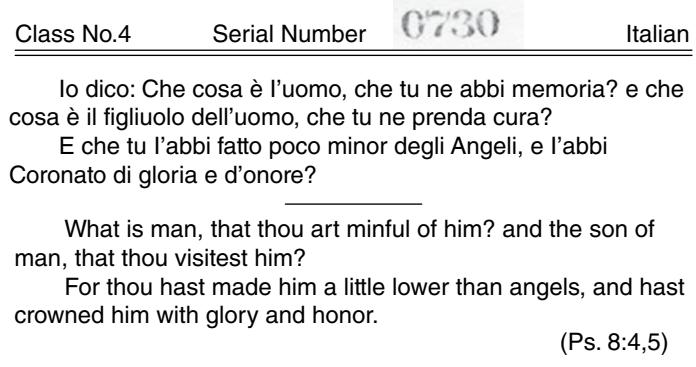
In 1906, the Bureau of Immigration and Naturalization was established, and standardized naturalization procedures were introduced. The Bureau changed the process from an informal approach conducted by county judges in each state. During these years, the Bureau sought and received correspondence from a variety of groups (judges, potential citizens, and immigrant aid groups) to determine the content of the process. These groups were concerned about potential citizens being denied citizenship for lack of knowledge about civics and US history. In response, the Bureau took measures to develop an educational program to help potential citizens learn about civics and history, although no standardized test or testing

procedures were set up. Soon, naturalization examiners, instead of judges, began asking the civics questions orally, with examiners choosing their own questions.

Pickus (2005), citing Thompson (1920), in *Schooling of the Immigrant*, explains that examiners considered the following questions:

Can the Constitution be changed?
Can Congress pass any law in regard to religion?
Can a citizen be deprived of life, liberty, or property?
What is the form of our government? Who rules this country?
What is a republic? What is a monarchy?

Language criteria were added for the first time in the 1893 Act, when Congress added *the ability to read and write* (but not specifically in English), and in the 1906 Act, Congress required applicants *to sign their petitions in their own handwriting and speak English*. In 1896, Congress passed the Literacy Test Bill, promoted by the Immigration Restriction League,[7] which wanted to have a barrier in place to restrict immigrants from the "inferior" races of Europe (immigrants from Eastern and Southern Europe). The test was expected to assess the ability of applicants for immigration to read at least 40 words in any language as a requirement for admission to the United States, but this bill was vetoed by President Cleveland. In 1917, as wartime hysteria fed American xenophobia, Congress passed the Literacy Test Bill again, overriding President Wilson's veto this time. This instrument of racial exclusion became the first literacy requirement that required potential citizens and immigrants over 16 years of age to read at least 30 words and not more than 80 words in ordinary use in any language. Thus, from 1917 onward, the test resulted in effectively restricting immigrants from Eastern and Southern Europe, Irish Catholics, and Asians.[8] Figure 6.1 shows a card that Italian immigrants had to read to the immigration officer on arrival. Similar cards were prepared for many languages, including Armeno-Turkish, French, Hebrew, Hindustani, Polish, etc. The literacy test was a precursor to the formal education requirements that were approved in 1952 and finally resulted in the naturalization test.


Class No.4 Serial Number 6730 Italian

Io dico: Che cosa è l'uomo, che tu ne abbi memoria? e che cosa è il figliuolo dell'uomo, che tu ne prenda cura?

E che tu l'abbi fatto poco minor degli Angeli, e l'abbi Coronato di gloria e d'onore?

What is man, that thou art minful of him? and the son of man, that thou visitest him?

For thou hast made him a little lower than angels, and hast crowned him with glory and honor.

(Ps. 8:4,5)

Figure 6.1 An example of the literacy test in Italian

Source: Ellis Island National Park Services, 2019. Retrieved from: www.nps.gov/elis/learn/historyculture/history-of-ellis-island-from-1892-to-1954.htm

The naturalization test

The first US Naturalization Test (USNT) was developed and implemented in 1998. It had three sections for the English language part (speaking, reading, and writing) and a history and government test (often referred to as the civic test). A recent revision of the test in 2007–2008 was reported in Kunnan (2009a). The redesigned test has not been changed in terms of test content or format.

Test content and format

The content and response format of the redesigned test is as follows:

(1) Applicants answer questions in English asked about their application (N-400) during the interview; this will form the (listening) and *speaking* test.
(2) Applicants will have three chances to *read* a sentence (aloud)
(3) Applicants will have three chances to *write* a sentence. [Vocabulary lists for reading, writing, and speaking tasks are available for study].
(4) Applicants will have to answer questions orally that will cover *U.S. history and civics*.

The new guidelines for the English test and the civics test are as follows:

Speaking: An applicant's (listening and) speaking skills are determined by the applicant's answers to questions asked by an examiner during the naturalization eligibility interview. Examiners are required to repeat and rephrase questions until the officer is satisfied that the applicant either fully understands the question or does not understand English. If the applicant generally understands and can respond meaningfully to questions relevant to the determination of eligibility, the applicant has demonstrated the ability to speak English. The test is administered by an examiner orally and individually and it could start with questions regarding the N-400 form and everyday matters (such as "Where do you live?" "How did you come to this office today?" etc.). No clear pass or fail criteria are listed for this section.

Reading: An applicant must read one sentence out of three sentences to sufficiently demonstrate the ability to read in English. The examiner should be able to determine that the applicant appears to understand the meaning of the sentence. Once the applicant reads one of three sentences correctly, the examiner will stop administering the reading test. Applicants will not be failed because of their accent when speaking English. The examiner will score the reading as follows:

Pass: Reads one sentence without extended pauses; reads all content words but may omit short words that do not interfere with meaning; may make pronunciation or intonation errors that do not interfere with meaning.

Fail: Does not read the sentence; omits a content word or substitutes another word for a content word; pauses for extended periods of time while reading the sentence; makes pronunciation or intonation errors that interfere with meaning.

Writing: An applicant must write one sentence (on a piece of paper) out of three sentences to sufficiently demonstrate the ability to write in English in a manner that

would be understandable to the examiner. An applicant must not abbreviate any dictated word in the written sentence. Once the applicant writes one of three sentences correctly, the examiner will stop administering the writing test. An applicant will not be failed because of spelling, capitalization, or punctuation errors unless the errors would prevent understanding the meaning of the sentence. The examiner will score the writing as follows:

> *Pass:* Has the same general meaning as the dictated sentence; may contain some grammatical, spelling, punctuation, or capitalization errors that do not interfere with meaning; may omit short words that do not interfere with meaning; numbers may be spelled out or written as digits.
>
> Fail: Writes nothing or only one or two isolated words; is completely illegible; writes a different sentence or words; written sentence does not communicate the meaning of the dictated sentence; writes an abbreviation for a dictated word.

Civics: Applicants will be asked at least 10 questions based on the 100 questions on U.S. history and civics list. Questions are related to American government and history are under the sections: *American government* (principles of American democracy, the system of government, rights, and responsibilities), *American history* (colonial period and independence, 1800s, recent American history and other historical information), *Symbols and holidays* (symbols, holidays).

> *Pass* and *Fail*: Applicants need to answer at least six questions correctly to pass.
>
> *(USCIS Scoring guidelines)*[9]

Test administration

The test is administered by an examiner orally and individually face to face with the applicant. For the writing test, applicants are required to handwrite a sentence on paper. No computer or tablet is used in the test. The order of the test could be different, but the guidelines for the content and procedure are the same for all applicants. The whole process is expected to take about 20 minutes, which includes both the test and the process of checking applications for eligibility for citizenship. Applicants have two opportunities to take both tests per application. If they fail any portion of the test during the first interview, they will be retested on the failed portion of the test between 60 and 90 days from the date of the first interview. If they fail the tests twice, they will have to start the application process all over again.

Is the US Naturalization Test beneficial?

One defensible way to examine and judge whether the USNT is beneficial (or if it is a barrier) to applicants is to use a framework that accounts for the principle of fairness of the assessment and the principle of justice of the assessment institution (see Kunnan, 2018, for detailed explication of these two principles; see also Walters, this volume). Briefly, the fairness principle is composed of the following sub-principles: (1) opportunity to learn, (2) meaningfulness, (3) absence of bias, (4) access and accommodations, and (5) consequences. The justice principle is composed of the following sub-principles: (1) reasoning or justifying the assessment and (2) advancing justice. I will take up each one of these sub-principles and examine how the USNT fares on each of them.

Fairness 1: opportunity to learn

Question: Do applicants have an opportunity to learn for the test? The evidence in support of the question is as follows:

Websites: The US Citizenship and Immigration Services (USCIS) websites provide guidance that is meant to be useful for the English test. Vocabulary practice for the reading and writing tests and topics for the speaking test covered on Form N-400, application for naturalization, and the interview are offered. For the civics test, topics such as American history and government and integrated civics with 100 questions and answers are available in Arabic, Chinese, English, Korean, Spanish, Tagalog, and Vietnamese. In addition, large print text-only formats, as well as in the 65/20 Exemption form with only 20 questions, are available.

Community college courses: Applicants can attend community college courses to prepare for the English and civics tests with a fee of up to to $35 per semester in some community colleges or for free at advocacy or community centers (such as Catholic Charities, Asian Americans Advancing Justice, etc.) that have received USCIS grants. These measures increase the opportunity for success.

Pass Rates: The pass rate was about 90 percent for first-time applicants from March to August 2019.

The expectation of USCIS is that applicants will have the resources to access websites or be able to go to a community college. There is no clear information on whether materials are made available to applicants through community centers for those whose household incomes are low. But an online program that is available on mobile phones, tablets, and computers would ease their learning burden as applicants can prepare for the test from home. In summary, all things considered, it could be concluded that *the opportunity to learn and opportunity for success are moderate to high*, although there is no research study that has documented the opportunity to learn.

Fairness 2: meaningfulness

Question: Do applicants find the test meaningful? The evidence in support of the question is as follows:

Constructs: The constructs of the English test (through the skills of speaking, reading, and writing) clearly *underrepresent* language proficiency. The tasks asked of the applicants are basic in listening and speaking tasks as well as in reading and writing. If we map the tasks onto the Common European Framework of Reference (CEFR), we could place the test at an A1 for listening and speaking and A2 for reading and writing. Passing this test would not ensure that the applicants were able to use English for a wide range of activities in the community.

The history and civics test, on the other hand, requires both memory of information (with 100 questions) and the ability to read the questions and multiple-choice options.

Consistency: Concern with the lack of consistency in the difficulty level of the questions administered was one of the reasons why the test was redesigned (Kunnan, 2009a). In the English test, sentences are created by using the vocabulary practice for reading and writing. However, it is not clear whether across applicants, the sentences for reading and writing will be of equal level of difficulty.

In terms of the administration of the history and civics test, the questions are supposedly selected carefully in terms of content. But we do not know if the sets are balanced in terms of difficulty; this could lead to some applicants being asked more difficult questions than others as different examiners may choose a different set of ten questions. It is also unclear whether test items are tagged with difficulty level indices so that a balance of "easy, medium, and difficult" items can be chosen for every applicant.

Besides, there are no questions on pivotal moments in American history that applicants may be familiar with, such as the treatment of Native Americans, the great migration of African Americans from the Jim Crow South to the North and West starting in 1915, the Depression of the 1930s, the 1945 atomic bombings of Hiroshima and Nagasaki, the House Un-American Activities Committee, the assassination of Martin Luther King, protests against the Vietnam War in the 1960s and 1970s, the Watergate scandal, Cesar Chavez and the farm movement, etc. Other events that are relevant to immigrants do not find a place either: for example, the Chinese Exclusion Act, the Japanese internment, the Mariel boatlift, the Vietnamese refugee crisis, etc.

Pickus (2005) concluded that the "questions deal with rights and freedoms, but none speaks of its obligations . . . civic knowledge does not make one a good citizen – native-born Americans do not have to take any such test" (p. 25a). Thus, in summary, it could be concluded that the test is not meaningful.

Fairness 3: absence of bias

Question: Is the test absent bias? The evidence in support of the question is as follows:

Item difficulty: As noted earlier, it is not clear that the applicants are administered items with the same level of difficulty on either the English test or the history and civics test. Thus, test bias cannot be ruled out. It is hard to know as no interview data is made available on whether there is adequate consistency across examiners in the country.

Item topics and content: The topics and content of the reading and writing sentences as well as the history and civics test items may also be biased. There is no information concerning whether the administrative procedures for both item difficulty and item topics and content are balanced. Thus, again, *bias cannot be ruled out*. Once again, there is no information regarding pass rates for applicants with low income and/or applicants from countries that do not share similar government structure with the US. Thus, in summary, it must be concluded that it is possible there is bias in the administration.

Fairness 4: access and accommodations

Question: Does the test offer enough access and adequate accommodations to applicants? The evidence in support of the question is as follows:

Costs: The cost of filing an application is $640 for Form N-400 and $85 for the biometric fee. A reduced filing fee of $320 plus the $85 biometric fees is available if the applicant's documented annual household income is greater than 150 percent but not more than 200 percent of the Federal Poverty Guidelines at the time of filing. However, the USCIS recently released a notice that it will raise the naturalization application fee from $640 to $1,170 and remove the reduced fee for applicants with low incomes.[10]

The Department of Homeland Security states in support of the fee changes,

> DHS proposed adjustments to USCIS' fee schedule to ensure full cost recovery. DHS did not target any group or class of individuals or propose changes with *the intent to deter requests from low-income immigrants* seeking family unity or deterring requests from any immigrants based on their financial or family situation or to block individuals from accessing immigrant benefits.
>
> *(USCIS, 2020 website, emphasis added)*

This is a big jump in fees. Even the current fee is more than one or two weeks' pay for many applicants; the latest fee increases and the cancellation of the reduced fee will clearly put the *citizenship application much behind in the immediate needs and desires of applicants with low income*.

Considerations: Due consideration exceptions are available for taking the English test for the following groups:

1. "50/20" exception: Applicants 50 years or older at the time of filing for naturalization and have lived as a permanent resident in the United States for 20 years.
2. "55/15" exception: Applicants 55 years or older at the time of filing for naturalization and have lived as a permanent resident in the United States for 15 years.
3. If applicants are exempt from the English test, they must take the civics test, but this may be done in the applicants' language of choice, with an interpreter, if necessary.
4. "65/20" exception: Applicants 65 years or older will be asked 10 questions from a total of only 20 US history and civics items marked with an asterisk in the official study materials. The test may also be taken in the language of the applicant's choice.

Accommodations: The following accommodations are available: for those with medical disabilities, the applicant's doctor must conduct an in-person examination of the applicant; explain the nature and extent of the medical condition on a form; explain how the medical condition relates to the applicant's inability to comply with the English and civics requirements; attest that the medical condition has lasted or is expected to last at least 12 months; attest that the cause of the medical condition is not related to the illegal use of drugs; and must complete the disability exception form using common terminology that a person without medical training can understand. However, there is no mention of American Sign Language for applicants who need this service.[11]

In summary, while considerations and accommodations available make the test accessible to most applicants, *the filing fees are extremely high*. There is no doubt that the fee increases will be a barrier to citizenship applications for many low- and moderate-income applicants despite the statement from the USCIS. The lack of American Sign Language is another barrier.

Fairness 5: consequences

Question: Is the test bringing about beneficial consequences to the applicant and their community? The evidence in support of the question is as follows:

Naturalizations and denials: Petitions filed for naturalizations and persons naturalized from 1907 to 2019 were 35,257,046 and 31,288,132, respectively, for an overall

88.74 percent rate of approval.[12] In the period from 2010 to 2019, the average number of applicants was over 700,000 for an overall 90.0 percent rate of approval. Table 6.1 provides details regarding applications, naturalizations, and denials for the last five years. While the applications denial rates have averaged less than 10 percent, the reasons for the denials could be many: failure to correct problems in personal documents indicated by the examiner, failure in the English or civics test (even after two attempts), or failure to continue the application process after initial filing of the petition. Table 6.2 shows naturalizations in from leading countries by year and percentage.

These official tables from the Department of Homeland Security provide a sense of the scale of the testing process and the overall success rate (of around 90 percent). Although the 90 percent pass rate is quite high, when we compute the 10 percent denial rate after the first interviews and tests, the actual number of denials could be about 80,000 a year. Even if some more applicants would be granted citizenships after the second round of interviews and tests, the number of denials is still high. In summary, the pass rates indicate *that the test is not difficult to pass for most applicants, but the high number of denials dampens the encouragement from the pass rate*. Further, there is no evidence that the successful applicants will go on to improve their English language ability and knowledge of American government and civics because of the USNT.

Table 6.1 Applications, naturalizations, and denials, 2015–2019

Year	*Applications filed*	*Persons naturalized*	*Applications denied*	*Denied percentage*
2015	783,062	730,259	75,810	9.68
2016	972,151	753,060	86,033	8.85
2017	986,851	707,265	83,176	8.43
2018	837,168	761,901	92,631	6.00
2019	830,560	843,593	97,789	11.77

Source: Table 20, US Department of Homeland Security,

www.dhs.gov/immigration-statistics/naturalizations

Note: In all DHS reports, "years" refer to US fiscal years, which run from October 1 to September 30. For example, fiscal year 2017 began on October 1, 2016 and ended on September 30, 2017. Thus, the number of naturalizations exceeds the number of applications in some years due to previously adjudicated application denials and delays in application processing.

Table 6.2 Naturalizations from six leading countries by year and percentage, 2015–2019

Country of birth	*2015*	*2016*	*2017*	*2018*	*2019*
Total	730,259	753,060	707,265	761,901	843,593
Mexico	105,958	103,550	118,559	131,977	122,286
India	42,213	46,188	50,802	52,194	64,631
China	31,241	35,794	37,674	39,600	39,490
Philippines	40,815	41,285	36,828	38,816	43,668
Dominican Rep	29,665	31,320	29,734	22,970	23,101
Cuba	25,770	32,101	25,961	32,089	36,246

Source: Yearbook of Immigration Statistics, 2019; U.S. Department of Homeland Security, www.dhs.gov/immigration-statistics/naturalizations

Justice 1: reasoning or justification

Question: Does the institution administering the test provide adequate reasoning for the test? The evidence in support of the question is as follows:

Civic nationalism: There is no recent reasoning or justification for the tests that were developed based on the 1952 Act. In a study on this question, Min (2010) investigated the level of political and civic activity among Korean Americans in Los Angeles before and after the taking of the tests for US citizenship. The main finding was that there was no statistically significant difference between political and civic activity before and after the test among these study participants: study participants who were active before taking the test continued to be that way, and individuals who were inactive before taking the test continued to be that way. The conclusion drawn was that *the US naturalization process did not help toward the goal of more civic nationalism*.

Reluctance to file applications: Martinez (2010) found that among Latinos in the Los Angeles area, there was significant reluctance in the community to apply for US citizenship. The main finding was that over 25 percent of the residents who were eligible to apply for citizenship do not apply. She found that among the many reasons for this reluctance was the educational requirement (the English language and history and government tests), which seems to be a big hurdle for this group of potential citizens who are non–English speaking adults. The number one reason listed is the language barrier – for the English test as well as the civics test. In summary, *the public reasoning and/or justification is lacking*, and the goal of turning immigrants into civic nationalists may not materialize this way.

Justice 2: advancing justice

Question: Does the institution administering the test advance justice related to the test? The evidence in support of the question is as follows:

Remedies: According to regulations, an applicant who passes the tests is informed by the examiner in person that they have passed the tests. Their application is then forwarded for eventual approval, leading to the citizenship ceremony. In the event the applicant fails any portion of the English test and/or civics test during their first examination, the applicant will be rescheduled to appear for a second opportunity to take the test(s) between 60 and 90 days after the first examination. In the second examination, the examiner is not permitted to administer the same English or civics test forms administered during the initial examination. In addition, the examiner is only to retest the applicant in the area(s) the applicant previously failed. If the applicant fails any portion of the test a second time, the examiner will deny the application based on the need to meet the educational requirements for naturalization. This means the applicant will have to start all over again with the filing of application, waiting for the tests to be scheduled, and so on. When all else fails, legal recourse through courts is also available.

Legal challenges: A legal challenge is another way of checking government and its policies. Garcia (2010) argued that the USNT can be challenged, citing the 14th Amendment to the US Constitution in terms of the way the test is administered – in terms of variability among the examiners in asking the questions. There is a general lack of an appeal process and a lack of public information regarding pass and fail rates by nationality, native language, or ethnicity. This, she argued, violates the equality and due process provisions of the Amendment.

In addition, as US-born citizens are not asked to demonstrate their ability in the English language or US history and government before they receive their driver's licenses or passports, the USNT could be in violation of laws against racial and ethnic discrimination, although courts are normally reluctant to review citizenship requirements as Article 1, Section 8 of the US Constitution leaves the power to establish rules regarding naturalization to Congress. The only action proven to force USCIS to a review in federal court is regarding delays or denials of applications. These appeals and lawsuits can be made for naturalization cases delayed before or after the interview and for naturalization denials. Federal law provides protection that applications should be decided within a reasonable amount of time, normally within six to twelve months.

Suggestions for reform

It should be clear from these analyses that the USNT is not clearly beneficial to the applicants for citizenship or to the community. Instead, there is enough evidence that the USNT is a barrier to citizenship. Here are the main reasons the USNT is a barrier:

1 No need for the test: Even if we accept the INA of 1952 and the English and US history and civics educational requirements as they stand, *there is no constitutional requirement that this be satisfied through a test*. Instead, applicants might be able to gain the basic English language proficiency and the memory-based history and civics knowledge by attending relevant courses at a community college or other advocacy group program. This would eliminate the need to assess applicants' skills and knowledge through these tests.
2 Non–English speaking applicants: *Both the English and history and civics tests are a barrier to applicants who are immigrants from non–English speaking countries*. This is particularly the case if their English proficiency is limited. In addition, the ability to remember facts and figures about US history and civics in English is a barrier. If knowledge is the main construct of the history and civics test, all applicants could be given the opportunity to meet this requirement in their first language (as is provided for the 50/20 or 55/15 exception applicants).
3 Costs of application: *Costs to applicants are a definite barrier to citizenship, especially to middle- and low-income applicants*. These applicants are likely to delay applying until their other needs are met. It is possible that full costs for the test are higher because of the face-to-face examination of every applicant by immigration examiners.

    There is a clear solution to this problem: the face-to-face examination, which includes the two tests, can be completed in community colleges or advocacy group centers. Another solution would be to exempt all holders of undergraduate or graduate degrees from a US university from the tests. Applicants who have studied at US universities are likely to have both English proficiency and maybe knowledge of US history and civics. Such applicants could be exempted from the English test and offered a discount on the application fee for a degree from a US university. This would reduce the time and costs for the interview and the naturalization process.[13]
4 Sub-constructs for the English test: Although the tests require a basic level of proficiency in the (listening and) speaking tasks, they are not based on a syllabus or any organized units. The tasks are random commands (please come in, please take a seat, please raise your hand, etc.). The reading and writing tasks, too, seem to be basic, but they are random sentences (based on the vocabulary list on the USCIS website).

Instead, a functional, notional, or task-based syllabus related to practical activities using English could be implemented (e.g., reading directions to their home to emergency staff, reading transport schedules, reading street signs or grocery labels, filling out forms, etc.). Of course, it is likely these interactive routines could be memorized. This should not be a problem as entry-level language learners are likely to memorize routines. Also, it is possible that these skills might be best achieved through a community college course eliminating the need for the test. The problems with the history and government section were discussed earlier.

5 Use of technology: Although most of today's applicants for citizenship are most likely technology savvy, the naturalization test is a paper-and-pencil test with no option for the use of a tablet or online test performance. In fact, the writing test that requires a handwritten sentence could very easily be changed to a typed sentence on a tablet or a computer. An online program for English language learning and for the history and government part would be highly beneficial to applicants.

6 Multilingual policy: While the US language policy is overtly monolingual English, it is the case that most cities have high numbers of speakers of other languages, and public services (e.g., driver's license tests, state and federal courts) are available in these languages (e.g., for cities in California and Florida). Therefore, a change in policy to go from monolingualism and promote multilingualism would be necessary. One way this could be operationalized would be to give applicants for citizenship credit for language proficiency in other languages in the form of reduced fees for the naturalization application.

7 Research: Much research needs to be conducted to understand the complex dimensions of how and when individuals and families decide to apply for citizenship. Who applies and who does not? Who succeeds and who does not? Why do they succeed; why do they not? Who participates in American democratic institutions (voting, signing petitions, attending local city council meetings, protesting injustices, etc.)? Much research is also needed to understand the consequences of the test. What happens to applicants who are unsuccessful? How many of them reapply, and how many of them are successful? For all this to happen, researchers need to have access to anonymized applicant test performance data.

Conclusion

In terms of the test, despite the clear guidelines and study materials for the redesigned test on the USCIS website, the process as it stands today is an unnecessary barrier to citizenship for non–English speaking applicants and low-income applicants. It is clearly not beneficial to the applicants or to the community. Etzioni (2007) summed it up this way 15 years ago: the "test hinders those who do not speak English and favors immigrants from English-speaking countries and persons who can afford extensive English education prior to their arrival, or once they are in the US" (p. 355). In addition, the whole process is expensive, untransparent, unnecessarily long, and fraught with anxiety.

A decade ago, I argued that the naturalization test is a barrier to citizenship (Kunnan, 2009a) and that it is unmeaningful and indefensible (Kunnan, 2009b, 2013). After ten years and after the revised test, I would like to argue much the same and add that it is more of a barrier now for the following reasons. It is also still unproven to be beneficial either for citizenship applicants or for the larger community. The xenophobia of the past (and perhaps the present in some political quarters) is reflected in some of these more recent manufactured

obstacles (such as excessive bureaucratic measures, meaningless tests, high application costs, and delayed processing). In addition, the overall process discourages many immigrants from applying for citizenship, and once they apply, the process is anxiety creating instead of being a welcoming and pleasant process. Right now, it is obvious the naturalization process is unnatural!

But despite these daunting obstacles, large numbers of legal permanent residents (green card holders) apply for citizenship each year because they believe in the aspiration of American democracy: a nation bound not by race or religion but by the shared values of freedom, liberty, and equality and for the practical benefits, which include permanent residence, employment, and opportunity to participate politically and have a voice in how the city, state, and nation are governed.

One clear way forward then would be to rethink the naturalization process for new transnational immigrants of the twenty-first century. This new conceptualization would have to include ways to cancel the xenophobia in all branches of federal and state governments, legislatures, and the judiciary; to embrace multilingualism and multiculturalism; and to welcome immigrants to apply with lower fees and smoother bureaucratic measures so that all residents would choose to become citizens, to become civic nationalists, and to have a political voice in the community. In brief, the rethink requires review and reflection of American political and moral theory of immigration and citizenship as well as the associated educational requirements.

Acknowledgements

I would like to acknowledge Gary J. Ockey and Luke Harding for their useful comments on earlier versions of this chapter and Don Yao and Fiona Yang for their work on building a database on immigration and citizenship testing around the world. Any deficiencies and inconsistencies are mine. This project did not receive funding from any organization.

Further reading

Canagarajah, S. (ed.). (2017). *The Routledge Handbook of Migration and Language*. New York, NY: Routledge.
This collection of 31 chapters is an exploration of concepts, contexts, methods, and policies related to immigration, citizenship, and language.

Cisneros, J. D. (ed.). (2013). *The Border Crossed U.S.* Tuscaloosa, AL: University of Alabama Press.
This collection of essays offers new ways of thinking of the border between the US and Mexico by examining race, coloniality and citizenship, and radical and hybrid rhetoric social movements related to immigration and citizenship.

Hartelius, E. Johanna. (2015). *The Rhetorics of U.S. Immigration: Identity, Community, Otherness*. State College, PA: Penn State University Press.
This monograph examines US immigration as a rhetorical process that invents "persons and communities with respect to space and place" (p. 1).

Notes

1 In the recent past, President Trump railed publicly against immigrants from Asia, Africa, the Caribbean, and Central America. He signed an executive ordering banning the entry of immigrants from Muslim-dominated countries, which has now been revoked by President Biden.

2 See Jones's *Birthrights Citizens* (2018) for a detailed examination of the history of race and rights in antebellum America.
3 This Act was repealed only in 1943.
4 See Ngai's (2004) *Impossible Subjects*, which explores the origins of the term *illegal alien* in American law and society.
5 In 1991, the Immigration and Naturalization Service contracted with six private testing services to administer a standardized test for this specific purpose. By 1996, these organizations operated about 1,000 testing sites through subcontractors. This arrangement was stopped and contracts canceled after an episode of the TV program *20/20* exposed testing fraud. The test then reverted to the control of the INS, and USCIS is now part of the naturalization process.
6 The Citizenship Resource Center provides information regarding the latest rules and regulations for applying for citizenship.
7 See "The case for the literacy test." *Immigration Restriction Leagu*e, Publication no. 66 (1906).
8 A variant of the literacy test in English was also used in both northern and southern states to determine eligibility for voting, effectively denying voting rights to large numbers of African Americans in many southern states and disenfranchising a million Yiddish speakers in New York (del Valle, 2003). Decades later, the Voting Rights Act of 1965 suspended the literacy test in all states; in 1970, Congress extended the prohibition of the literacy tests and in 1975 made the ban indefinite and national in scope.
9 From www.uscis.gov/sites/default/files/document/guides/Test_Scoring_Guidelines.pdf.
10 For details regarding new fee proposals, see www.govinfo.gov/content/pkg/FR-2020-08-03/pdf/2020-16389.pdf.
11 The Americans with Disabilities Act of 1990 states "that people with disabilities have the same opportunities as everyone else to participate in the mainstream of American life – to enjoy employment opportunities, to purchase goods and services, and to participate in State and local government programs and services."
12 Data on naturalizations were first compiled by a single federal agency with the establishment of the Bureau of Immigration in 1906. The year 1907 includes naturalizations from September 27, 1906, to June 30, 1907.
13 The author of this chapter was asked to write the sentence "Today is a beautiful day" as part of the writing test. This despite the author's completion of a doctoral degree at a US university with a dissertation written in English. The dissertation was on the officer's desk.

References

Americans with Disabilities Act of 1990, Pub. L. No. 101–336, § 2, 104 Stat. 328. (1991).

Crawford, J. (1992). *Language loyalties*. Chicago, IL: University of Chicago Press.

De Valle, S. (2003). *Language Rights and the Law*. Bristol, UK: Multilingual Matters.

Etzioni, A. (2007). Citizenship tests: A comparative, communitarian perspective. *Political Quarterly* 78: 353–63. https://doi.org/10.1111/j.1467-923X.2007.00864.x.

Gambino, C., Acosta, Y. and Grieco, E. (2014). *English-Speaking Ability of the Foreign-Born Population in the United States*. American Community Survey Reports 26. Washington, DC: US Census Bureau.

Garcia, J. (2010). *Is the U.S. Naturalization Test in Violation of the 14th Amendment of the U.S. Constitution?* Paper presented at the SCALAR 13 Conference, UCLA, Los Angeles.

Huntington, S. (2005). *Who Are We? America's Great Debate*. New York, NY: Free Press.

The Immigration Act of 1924. United States. Congress. Senate. Committee on Immigration. (1928). Restriction of western hemisphere immigration, etc. February 1, 27–29; March 1 and 5, 1928. US Government Printing Office.

The Immigration and Nationality Act of 1952. Pub. L. No. 82-414, Stat. 163, Title 8 of the U.S. Code, Chapter 12. USCIS, Department of Homeland Services.

Jones, M. (2018). *Birthright Citizens*. Cambrige, UK: CambridgeUniversity Press.

Kennedy, J. F. (1964). *A Nation of Immigrants*. New York, NY: Harper.

Kunnan, A. J. (2009a). Politics and legislation in immigration and citizenship testing: The U.S. case. *Annual Review of Applied Linguistics* 29: 37–48. https://doi.org/10.1017/S0267190509090047.

———. (2009b). The U.S. Naturalization Test. *Language Assessment Quarterly* 6: 89–97. https://doi.org/10.1080/15434300802606630.
———. (2013). Language assessment for immigration and citizenship. In G. Fulcher and F. Davidson (eds.), *The Handbook of Language Testing*. New York, NY: Routledge, 152–66.
———. (2018). *Evaluating Language Assessments*. New York, NY: Routledge.
Kunnan, A. J., Yao, D. and Yang, F. (2019). *Language Requirements for Immigration and Citizenship: A Database*. Unpublished manuscript, University of Macau.
Kymlicka, W. (1995). *Multicultural Citizenship: A Liberal Theory of Minority Rights*. Oxford, UK: Clarendon Press.
Lee, E. (2019). *America for Americans: A History of Xenophobia in the United States*. New York, NY: Basic Books.
Lew-Williams, B. (2018). *The Chinese Must Go*. Harvard, MA: Harvard University Press.
Martinez, J. (2010). *Why Are Eligible Latinos Not Becoming U.S. Citizens?* Seminar paper, TESL 567b, California State University, Los Angeles.
May, S. (2005). Language rights: Moving the debate forward. *Journal of Sociolinguistics* 9: 319–47. https://doi.org/10.1111/j.1360-6441.2005.00295.x.
———. (2008). Language education, pluralism, and citizenship. In S. May and N. Hornberger (eds.), *Encyclopedia of Language and Education. Language Policy and Political Issues in Education*, 2nd edn., vol. 1. Amsterdam, The Netherlands: Springer Science, 15–29.
Min, K. (2010). *Is the U.S. Naturalization Test Meaningful in Achieving its Purposes of 'Civic Nationalism,' 'Social Integration,' and 'Political Allegiance' with Citizens of Korean Origin?* Seminar paper, TESL 567b, California State University, Los Angeles.
Ngai, M. (2004). *Impossible Subjects*. Princeton, NJ: Princeton University Press.
The Naturalization Act of 1790. A Bill to Establish a Uniform Rule of Naturalization, and Enable Aliens to Hold Lands under Certain Conditions; 3/4/1790, etc. U.S. Senate, Record Group 46; National Archives Building, Washington, DC.
Parker, K. (2015). *Making Foreigners*. Cambridge, UK: Cambridge University Press.
Pavlenko, A. (2002). 'We have room for but one language here': Language and national identity in the U.S. at the turn of the 20th century. *Multilingua* 21: 163–96. https://doi.org/10.1515/mult.2002.008.
Pickus, N. (2005). *True Faith and Allegiance: Immigration and American Civic Nationalism*. Princeton, NJ: Princeton University Press.
Pogge, T. (2003). Accommodation rights for Hispanics. In W. Kymlicka and A. Patten (eds.), *Language Rights and Political Theory*. Oxford, UK: Oxford University Press, 105–22.
Ricento, T. and Wright, W. (2008). Language policy and education in the United States. In S. May and N. Hornberger (eds.), *Encyclopedia of Language Education*, 2nd edn., vol. 1. Amsterdam, The Netherlands: Springer, 285–300.
Schlesinger, A. (1992). *The Disuniting if America: Reflections on a Multicultural Society*. New York, NY: WW Norton.
Thompson, F. (1920). *Schooling of the immigrant*. [Reprinted 2015]. Andesite Press.
U.S. Citizenship and Immigration Services. (2020). Retrieved December 10, 2020, from http://www.uscis.gov.

Websites

Citizenship Resource Center. Retrieved from www.uscis.gov/citizenship.
Scoring guidelines of the test. Retrieved from www.U.S.cis.gov/sites/default/files/U.S.CIS/Office%20of%20Citizenship/Citizenship%20Resource%20Center%20Site/Publications/PDFs/Test_Scoring_Guidelines.pdf.
Civics (history and government) questions for the naturalization test. Retrieved from www.U.S.cis.gov/sites/default/files/document/questions-and-answers/100q.pdf.

Court cases

Dred Scott v. Sandford, 60 U.S. 393 (1857).
Farrington v. Tokushige, 273 U.S. 284 (1927).
Lau v. Nichols, 414 U.S. 563 (1974).
Meyer v. Nebraska, 262 U.S. 390 (1923).

Section 3

Classroom assessment and washback


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

7

Classroom-based assessment

Janna Fox, Nwara Abdulhamid, and Carolyn E. Turner

Introduction and definitions

Classroom-based assessment (CBA) in language learning contexts was defined in the first edition of this chapter as "assessment internal to the classroom and managed by the teacher" (Turner, 2012: 65). CBA practices have traditionally been classified as summative and formative (Heitink *et al*., 2016); however, increasing awareness of the crucial role of such assessment in describing, supporting, and documenting learning (Andrade and Brookhart, 2019; Panadero *et al*., 2018) has prompted a refinement of distinctions in CBA practices. We begin this chapter by considering both traditional and new definitions that clarify these distinctions and their historical evolution in practice.

What is CBA?

Although there is recognition that teachers use many different methods in their classrooms to gauge the learning taking place (including tests and quizzes), many books, articles, and presentations continue to distinguish between "testing" and "assessment," with assessment more often than not referring to anything *but* tests. In this chapter, classroom-based assessment is viewed as the umbrella term which encompasses the broad array of methods – including tests – teachers use in their language classrooms to describe, support, and document learning.

Traditional definitions

The traditional distinction between summative and formative purposes for the evaluation of learning first appeared in the work of Bloom (1956) and Scriven (1967). Summative assessment practices elicit evidence of students' achievement and/or confirm that they have met or exceeded requirements at the end of a period of study. They are (1) typically associated with products of learning (e.g., tests, essays, presentations); (2) used by teachers to account for grades on final assignments, quizzes, or report cards; and, (3) used by institutions in awarding school diplomas and as tools for selection.

DOI: 10.4324/9781003220756-11

Formative assessment practices are embedded in day-to-day classroom activity for the purpose of promoting learning and informing teaching. They take many forms, as teachers collect and reflect on evidence of their students' learning through a "repertoire of methods and reflective procedures" (Turner, 2012: 66), including observations, conferences, discussions, projects, task sheets, quizzes, and student self- and peer assessment (e.g., Andrade and Brookhart, 2019; Cheng and Fox, 2017). Within a classroom, "teachers and students interweave classroom interaction, formative assessment, and teaching and learning in a complex but on-flowing process" (Leung and Mohan, 2004: 343). The evidence generated in this process informs next steps to help fill gaps toward learning goals.

Newer distinctions in CBA

Since these early definitions were introduced, a plethora of new terminology has emerged. It should be noted that differences in CBA definitions may be attributed in part to the particular roles played by teachers and students in classroom assessment practices. As Heitink *et al*. (2016) point out, teachers and students can alternate roles as providers, recipients, and/or users of evidence (e.g., in needs analysis, feedback, self-reflection, or reporting). How CBA practices are defined also depends on how the evidence of learning is used and the particular curricular context that informs those practices. Further, different disciplines refer to the same or similar concepts with different terminology. However, over the past decade, greater focus on CBA has been accompanied by increased clarity with regard to distinct approaches for gathering and using evidence from assessment to describe and/or promote learning in the language classroom:

Assessment of learning (AoL) is summative assessment which focuses on achievement or outcomes of teaching and learning and accumulates evidence of the products of teaching and learning internal to the classroom. The term *AoL* tends to be used in discussions of pedagogical practices within the classroom, whereas *summative assessment* appears more typically in educational measurement and may be both internal to the classroom (e.g., class quizzes, final essays, examinations) or external to it (e.g., matriculation examinations, regional or national achievement tests).

Assessment for learning (AfL) is a term that originated in retheorization and reconceptualizaton of classroom-based formative assessment (e.g., Black and Wiliam, 1998; Rea-Dickins, 2001). In 2002, the Assessment Reform Group provided an early, comprehensive definition of AfL as the "process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there" (p. 2).

AfL practices create collaborative space and time for "useful and helpful teacher feedback, so that students can engage with further learning and/ or revise what has been learned" (Leung and Mohan, 2004: 338). Such practices are "interactionally realized through teacher-student talk" (Leung and Mohan, 2004: 336) – both planned and spontaneously arising in the act of teaching itself or in teachers' written responses to their students' work. AfL stimulates reflection, raises awareness, and encourages students' realistic goal setting and autonomy (Little, 2005). AfL activities are a fundamental learning resource which supports motivation (cf. Cheng and Fox, 2017; Little, 2005; Wiliam, 2011). At the same time, AfL provides teachers with a window on student thinking, knowledge, and understanding, Currently, there appears to be universal agreement that AfL plays a pivotal and critical role in learning (cf. Cizek, 2010; Vogt and Tsagari, 2014).

Assessment as learning (AaL) is defined by Earl (2013) as "a subset of AfL [that] emphasizes the important role of students as active agents in the assessment process." She argues that AaL creates the possibility for students to "not only contribute [to] but also connect assessment with their previous learning [and] to set up individualized goals for progress" (p. 553). Examples of AaL are *can-do inventories*, which ask students to self-assess their language proficiency for the purpose of identifying strengths and weaknesses; student selection of their best work for final evaluation in a portfolio collection; and peer assessment of a draft composition produced by a fellow classmate. Cheng and Fox (2017) identify AaL tasks and activities as those which "allow students the opportunity to use assessment to further their own learning" – to reflect on, monitor, define their own goals, and "advocate for their own learning" (p. 182). AaL is related to the concept of self-regulated learning (SRL). Proponents of SRL (e.g., Andrade and Brookhart, 2019) emphasize the key role of "students making their own meaning by actively connecting new information and experiences with prior understandings" (p. 1). SRL is, therefore, "self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals" (Zimmerman, 2000: 14).

Having introduced some of the key definitions that currently inform discussions of CBA, we next consider its evolution in relation to language teaching methods and classroom assessment practices.

Historical perspectives

How we define what it means to know a language has influenced how we have taught it. How we have taught it at the classroom level has influenced how we have assessed it. Traditionally, knowing a language was viewed as knowing about it. Grammar translation methods focused on mastering knowledge of grammar and vocabulary, often through direct translation of texts that were highly valued for cultural, historical, or literary merits. Assessment in the language classroom was therefore based on accuracy and the production and recognition of correct forms. Teachers typically used tests of grammar and vocabulary, word-by-word translations, and dictation to assess achievement. The focus was on summative outcomes while formative feedback was provided to support students' development of correct and accurate language forms. In general, teaching was dominated by a *transmission model*; the teacher's role as a repository of knowledge was to provide accurate and precise information about the language. The student's role was to "receive, store . . . [and] to commit the facts and procedures to memory" (Tishman *et al*., 1993: 149). Most CBA was summative – designed to evaluate the precision of the student's recall through translation, recitation, and tests of vocabulary and grammar.

As definitions of language evolved, so too did CBA practices. In a state-of-the-art review, Breen (1987a, 1987b) recounts the progression of methods in language teaching, documenting the transition from the teaching of language as content to the teaching of language use for meaningful communication. He examines the rise of communicative language teaching (CLT) approaches in the 1970s (cf. Hymes, 1972; Savignon, 1987), which were accompanied by radical shifts in classroom assessment practices, characterized by performance-based, direct assessment of speaking and writing; increased value for the meaningful use of language; and a decreased value for form (e.g., grammar, spelling).

By the 1980s, some CLT approaches, informed by theories of communicative competence and second language acquisition, had begun to systematically include task-based classroom activity, which required meaningful interaction and communication for completion. Breen

(1987a, 1987b) highlights the contribution of Prabhu's (1987) *Bangalore Project* in this regard, because it signaled a shift away from transmission- and grammar-based instruction toward meaningful communication and language use in context and is an early example of task-based language teaching (TBLT). While Prabhu investigated foreign language learning in India, Bygate *et al*. (2001) focused on the use of pedagogical tasks in second language learning contexts. They defined tasks as "a central element of language pedagogy" (p. 1) and highlighted the value of "tasks to provide formative information, during instruction, so that learners and teachers are better informed about the progress that has been made" (p. 8). Further, they noted the potential of tasks "to generate feedback . . . [or] to provide diagnostic information to indicate to learners where their strengths and weaknesses lie and how they might be improved" (p. 8).

It is important to note that the work of Prabhu (1987) and Bygate *et al*. (2001) was underway at approximately the same time as that of the Assessment Reform Group (2002). Efforts to refine understanding of formative assessment practices, particularly AfL practices in the classroom, were all part of a progressive movement toward learner-centred learning. For example, the increased focus on formative assessment coincided with more learner-centred language teaching approaches (e.g., Nunan, 1988) and the increasing prevalence of needs analysis as a recurring AfL practice undertaken by teachers at the beginning of a new course and at intervals throughout a course. Needs analysis was both formal (e.g., questionnaires, reflective essays, interest inventories, can-do statements) and informal (e.g., observations, question-and-answer sessions, group discussions) and, like all AfL practices, was used to inform teachers' planning and decision making.

Student-led approaches to self- and peer assessment were also widely discussed in the language assessment literature (e.g., Hamp-Lyons and Condon, 2000), in particular within writing studies and writing assessment (e.g., Elbow and Belanoff, 1986). Thus, Black and Wiliam's (1998) seminal paper on formative assessment for the Assessment Reform Group (2002) was part of a proliferation of research on CBA (e.g., Leung, 2004; Rea-Dickins, 2006). Formative assessment secured its place as a "pedagogically desirable approach to assessment which is capable of promoting learning" (Leung and Mohan, 2004: 335); it was generally acknowledged that teachers' assessment should be both summative and formative, and assessment should be fully integrated in day-to-day teaching (Leung and Mohan, 2004). At this time, discursively derived, empirically based rating scales (Upshur and Turner, 1995; Fulcher, 1996) provided teachers with methods for rating scale development derived from their students' work. Teachers' active participation in the development of the scales not only increased consistency in their assessment but was also a context for professional development.

Early proponents of formative assessment (Harlen and Winter, 2004; Leung and Mohan, 2004; Rea-Dickins, 2006) called for the expansion of theoretically and empirically informed research to investigate how formative assessment could be best accomplished at the classroom level. Since that time, as noted by Hill and McNamara (2012), CBA studies have focused on multiple issues, including validity and reliability (Andrade and Cizek, 2010; Gipps, 1994; Moss, 1994, 2016), the washback of high-stakes testing on classroom teaching and learning (e.g., Cheng and Sultana, this volume; Cheng *et al*., 2004; Turner, 2009), and teacher assessment literacy (Taylor, 2009; Vogt and Tsagari, 2014).

Over the past decade, benchmarks and standards movements – e.g., Common European Framework of Reference (CEFR), Canadian Language Benchmarks (CLB), and China's Standards of English Language Ability (CSE) – have dramatically influenced teaching, learning, and assessment in language classrooms (see also Deygers, this volume). Benchmarks

are criterion-referenced descriptors that define degrees of proficiency (e.g., knowledge, skill, competence) at "successive levels on a continuum of language ability" (*On CLB*, retrieved from www.language.ca/resourcesexpertise/on-clb/). They are often used for placement and promotion if they are referenced to levels in programs or educational systems.

When benchmarks are referenced to educational programs/systems in terms of what students should know or be able to do at successive levels, they are often referred to as standards. Standards translate readily into constructs, which can be measured by at-a-distance, centralized tests that take *snapshots* of students', schools', or systems' achievement at one moment in time. However, they do not capture a contextually rich, fine-grained understanding of individual students' strengths and weaknesses, developmental growth, challenges, and potential, which are evident to teachers engaged in CBA over time (Birenbaum *et al*., 2015; Macqueen *et al*., 2019).

Although currently formative assessment/AfL is emphasized in educational policies around the world and its positive impact on learning is supported by extensive research, according to Birenbaum *et al*. (2015), it appears that, globally, "ongoing tensions between formative and summative forms of assessments, as reflected in large-scale testing programs, continue to pose a significant risk to the uptake of authentic and sustained AfL practices in school systems" (p. 135). In the following section, we further explore these tensions.

Critical issues and topics

In 2012, Turner expressed concern about tensions between CBA and external large-scale, standards-driven tests, pointing out that teachers continued "to grapple with the relationship between what they are doing in the classroom and . . . the external large-scale tests their students . . . need to take" (p. 69). She observed that while external tests were "recording progress and achievement" (p. 69) at a single point in time, CBA was drawing on a large repertoire of assessment practices (including tests) in amassing ongoing evidence of learning: of "how students learn and the processes to help them learn" (p. 69). Further, she questioned how alignment of assessment at these two levels might unfold in practice.

Early discussions of *alignment* (Pellegrino *et al*., 2001) viewed it as the degree of agreement among policy, curricula, large-scale external tests, and CBA practices. Alignment was envisioned as a "comprehensive, coherent, and continuous" (p. 9) assessment system. It created the potential for evidence drawn from both CBA practices in unique, local contexts and large-scale external tests to form a relationship and work together to describe and promote learning. As Pellegrino *et al*. (2001) described it: "Assessment at all levels would be linked back to the same underlying model of student learning and would provide indications of student growth over time" (p. 9). The problem, which Turner foresaw, is that the models of student learning at play in CBA and those that prevail in external large-scale testing are generally not the same, nor do these two levels of the assessment system share the same mandates or the same collections of evidence.

The tension between CBA and increasingly pervasive external testing is becoming more alarming as evidence mounts of its impact on classroom teaching, learning, and assessment (e.g., Birenbaum *et al*., 2015; Kempf, 2016; Macqueen *et al*., 2019; Torrance, 2015). External-to-the-classroom testing takes place not only within countries but also internationally, in large-scale assessments such as the Organisation for Economic Co-operation and Development's (OECD) Programme for International Student Assessment (PISA), which evaluates, compares, and ranks educational systems around the world.

Policy makers typically promote external standardized testing as a means of assuring accountability, identifying and responding to underachievement, and monitoring progress. The ostensible intent of such testing is fairness, consistency of opportunity, and improved outcomes for all. However, underperformance on external tests has also resulted in:

1 criticism of teachers and schools, regardless of local needs and contextual factors (Torrance, 2015);
2 undermining of locally effective teaching and learning approaches (Kempf, 2016; Macqueen *et al*., 2019);
3 implementation of unwarranted changes in educational policy; and/or
4 ill-advised educational reforms (Chalhoub-Deville, 2016) due, for example, to a published decline in ranking or a lower-than-expected performance on a league table comparison with other countries.

In a study of education in the US and Canada, Kempf (2016), reports that external, high-stakes standardized testing is changing how teachers teach and students learn. Eliciting interview and survey data from thousands of teachers, he identifies negative consequences of external, standardized testing on classroom practices, concluding, "Teachers' work was changing . . . and the testing question for many was trumping more advanced questions of practice" (p. 100). As is the case in much other recent research on the impact of external high-stakes testing on classroom teaching and assessment practices (e.g., Birenbaum *et al*., 2015; Macqueen *et al*., 2019; Torrance, 2015), Kempf (2016) reported recurrent and dominant themes of "deprofessionalization, alienation, and frustration" (p. 100) as "class time spent on testing consumes scarce time and attention needed to support the success of all students – further disadvantaging [English language learners], students with exceptionalities, low income, and racially minoritized students" (p. 100).

Given the current global emphasis on assessment, teachers' language assessment literacy (LAL) "has taken on new importance in educational systems" (Vogt and Tsagari, 2014: 375). Some have argued for increased assessment literacy as one means of combatting the impact of external tests on teachers' assessment in the classroom. However, as Abdulhamid (2018) and Sultana (2019) report, increased knowledge, skill, and practice in assessment may be insufficient when classroom teachers are working in language teaching contexts governed by policies and educational cultures which are test intensive. In many cases, students' needs are best addressed by preparing them to pass powerful, external tests, and language learning is relegated to a secondary and somewhat incidental goal. In order to address negative consequences, many (Chatterji, 2013; Macqueen *et al*., 2019) have been calling for a reconsideration of what counts as evidence of learning, noting that CBA evidence should inform inferences drawn from external tests and, indeed, work seamlessly with them in supporting learning (Turner and Purpura, 2016), just as Pellegrino *et al*. (2001) suggested it should in their early considerations of alignment.

The recent unifying concept of *learning oriented assessment (LOA)* (Turner and Purpura, 2016; Purpura and Turner, forthcoming) may provide a positive means of addressing these issues. LOA prioritizes learning and focuses on how assessment fits into the larger concept of learning. Purpura and Turner are developing a working framework with interrelated dimensions to help tease apart LOA and demonstrate how appropriate assessments can help guide next steps for further learning in relation to language-learning goals and, concomitantly, provide useful evidence to inform inferences drawn from external tests (as discussed in Moss, 2016). The framework is conceptualized in terms of seven dimensions.

Two of the dimensions are viewed as performance indicators: elicitation/evidence and proficiency. The other five dimensions are viewed as potential performance moderators: contextual, socio-cognitive, instructional, social-interactional, and affective. The multiple dimensions converge to address synergies among assessment, teaching, and learning and provide a richer description of assessment information, thus making it more accessible for use. LOA is particularly concerned with how information from assessments can be used to close learning gaps.

Current contributions and research

Of the many current contributions to the assessment research literature, we highlight the following:

Assessment-centred language teaching: learning outcomes and assessment tasks

New models of course design (Biggs and Tang, 2011) have placed assessment at the centre of a teacher's planning as learning outcomes and assessment tasks precede the development of ongoing classroom activities. *Learning outcomes* are defined in advance of a new course and state what a successful student should know, be able to do, and/or value at the end of a course. They may be explicitly identified in policy or curricular documents (e.g., benchmarks, a course syllabus, teachers' guides, textbooks) or developed by teachers themselves. They are linked through *backward design* (Wiggins and McTighe, 2005) directly to assessment tasks, which are, in turn, supported by ongoing classroom activity (Cheng and Fox, 2017). A learning outcome by definition requires evidence, and thus, teachers' CBA practices result in collections of evidence that a student has met, not met, or exceeded the predetermined outcomes (e.g., from group projects, presentations, portfolios, reflective journals, essays, tests).

One consequence of this curricular approach has been the increased focus of language teachers on the development of assessment tasks and criterion-referenced rubrics (i.e., rating scales) as a means of providing evidence that a student has met, not met, or exceeded an outcome. Abdulhamid and Fox (2020) report that this increased focus has had a positive impact on teachers' organization, planning, and awareness of learning development. However, some teachers have become concerned about the unintended, negative consequences of outcomes-based approaches in contexts where rigid evidentiary requirements are in place. In such contexts, teaching can become *test driven* rather than *learning driven*.

Another consequence of an outcomes-based approach at the classroom level has been a revitalization of language functions (e.g., greetings, apologies, requests), which are often linked in benchmark documents to tasks by levels (see also functional language teaching approaches in Breen, 1987a, 1987b; Curtis, 2017). This model of course design is most easily implemented through task-based language teaching (TBLT).

CBA and diagnostic assessment

Although needs analysis is frequently used as a CBA approach, classroom-based diagnostic assessment (CBDA) has become increasingly prevalent in recent years (cf. Fox, 2009; Fox and Artemeva, 2017; Jang and Sinclair, this volume; Jang and Wagner, 2013). CBDA is linked to specific pedagogical support in the classroom (e.g., individualized in-class learning

plans, instruction, and feedback; one-on-one tutorials; negotiated learning contracts). It is more complex, scientific (e.g., Alderson, 2005; Jang, 2009; Lee, 2015), and systematically principled (Alderson *et al*., 2015; Harding *et al*., 2015) than needs analysis in eliciting information specific to learning support for an individual student. It is used across all educational levels, with recent increases in its use at tertiary levels (Read, 2016) in both language (e.g., Doe, 2015; Fox and Hartwick, 2011) and discipline-specific classrooms (Fox and Artemeva, 2017). Although the uptake of diagnostic assessment information by teachers and students varies (Doe, 2015; Fox, 2009), individual learning profiles can provide an important source of information for teachers' planning and organization of classroom activity. Further, CBDA projects often involve partnerships between testers, charged with developing sophisticated diagnostic assessments, and teachers, who actively contribute to the development of the assessment and apply and refine its use through practice in their classrooms.

The rise of technologically mediated CBA

Although the pandemic of the 2020s triggered a global, sudden, and dramatic move to online teaching and learning, during the previous decade, there had been an ever-increasing trend toward technologically mediated CBA (e.g., Chapelle and Sauro, 2017; see also Sawaki, this volume). Technology offers teachers greatly expanded potential for diagnosing, monitoring, and responding to individual student needs; collecting more precise evidence of learning over time; and systematically assessing incremental language development. Technological advances have broadened and deepened potential learning spaces (Savin-Baden, 2008), which are no longer defined by four walls, desks, and chalkboards or traditional time frames imposed by a lesson, unit, or course. Technology is also transforming classroom assessment. Computer programs can automatically generate individual learning profiles, drawing on samples of student work (e.g., recordings, writing, projects) collected over time. Data analytics that are built into online activities can be applied to document changes, issues, and development. For example, speech analysis software can provide information on emphasis, pitch, speed, and spectrographic comparisons between a student's speech and that of a fully intelligible speaker. At intervals, the software can generate a learning profile which documents a student's development, is a resource for reflection and increased self-awareness, and is a nexus of ongoing interactive feedback and discussion. Computerized feedback can be immediate, scientific, and available on demand for just-in-time learning (Hartwick, 2018; Hartwick and Nowlan, 2018; Purpura, 2019).

The ePortfolio, defined by Eynon and Gambino (2018) as "[w]eb-based, student-generated collections of learning artifacts (papers, multimedia projects, speeches, images, etc.) and related reflections, focused on learning and growth" (p. xix), provides a dynamic, interactive site for teacher-student and student-student interactions. Eynon and Gambino argue that ePortfolios are potential learning catalysts for AfL when they are "done well": that is, when they are managed by teachers whose "sophisticated" (p. xix) teaching and assessment practices place their students at the heart of assessment processes in the classroom. Learning processes (e.g., planning documents, observational checklists, progress reports on key events, video clips of a rehearsal, pronunciation exercises) and products (e.g., poster presentations, oral reports, written commentaries, tests) may be housed in ePortfolios and used on an ongoing basis to stimulate reflective dialogue on learning between teachers and students and between students and their peers (cf. Barrett, 2007).

Technology is also supporting our understanding of language development – of what to assess as evidence of language learning. Alongside their large-scale corpus counterparts

(e.g., COCA, MICASE), specialized *learner corpora*, gathered as a result of expanding technologically mediated CBA, are increasingly supporting greater understanding of acquisition, language progression, teaching, learning, and assessment (Callies and Götz, 2015; see also Cushing, this volume). Spurred on by the pandemic of the early 2020s, learner corpora are often the result of working partnerships between classroom teachers, researchers, and other stakeholders (e.g., language testers, educational developers, computer software specialists). They are developed over time from collections of student work (spoken and written) across program levels and are often stored for analysis on classroom management systems. Learner corpora can be used to inform program organization (e.g., levels, focus), curricula, learning materials, and assessment tools. For example, they provide an empirical resource for interventions used in *dynamic assessment* approaches (Anton and Garcia, this volume; Poehner and Lantolf, 2013), whereby an assessor interacts with and intervenes in a test taker's/student's engagement with a pre-selected activity/task in order to describe "the emergence of capacities" and "the results of those interventions in detail" (van Lier, 2000: 256).

New online learning spaces have prompted some particularly innovative language teachers (e.g., Hartwick, 2018) to ask their students to create avatars, invent digital identities, and undertake classroom activities in 3D virtual learning environments (3DVLEs) in order to complete an assignment, contribute to a project, or retrieve information for a problem-solving activity. Such 3DVL environments are typically housed on existing platforms, which can be modified by language teachers working in partnership with educational technology experts. Features can be embedded within the activities which automatically collect *learning/data analytics* or *telemetrics* as students engage in learning tasks or activities, alone or in collaboration with others. Such automated telemetric features measure, analyze, and report data about learners and their learning to inform teachers' understanding, planning, and decision making and raise learners' awareness of how well they are doing and how they might improve.

Hartwick and Nowlan (2018) provide examples of how such analytics in a classroom-based 3DVLE can support and refine assessment in an EAP context by supplying immediate automated feedback to students on their responses to a problem-solving activity (e.g., "Stop. Did you read the instructions first?") or information for their teacher regarding how much time a student took to solve the problem, the sequence of steps the student employed, the strategies they used, and the choices they made. New virtual learning environments allow language students to interact with students studying in other classes, schools, cities, or countries; to attend online lectures delivered at a distance; and to participate in webinars or workshops specific to their individual learning interests or needs.

Although there has been extraordinary growth in research related to CBA practices and processes over the past decade (Vogt and Tsagari, 2014), tensions remain regarding the value of such contextually rich research and the role it should play in shaping policy and decision making. We briefly discuss these tensions next.

Reaffirming the value of CBA evidence of learning

In 2012, Turner asked whether reliability and validity (qualities that are associated with large-scale assessment, the measurement community, and quantitative methodological approaches) were "relevant" (p. 68) for CBA and identified the need to "reach a balance between CBA strategies and technical test development strategies" (p. 70). As Turner pointed out, CBA has as its purpose to "support learning and through feedback to help

learners understand where they are in their learning, where they need to go and how best to get there" (p. 68). This purpose is directly reflective of the local context of the classroom (Van den Branden, 2016), which is situated, unique, and variable; "where teachers know their students and work with them, where student reflection and performance promote and encourage further learning via self, peer, and teacher feedback" (Turner, 2012: 68). Reaffirming the value of local, variable, unique evidence of learning, which arises in and from the day-to-day interaction between teachers and students in the classroom, can only enhance our understanding of how much, what kind, and for whom language learning is taking place.

At the present time, there is evidence that the pervasive reliance on external standardized testing, which speaks to a preference of some policy and decision makers for quantitative measures of learning over a combination of both qualitative and quantitative evidence collected over time through CBA practices, can have a negative and substantive impact on language teachers' current classroom practices and students' learning. Over-valuing external tests can demean teachers' professional judgment and worth (Kempf, 2016), create feelings of "disempowerment and alienation" (Blazer, 2011: 6), and undermine teaching and learning practice (Macqueen *et al*., 2019). There is counter evidence that suggests that collaborating with teachers in the promotion of learning-oriented assessment improves learning quality (Moss, 2016; Van Viegen Stille *et al*., 2016). Re-valuing CBA as an important and legitimate source of evidence about language learning re-establishes teachers as trusted professional partners.

Recommendations for practice

Which CBA practices will work best in describing or supporting learning depends directly on a myriad of contextual and cultural factors, both macro (e.g., policy, curriculum, resources) and micro (e.g., unique classroom settings, individual teachers and students). However, ultimately, it is teachers and students, interacting in individual classrooms, who are the principal agents of learning and best situated to identify effective assessment practices. We are encouraged by the unifying focus on LOA, which should increase the value and use of teachers' evidence of learning to inform policy, decision making, and inferences drawn from external large-scale tests.

Where resources exist, an ePortfolio portal housed on a learning management system can be the locus of interactive, learning-oriented assessment which supports student reflection, self-awareness, and realistic goal setting (Eynon and Gambino, 2018). New software programs and downloadable apps allow teachers to develop *interactive rubrics*. Embedded links can render rubric criteria interactive, as students "click" to access supplementary support for their work as needed (e.g., through automated links to feedback, diagnostic quizzes, benchmarked examples of student writing, mini-lectures, self-assessment checklists).

Future directions

Looking ahead to the future, we see both the extraordinary potential of technologically mediated CBA and the challenges arising that will need to be addressed. New, technologically mediated, non-traditional learning spaces will continue to allow for a myriad of new CBA tools which attend to the multimodal nature of communication and language in use and incorporate embedded or explicit measures to inform teaching and support learning. As noted earlier, continuing advances in data/learning analytics and corpus analysis will directly support and inform the use of dynamic assessment in CBA. However, as a field,

language assessment is only beginning to understand and harness the potential of such virtual learning spaces (see Hartwick, 2018, for a discussion) and further advances in technologically mediated CBA depend directly on the availability of technological resources, time for teacher-initiated and program-initiated innovation, and mandates for both. When used as *showcase portfolios*, ePortfolios provide a digital, archival storage alternative that can be used for AoL/summative assessment of achievement. If retained over time, they can offer a source of formative information for those students who look back at the work they produced from the vantage point of additional study.

However, in addition to the cost and unequal access to technologically mediated CBA approaches, ePortfolios may create additional burdens for language teachers who are underprepared and under-resourced. Teachers may be unable to adequately or actively monitor students' work because of large class sizes and little preparation time. In some cases, teachers may be pressured to provide evidence of student work in portfolios which are subject to external bureaucratic review and be concerned with potential negative consequences for their program, themselves, and their students alike (Fox, 2014). In such contexts, teachers typically select what is to be showcased in a portfolio, undermining the potential for increasing student autonomy and self-awareness that results from students' active participation in the selection of their work.

Further, portfolio assessment may not be well understood by students as the increased personal autonomy and focus on reflection may be unfamiliar to students who have experienced more traditional assessment methods. To prepare students for portfolio assessment, it is critical that teachers have skills to support student understanding, motivate active use, and emphasize the goals of increasing students' self-awareness and reflection. These are challenging and difficult requirements for effective portfolio use (whether technologically mediated or not), and without adequate professional development and support, teachers may feel they are trapped in *portfolio prisons* (Hargreaves *et al*., 2002) with unrealistic and time-consuming demands imposed on them.

Other potential disadvantages of ePortfolios relate to challenges around privacy and surveillance. Some (e.g., McAlpine, 2005; Shohamy, 2001) have expressed concern that digital records housed in ePortfolios may be permanent and could violate a student's rights to privacy. In response to this concern, many educational authorities provide guidelines regarding privacy (e.g., Information and Privacy Commissioner of Ontario, 2019) and promote awareness of legislation which protects a student's personal information.

In the future, we envision an expanded role for CBA as a source for *transdisciplinary* research (Moss, 2016) within an LOA framework (Purpura and Turner, forthcoming). Moss argues for research partnerships that encourage transdisciplinary collaboration across traditional disciplinary boundaries (e.g., Andrade and Brookhart, 2019; Artemeva and Fox, 2010; Fox and Artemeva, forthcoming) and reaffirm collaborative partnerships among researchers and teachers (see Poehner and Inbar-Lourie, 2020), who share a common goal of improving the quality of teaching and learning. Such partnerships are rapidly developing, as the examples here regarding technologically mediated CBA illustrate. Moss (2016) also promotes *methodological pluralism*: "new relationships between measurement specialists, researchers who study professional and organizational learning as well as students' learning, and professionals [teachers] who use test-based and other information in their daily practice" (p. 248). Her perspective supports an expanded role for CBA within the broader all-encompassing framework of LOA, which can provide richer descriptions of assessment and better understanding of its use in supporting learning and closing learning gaps. In addition, the multifaceted nature of classroom assessment evidence (which is by definition,

local, unique, and contextual) can work together with external-to-the-classroom measures: one source of evidence should inform and improve inferences drawn from the other (Fox and Artemeva, forthcoming).

There are early examples of such a relationship in Rea-Dickins (2001), whose research illustrates "classroom assessment as a multifaceted phenomenon with distinct identities linked to learning, teaching and bureaucratic functions" (p. 429). Internal-to-the-classroom CBA purposes and external bureaucratic purposes should *align* in terms of learning goals within the LOA framework. The strength of methodological pluralism, which is increasingly evident in language assessment research (e.g., Jang *et al*., 2014; Moeller *et al*., 2016; Turner, 2014), is that it provides an epistemological and methodological context for a more balanced and more comprehensive approach to the questions language assessment addresses:

- Who is learning?
- What are they learning?
- How much are they learning?
- Why?
- What should we do next to improve their learning?

Following Moss (2016), in order to better address these questions in the future, we argue for (1) increasing use of CBA evidence to extend and clarify inferences drawn from system-wide external tests; (2) the elaboration of more comprehensive theoretical frameworks to inform CBA practices (Purpura and Turner, forthcoming); and (3) increased use of social theories and methodological pluralism in framing, conducting, and interpreting language assessment research (Fox and Artemeva, forthcoming). As Pellegrino *et al*. (2001) foresaw, alignment requires that the contextually rich findings from CBA are systematically related to those of external, large-scale tests. Taking all available evidence into account will result in inferences about learning that are more "robust" (Moss, 2016: 248): that is, more defensible, appropriate, meaningful, and useful. Increased valuing and use of teachers' evidence of learning can (and should) inform policy and decision making.

Further reading

Cheng, L. and Fox, J. (2017). *Assessment in the Language Classroom*. New York, NY: Palgrave Macmillan.

This practical resource for language teachers provides an overview of assessment practices which support learning. Informed by empirical research, it combines detailed explanations with hands-on examples from classroom assessment experience.

Hartwick, P. and Nowlan, N. S. (2018). Integrating virtual spaces: Connecting affordances of 3D virtual learning environments to design for twenty-first century learning. In Y. Qian (ed.), *Integrating Multi-User Virtual Environments in Modern Classrooms*. Hershey, PA: ICI Global, 111–36.

Hartwick and Nowlan examine the importance of social interaction, task design, and teacher feedback, given the *affordances* of online learning.

Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice* 23: 236–51. https://doi.org.10.1080/0969594X.2015.1072085.

This landmark article proposes a long-term, transdisciplinary research agenda "to support routine conceptual uses" (p. 248) of external test data and suggests strategies for using such "data well" (p. 248), through evidence-rich, empirical partnerships that connect testers with test users (e.g., teachers, administrators, policy makers).

Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin, Germany and Boston, MA: DeGruyter Mouton.

Turner and Purpura propose a multi-dimensional framework for learning oriented assessment (LOA) that prioritizes learning when considering the interrelationships across instruction, assessment, and learning.

References

Abdulhamid, N. (2018). *What Is the Relationship Between Alignment and Washback? A Mixed-Methods Study of the Libyan EFL Context*. Unpublished doctoral dissertation, Carleton University Ottawa, Canada.

Abdulhamid, N. and Fox, J. (2020). Portfolio based language assessment (PBLA) in language instruction for newcomers to Canada (LINC) programs: Taking stock of teachers' experience. *Canadian Journal of Applied Linguistics* 23: 168–92. https://doi.org/10.37213/cjal.2020.31121.

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. London, UK: Continuum.

Alderson, J. C., Brunfaut, T. and Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics* 36: 236–60. https://doi.org/10.1093/applin/amt046.

Andrade, H. L. and Brookhart, S. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice* 27: 350–72. https://doi.org/10.1080/0969594X.2019.1571992.

Andrade, H. L. and Cizek, G. J. (2010). *Handbook of Formative Assessment*. New York, NY: Routledge.

Artemeva, N. and Fox, J. (2010). Awareness vs. production: Probing students' antecedent genre knowledge. *Journal of Business and Technical Communication* 24: 476–515. https://doi.org/10.1177%2F1050651910371302.

Assessment Reform Group. (2002). *Assessment for Learning: 10 Principles*. London, UK: Assessment Reform Group. Retrieved from www.aaia.org.uk/storage/medialibrary/o_1d8j89n3u1n0u17u91fdd1m4418fh8.pdf.

Barrett, H. C. (2007). Researching electronic portfolios and learner engagement: The REFLECT initiative. *Journal of Adolescent & Adult Literacy* 50: 436–49. https://doi:10.1598/JAAL.50.6.2.

Biggs, J. and Tang, C. (2011). *Teaching for Quality Learning*. Berkshire, UK: McGraw Hill.

Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A. and Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education* 13: 117–40. https://doi.org/10.1177/1478210314566733.

Black, P. and Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan* 92: 81–90. https://doi.org/10.1177/003172171009200119.

Blazer, C. (2011). *Unintended Consequences of High-Stakes Testing: Information Capsule, Volume 1008: Research Report for the Miami-Dade County Public Schools Research Services*. Retrieved from https://archive.org/details/ERIC_ED536512.

Bloom, B. S. (ed.). (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals*. New York, NY: McKay.

Breen, M. (1987a). Contemporary paradigms in syllabus design: Part I. *Language Teaching* 20: 81–92. https://doi.org/10.1017/S0261444800004365.

Breen, M. (1987b). Contemporary paradigms in syllabus design: Part II. *Language Teaching* 20: 157–74. https://doi.org/10.1017/S026144480000450X.

Bygate, M., Skehan, P. and Swain, M. (eds.). (2001). *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. Harlow, UK: Longman.

Callies, M. and Götz, S. (eds.). (2015). *Learner Corpora in Language Testing and Assessment*. Amsterdam, The Netherlands: John Benjamins.

Chalhoub-Deville, M. (2016). Validity theory: Reform policies, accountability testing, and consequences. *Language Testing* 33: 453–72. https://doi.org/10.1177/0265532215593312.

Chapelle, C. A. and Sauro, S. (eds.). (2017). *The Handbook of Technology and Second Language Teaching and Learning*. Hoboken, NJ: John Wiley & Sons, Inc.

Chatterji, M. (ed.). (2013). *Validity and Test Use: An International Dialogue on Educational Assessment, Accountability and Equity*. Bingley, UK: Emerald Group.

Cheng, L. and Fox, J. (2017). *Assessment in the Language Classroom*. New York, NY: Palgrave Macmillan.

Cheng, L., Watanabe, Y. and Curtis, A. (2004). *Washback in Language Testing: Research Contexts and Methods*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. (2010). An introduction to formative assessment: History, characteristics and challenges. In H. L. Andrade and G. J. Cizek (eds.), *Handbook of Formative Assessment*. New York, NY: Routledge, 3–17.

Curtis, A. (2017). *Methods and Methodologies for Language Teaching*. London, UK: Palgrave Macmillan.

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly* 12: 110–35. https://doi.org/10.1080/15434303.2014.1002925.

Earl, L. (2013). *Assessment as Learning: Using Classroom Assessment to Maximize Student Learning*, 2nd edn. E-reader version. Thousand Oaks, CA: Corwin Press.

Elbow, P. and Belanoff, P. (1986). Portfolios as a substitute for proficiency examinations. *College Composition and Communication* 37: 336–39. Retrieved from www.jstor.org/stable/358050.

Eynon, B. and Gambino, L. (2018). *Catalyst in Action: Case Studies of High Impact EPortfolio Practice*. Sterling, VA: Stylus Publishing.

Fox, J. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes* 8: 26–42. https://doi.org/10.1016/j.jeap.2008.12.004.

———. (2014). Portfolio based language assessment (PBLA) in Canadian immigrant language training: Have we got it wrong? *Contact* 40: 68–83. Retrieved from www.teslontario.org/uploads/publications/researchsymposium/ResearchSymposium2014.pdf.

Fox, J. and Artemeva, N. (2017). From diagnosis toward academic support: Developing a disciplinary, ESP-based writing task and rubric to identify the needs of entering undergraduate engineering students. *ESP Today* 5: 148–71. https://doi.org/10.18485/esptoday.2017.5.2.2.

———. (forthcoming). *Reconsidering Context in Language Assessment: Transdisciplinary Perspectives, Social Theories, and Validity*. New York, NY: Routledge.

Fox, J. and Hartwick, P. (2011). Taking a diagnostic turn: Reinventing the portfolio in EAP classrooms. In D. Tsagari and I. Csépes (eds.), *Classroom-Based Language Assessment*. Frankfurt, DE: Peter Lang, 47–62.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13: 208–38. https://doi.org/10.1177%2F026553229601300205.

Gipps, C. V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment*. London: The Falmer Press.

Hamp-Lyons, L. and Condon, W. (2000). *Assessing the Portfolio: Principles for Practice, Theory, and Research*. Cresskill: Hampton.

Harding, L., Alderson, J. C. and Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing* 32: 317–36. https://doi.org/10.1177%2F0265532214564505.

Hargreaves, A., Earl, L. and Schmidt, M. (2002). Perspectives on alternative assessment reform. *American Educational Research Journal* 39: 69–95. https://doi.org/10.3102%2F00028312039001069.

Harlen, W. and Winter, J. (2004). The development of assessment for learning: Learning from the case of science and mathematics. *Language Testing* 21: 390–408. https://doi.org/10.1191%2F0265532204lt289oa.

Hartwick, P. (2018). *Exploring the Affordances of Online Learning Environments: 3DVLs and ePortfolios in Second Language Learning and Teaching*. Unpublished doctoral dissertation, Carleton University, Ottawa, Canada.

Hartwick, P. and Nowlan, N. S. (2018). Integrating virtual spaces: Connecting affordances of 3D virtual learning environments to design for twenty-first century learning. In Y. Qian (ed.), *Integrating Multi-User Virtual Environments in Modern Classrooms*. Hershey, PA: ICI Global, 111–36.

Heitink, M. C., Van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K. and Kippers, W. B. (2016). A systematic review of prerequisite for implementing assessment for learning in classroom practice. *Educational Research Review* 17: 50–62. https://doi.org/10.1016/j.edurev.2015.12.002.

Hill, K. and McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing* 29: 395–420. https://doi.org/10.1177%2F0265532211428317.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride and J. Holmes (eds.), *Sociolinguistics: Selected Readings*. Harmondsworth, UK: Penguin, 269–93.

Information and Privacy Commissioner of Ontario. (2019). *A Guide to Privacy and Access to Information in Ontario Schools*. Retrieved from www.ipc.on.ca/wp-content/uploads/2019/01/guide-to-privacy-access-in-ont-schools.pdf.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for fusion model application to LanguEdge assessment. *Language Testing* 26: 31–73. https://doi.org/10.1177%2F0265532208097336.

Jang, E. E. and Wagner, M. (2013). Diagnostic feedback in language classroom. In A. Kunnan (ed.), *Companion to Language Assessment*. Hoboken, NJ: Wiley-Blackwell, 693–711.

Jang, E. E., Wagner, M. and Park, G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics* 34: 123–53. https://doi.org/10.1017/S0267190514000063.

Kempf, A. (2016). *The Pedagogy of Standardized Testing: The Radical Impacts of Educational Standardization in the US and Canada*. New York, NY: Palgrave Macmillan.

Lee, Y. W. (2015). Future of diagnostic language assessment. *Language Testing* 32: 295–98. https://doi.org/10.1177%2F0265532214565385.

Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice, and change. *Language Assessment Quarterly* 1: 19–41. https://doi.org/10.1207/s15434311laq0101_3.

Leung, C. and Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing* 21: 335–59. https://doi.org/10.1191%2F0265532204lt287oa.

Little, D. (2005). The common European framework and the European language portfolio: Involving learners and their judgements in the assessment process. *Language Testing* 22: 321–36. https://doi.org/10.1191%2F0265532205lt311oa.

Macqueen, S., Knoch, U., Wigglesworth, G., Nordlinger, R., Singer, R., McNamara, T. and Brickle, R. (2019). The impact of national standardized literacy and numeracy testing on children and teaching staff in remote Australian indigenous communities. *Language Testing* 36: 265–87. https://doi.org/10.1177%2F0265532218775758.

McAlpine, M. (2005). E-portfolios and digital identity: Some issues for discussion. *E-Learning* 2: 378–87. https://doi.org/10.2304%2Felea.2005.2.4.378.

Moeller, A. J., Creswell, W. and Saville, N, (eds,). (2016). *Second Language Assessment and Mixed Methods Research*. Cambridge, UK: Cambridge University Press.

Moss, P. A. (1994). Can there be validity without reliability? *Educational researcher* 23: 5–12. https://doi.org/10.3102%2F0013189X023002005.

———. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice* 23: 236–51. https://doi.org/10.1080/0969594X.2015.1072085.

Nunan, D. (1988). *The Learner-Centred Curriculum*. Cambridge, UK: Cambridge University Press.

Panadero, E., Andrade, H. and Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: A roadmap of where we are, how we got here, and where we are going. *Australian Educational Researcher* 45: 13–31.https://doi.org/10.1007/s13384-018-0258-y.

Pellegrino, J. W., Chudowsky, N. and Glaser, R. (eds.). (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press.

Poehner, M. E. and Inbar-Lourie, O. (eds.). (2020). *Toward a Reconceptualization of Second Language Classroom Assessment: Praxis and Researcher-Teacher Partnership*. Cham, CH: Springer.

Poehner, M. E. and Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research* 17: 323–42. https://doi.org/10.1177/1362168813482935.

Prabhu, R. N. S. (1987). *Second Language Pedagogy*. Toronto, ON: Oxford University Press.

Purpura, J. E. (2019). *Questioning the Currency of Second and Foreign Language Certification Exams*. Paper presented at the A&H Distinguished Lecture Series at Teacher College, Columbia University, New York, NY. Retrieved from www.tc.columbia.edu/arts-and-humanities/ah-distinguished-speaker-series/.

Purpura, J. E. and Turner, C. (forthcoming). *Learning-Oriented Assessment in Language Classrooms: Using Assessment to Gauge and Promote Language Learning*. London, UK: Taylor & Francis Ltd.

Read, J. (ed.). (2016). *Post-Admission Language Assessment of University Students*. Cham: Springer International.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing* 18: 429–62. https://doi.org/10.1111/j.1473-4192.2006.00112.x.

———. (2006). Currents and eddies in the discourse of assessment: A learning focused interpretation. *International Journal of Applied Linguistics* 16: 164–88. https://doi.org/10.1111/j.1473-4192.2006.00112.x.

Savignon, S. J. (1987). *Initiatives in Communicative Language Teaching II: A Book of Readings*. Reading, MA: Addison-Wesley.

Savin-Baden, M. (2008). *Learning Spaces: Creating Opportunities for Knowledge Creation in Academic Life*. New York, NY: McGraw-Hill, Society for Research into Higher Education & Open University Press.

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne and M. Scriven (eds.), *Perspective Son Curricular Evaluation*. Chicago, IL: Rand-McNally, 39–83.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. London, UK: Pearson.

Sultana, N. (2019). *An Exploration of Alignment of the Secondary School Certificate (SSC) English Examination with Curriculum and Classroom Instruction: A Washback Study in the Context of Bangladesh*. Unpublished doctoral dissertation, Queen's University, Kingston, Canada.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics* 29: 21–36. https://doi.org/10.1017/S0267190509090035.

Tishman, S., Jay, E. and Perkins, D. N. (1993). Teaching thinking dispositions: From transmission to enculturation. *Theory into Practice* 32: 147–53. Retrieved from www.jstor.org/stable/1476695.

Torrance, H. (2015). Blaming the victim: Assessment, examinations, and the responsibilisation of students and teachers in neo-liberal governance. *Discourse: Studies in the Cultural Politics of Education* 38: 83–96. https://doi.org/10.1080/01596306.2015.1104854.

Turner, C. E. (2009). Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal on Pedagogies and Learning* 5: 103–23. https://doi.org/10.5172/ijpl.5.1.103.

———. (2012). Classroom assessment. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. New York, NY: Routledge, 65–78.

———. (2014). Mixed methods research. In A. J. Kunnan (ed.), *The Companion to Language Assessment*. Chichester, UK: John Wiley & Sons Ltd., 1403–17. doi:10.1002/9781118411360.wbcla142.

Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin, Germany and Boston, MA: DeGruyter Mouton.

Upshur, J. A. and Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal* 49: 3–12. https://doi.org/10.1093/elt/49.1.3.

Van den Branden, K. (2016). The role of teachers in task-based language education. *Annual Review of Applied Linguistics* 36: 164–81. doi:10.1017/S0267190515000070.

van Lier, L. (2000). From input to affordance: Social-interactive learning from an ecological perspective. In J. P. Lantolf (ed.), *Sociocultural theory and second language learning*. Oxford, UK: Oxford University Press, 245–59.

Van Viegen Stille, S., Bethke, R., Bradley-Brown, J., Giberson, J. and Hall, G. (2016). Broadening educational practice to include translanguaging: An outcome of educator inquiry into multilingual students' learning needs. *Canadian Modern Language Review* 72: 480–503. https://doi.org/10.3138/cmlr.3432.

Vogt, K. and Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly* 11: 374–402. https://doi.org/10.1080/15434303.2014.960046.

Wiggins, G. and McTighe, J. (2005). *Understanding by Design*. Alexandria, VA: Association for Supervision and Curriculum Development.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation* 37: 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001.

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich and M. Zeidner (eds.), *Handbook of Self-Regulation*. San Diego, CA: Academic Press, 13–39. doi:10.1016/b978-012109890-2/50031-7.

8

Washback

Looking backward and forward

Liying Cheng and Nasreen Sultana

Introduction: definitions and review scope

Washback, the influence of testing on teaching and learning, has been a consistent research area in language testing since the seminal work of Alderson and Wall (1993) and Wall and Alderson (1993). This growing trend resonates with the global shift from large-scale testing to classroom assessment (Volante *et al*., 2020; see also Fox *et al*., this volume). As a result, a large body of empirical work has been conducted worldwide (see Cheng *et al*., 2015 for a recent review), especially in countries where English is not the dominant language. This chapter shifts the focus from research conducted in Europe and East Asia to washback studies conducted and contextualised in South Asia and the Middle East, where we have witnessed increasing empirical work over the past ten years.

We live in an age when test results are used for multiple purposes globally. Tests are not only gatekeepers; they are also associated with students' progress, growth, opportunities, and often status. Further to this, test results are also used to measure the effectiveness of teachers, schools, and educational interventions. Tests with greater power have more significant consequences for the lives of students, teachers, and other stakeholders. This consequential nature of testing is known as washback (Alderson and Wall, 1993; Bailey, 1996; Cheng, 1997, 2005; Hughes, 1993). The rise of washback is indebted to the long-standing "encroaching power" (Latham, 1877: 2) associated with high-stakes examinations in human history. Among the range of terms that has been used to refer to the influence of tests, such as "test impact" (Bachman and Palmer, 1996), "consequential validity" (Messick, 1989), "measurement-driven instruction" (Popham, 1987), "curriculum alignment" (Shepard, 1990), and "backwash" (Biggs, 1995), *washback* is the one term which has been sustained in the literature for both language testing and educational assessment.

This chapter, as a way of looking at washback both backward and forward, focuses on washback studies that were conducted between 2011 and 2019. This time period was chosen because the first version of this chapter by Wall (2012) covered studies until 2011. We broadly searched research databases for two types of washback studies according to their publication formats: empirical research articles and recent doctoral dissertations. The

 DOI: 10.4324/9781003220756-12

selection of the second type of studies was based on two criteria: (1) studies for which full texts were available in English and (2) studies that contained conceptual or theoretical frameworks with explicitly stated methodology. Studies before 2011 were only included when the studies were historically and empirically significant in washback research.

Historical perspectives

Washback research started with Alderson and Wall's (1993) pioneering publication "Does washback exist?" in which the authors offered 15 hypotheses regarding the potential aspects of teaching and learning that could be influenced by testing. These hypotheses have been tested by washback research studies over the past three decades. Broadly, these hypotheses discussed three types of test influence: the influence of tests on teaching and learning, the influence of tests on teachers and students, and the variable influence of washback depending on the significance of test consequences. The bulk of washback research that took place in the post-1993 era was primarily based on these hypotheses (Cheng, 2005; Green, 2007; Wall, 2005). Around the same time period, Hughes (1993) suggested three vital components: that is, the "participants," the "processes," and the "products" of an educational system that are responsible for producing washback. Based on Hughes's (1993) trichotomy and Alderson and Wall's (1993) 15 washback hypotheses, Bailey (1996) developed an operational washback model illustrating not only that the test stimulates the products (e.g., teaching, learning, teaching materials) through the participants and the processes of teaching and learning, but also that feedback provided by participants may, in turn, bring changes in the test.

Bailey's (1996) model turned out to be one of the most powerful conceptual frameworks over many years of washback research. This model visualized the non-linear relationship among the education components and illustrated the complexities that washback research carries forward. However, Bailey did not elaborate on the "processes" between products and participants. Cheng later applied Bailey's washback model and created her own exploratory model to study the Hong Kong Certificate of Education Examination (HKCEE) (1997, 1998, 1999). Cheng is one of the earlier researchers who realized that gaps in understanding among key stakeholders concerning the relationships between testing, teaching, and learning created unintended washback on classroom teaching and learning. Her influential work indicated that, although the test triggered an immediate change in teaching materials, it failed to bring about changes in teaching methodology. What remained a question in Cheng's study was *why* teachers were reluctant to make changes to their teaching methodology, a question which even now has implications for teacher education and teacher assessment literacy. This prompted a larger question: Why do some elements of teaching and learning remain immune to the influence of tests? This question was examined from various perspectives in subsequent washback studies around the world (Cheng *et al*., 2004). Later Green's (2007) three-dimensional model (direction, variability, and intensity) addressed the void in the washback literature on students' learning and offered a comprehensive model focusing on learners and learning linking back to test validity. Green's model remains, to this date, the most dominant learner-centered washback model.

Washback studies in language testing began with an agenda to explore how *new* English national examinations were successful in bringing about desired changes in classroom teaching and learning (Alderson and Wall, 1993; Shohamy *et al*., 1996; Wall, 2005) –

"changing teaching and learning through language testing" (as per the book title in Cheng, 2005). Gradually the research expanded to study test impact of international tests, such as the Test of English as a Foreign Language (TOEFL) (Alderson and Hamp-Lyons, 1996), the International English Language Testing System (IELTS) (Green, 2007), and university entrance examinations in Japan (Watanabe, 1996) and in China (Qi, 2005). In subsequent years, many studies have been carried out examining the washback effects of various *existing* examinations at the national and international levels.

Critical issues: washback complexities and validity

More recently, a new generation of washback researchers (e.g., Booth, 2018; Sultana, 2019a; Sun, 2016) have suggested studying washback by using validation frameworks to make washback research more viable and closely linked to the core nature of tests and test uses. This direction has been reiterated over the years by Cheng (2008, 2014) and Cheng *et al*. (2015). Messick (1996) was the first scholar who made the explicit connection between washback and validity for the field of language testing. Almost three decades of washback studies have since proved that a wide range of factors work together to bring about washback, and the actions of teachers and students in achieving washback are intrinsically connected with the test itself. Washback is not only limited to its connection with validity, but it is also vital to track the nature and scope of washback studies to understand the unpredictable and complex nature of the testing system.

Early washback studies unpacked a wide range of complexities in washback research, but washback was considered independently of the phenomenon of test validity (Bachman, 2005; Cheng, 2008). Bachman (2005) observed that "the extensive research on validity and validation has tended to ignore test use, on the one hand, while discussions of test use and consequences have tended to ignore validity, on the other" (p. 7). Many early researchers, including Alderson and Wall (1993), considered washback too complex to be considered as a component of test validation. Tsagari and Cheng (2017) challenged researchers to prove the connection between validity and consequences theoretically and empirically. The relationship between washback and validity is not straightforward (Sultana, 2019a; Wang, 2018; Xie, 2010). There may not be a linear relationship between a test and teaching and learning as tests may influence other mediating factors in the educational system and vice versa.

In a recent interview with Brunfaut, Alderson reiterates,

> I don't think washback is necessarily caused by the test. It is the use or misuse of the test. . . . What is important is who uses the test and how they use it, and whether it is used for high or low stakes.
>
> *(Brunfaut, 2014: 112–13)--*

The argument on the use or misuse of tests, however, implies that washback *is* an essential component of validity. If a test is used to guide teaching and learning, the ultimate purpose is to create better test performance that is associated with the score and the performance. Going back to the early discussion of the consequences of test use, Cronbach (1971) counted negative consequences as a threat to validity even if they were not produced due to a flaw in a test. Later, Messick (1989, 1996) narrowed down the role of consequences by connecting it with the test design, which makes more sense from a washback perspective. Similarly, Kane's (2006, 2013) interpretation/use argument (IUA) framework accepts test uses as an essential validity issue, which Alderson also acknowledged in his interview (see

also Kane, this volume). The evolution of washback research has led us to see the influence of tests using a range of lenses, yet those influences have always been connected to the test itself. This connection between washback and validity, however, has the potential to be very fruitful as argued by Bachman (2005), and as Messick has asserted, "rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback" (1996: 252). This is why recent washback researchers (Anand, 2018; Booth, 2018; Sultana, 2019a; Sun, 2016) have advocated washback research using validity frameworks indicating a corollary that examining washback and its impact on teaching and learning is a process leading to the test validation.

Most recently, in the introduction to *Language Testing*'s first virtual special issue: "Investigating Consequences of Language Test Use," Chapelle (2020: 6) argues that

> It remains to be seen whether future studies of consequences will be interpreted with respect to validity, but in view of the findings presented in these articles [included in this special issue] investigating language tests throughout the world, it seems clear that consequences will remain an important area of investigation in language testing.

After nearly three decades of research, the controversy regarding "whether and how consequences should be integrated into test validation or even whether they belong to test validation or not" (Tsagari and Cheng, 2017: 11) has shifted to the focus of finding the right validation framework to study washback.

Current contrubitions and empirical studies

Post-1993 washback research primarily examined (1) how teaching and learning were influenced by tests and (2) the different factors which produced positive and/or negative washback. Three major shifts in washback research have been noticed within the last three decades of research. First, an increasing number of washback studies have been conducted in the context of South Asian countries and the Middle East, which was not the case previously. Many of these studies have focused on the issues of "alignment" and "misalignment": i.e., the degree to which curriculum standards and policy, classroom teaching, and testing are in agreement with one another (Webb, 2007). Second, many of these recent studies have explored how the washback effect is directly or indirectly linked with the testing system and other contextual factors of the education system. Third, more recently, and especially in the last few years, researchers have adopted a wide range of conceptual and theoretical frameworks to study factors nested in teaching and learning which could induce washback. The following sections elaborate the aforementioned three points.

Alignment and misalingment

Recent research on the phenomenon of washback has continued to demonstrate the complexity of factors that determine outcomes. Umashankar's (2017) doctoral research, for example, examined the washback effect of a new system of English language speaking assessment in Sri Lanka. Because of an examination reform, teachers' and students' perceptions of teaching and learning and teachers' instructional practices changed. In this case, the washback was positive; many of the amendments following the introduction of the new speaking tests were intended. However, Froetscher's (2016) study examining the washback on classroom assessment in the context of the reformed Austrian secondary school leaving examination

raised concerns about the extent to which the intended washback outcomes were achieved. The study reported how teachers instantly changed their class tests, mirroring the reformed examination but not the underlying the curriculum envisioned by the policymakers. Similarly, Bhattarai (2019) studied the multiple facets of the recently reformed School Leaving Certificate (SLC) examination in Nepal, which indicated that the visions of the policymakers were not well matched with the perceptions of the users. Froehlich (2016) examined the newly introduced English oral test at the secondary level in German aiming at achieving communicative teaching and learning. She found that although the test created positive washback in teaching methodology, materials, and learning in general, it failed to achieve overall positive feedback from students and teachers. A crucial factor behind the disappointment could be the "top-down approach" of introducing a new test without consulting the teachers, as pointed out by the researcher, and that the "test designer had forgotten what it was like to be in a classroom" (p. 175). This concern was also pointed out earlier in Cheng's (2005) study and later by Bhattarai (2019) and Sultana (2019a): the intended washback was not achieved because there were gaps in the process of policy implementations.

Even though washback research was traditionally believed to focus on the impact of tests on teaching and learning, most washback studies have pointed out how inconsistencies and misalignments within the educational system hindered the expected washback on teaching and learning. This indicates that washback researchers need to see washback research as a way in which to investigate the impact of tests on teaching and learning while also considering the larger educational, contextual, and social effects.

In connection with this point, Sultana (2019a) studied an existing secondary school leaving English public examination in Bangladesh, which has been in operation for the last two decades. The study explored the extensive, intricate, and complex facades of washback. Contextual and bureaucratic factors were found to be primarily responsible for the existing misalignment within the educational components, which were in turn responsible for creating unintended washback on teaching and learning. Specifically, classroom teaching and learning were primarily shaped by the expectations of the examination, which was misaligned with the expectations of the curriculum and the textbook. On the surface level, it seemed that negative washback was the direct outcome of the examination and an extremely exam-driven approach to classroom instruction. Another study conducted by Abdulhamid (2018) in the context of Libya also reached the conclusion that lack of alignment between the examination and the curriculum standards not only produced negative washback but also possibly misled key stakeholders about the uses and expectations of the examination.

Earlier research demonstrated that the extent to which assessment goals and curriculum goals are aligned can constrain or promote innovative practices. For example, Chen (2002), in an investigation of curriculum innovation in a junior high school EFL context in Taiwan, found misalignment between the test and the curriculum. The MCQ-based test content failed to address the communicative objectives of the curriculum. Thus, the targeted communicative goals of the curriculum were not achieved through the test. The author suggested that "positive washback effects are more likely to occur when a curriculum and a test are highly matched" (p. 5). This assertion was echoed in Tan and Turner's (2015) study on the context of high-stakes secondary five ESL exit writing examinations in Quebec, Canada. They found that, since the examination closely reflected the requirements of the curriculum, the intended washback effect was achieved. On the other hand, Fox and Cheng's (2007) study on the Ontario Secondary School Literacy Test (OSSLT) in Ontario, Canada, reported that the test failed to address classroom practices and curricular goals and caused increased anxiety among the second language test takers.

Most of the studies conducted in South Asian countries have persistently pointed out existing misalignment in the education system as the primary cause of negative washback (see Sultana, 2018). For example, Rind and Mari's (2019) recent study investigating the higher secondary state-level English examination in Pakistan noted that while the learning and teaching of English, the test preparation materials, and the English examination were in alignment, they largely contradicted the objectives of the English curriculum of the country. The same was found in Hoque's (2011) doctoral study in Bangladesh exploring the washback of the national higher secondary English examination on teaching and learning. Hoque's research investigated how misalignment within the education system exerted negative washback on the curriculum, teaching, learning, and overall English education in the country. For example, the policy makers aimed at developing the overall English proficiency of the learners, which was not possible to achieve within the context because of the way English was taught and learned in and out of classrooms and the way textbooks were set up. Mathew's (2012) study in India, based on document analysis, revealed that there was non-conformity between the underlying principles of the assessment scheme of the Central Board of Secondary Education (CBSE) examination and teachers' principles of classroom teaching. The CBSE English examination acted as a de facto curriculum in classroom instruction since teachers tried to teach the examination without perceiving the underlying assumptions of the examination.

These studies suggest that in this South Asian region, there is often a high rate of misalignment between curriculum, teaching and learning, and the associated examination. In this case, considering the extreme importance that examinations have in these countries, teaching and learning are molded by the examination expectations. Teachers and students do what they need to do to pass the examination, which is the direct, and only, route to higher education, better jobs, and better prospects in life.

Contextual factors within the testing and education system

It is empirically proven that there exists a powerful relationship between language teaching and learning and language testing, and thus, there exists washback, be it positive or negative, intended or unintended, neutral or active (Alderson and Wall, 1993; Bailey, 1996; Cheng *et al*., 2004; Hughes, 1993; Messick, 1996). However, washback researchers continue to be challenged with the multi-layered and multi-dimensional nature of washback, which does not let us pinpoint the exact factors which may promote or hinder washback within the testing and education system. Two layers of interactions at the education system level and at the testing system level are discussed next.

Recent studies on washback, especially those conducted in the South Asian countries, have revealed that contextual factors, cultural understanding of the examinations and the uses of scores, limited resources, gaps in the policy level, and lack of teacher training interact to create negative washback on teaching and learning. Sultana's (2019a) study indicated that washback researchers need to explore how stakeholders view the examination and the uses of its scores. In this study, washback took a more intricate route because it was not accidental; rather, students and teachers worked to bring about the washback, which was unintended from the policy makers' point of view or from the perspective of the curriculum goals. In this context, obtaining an examination score was considered more important than how much students were able to perform or learn. In addition, socio-political forces can contribute to the washback effect. Political instability in Bangladesh caused delays in delivering

instructional materials, which, by extension, contributed to the unintended washback on classroom instruction.

Abdulhamid's (2018) study in Libya and Bhattarai's (2019) study in Nepal also explored the complex sociocultural complexities that exist in these countries where situations are not ideal. In these studies, even though the national examination worked as the key trigger, many other financial, political, and social issues were interconnected in the education system, which indirectly or directly influenced washback on learning and teaching. The studies indicated the influence of politics on educational systems in the respective countries in South Asia and the Middle East, signaling the possibility of a more complex washback situation at stake. This raises a question: Does washback look different in a political context of unrest? Although several studies (e.g., Khaniya, 1990; Memon, 2015; Onaiba, 2013; Sultana, 2019a) have mentioned the adverse effects of political unrest during their data collection phases, existing literature contains very few studies systematically documenting the impact of political unrest on washback. This issue overlaps with other recent discussions in language testing: e.g., the relationship between politics and test consequences (Shohamy, 2001), geopolitics of assessment (Cheng, 2018), uses of test as instruments of policy (McNamara, 2010), politics and citizenship testing (Kunnan, 2009), and the politics behind centralized examinations (Marshall, 2017). In discussing the consequences of test use, Alderson emphasised how educational testing is a political issue (Brunfaut, 2014).

As seen earlier, political unrest can sometimes directly affect classroom practice (e.g., Sultana's study). In other cases, unrest can create conditions in which negative washback is more likely to occur. For example, within the context of political unrest in Pakistan, Memon's (2015) study indicated that there was a heightened interest in taking IELTS as a way to emigrate (due to political instability) or as a way to secure a job or to gain higher status in the society. As a result, misuse of and misperceptions about the test were prevalent within the context of the IELTS preparation courses in Pakistan that Memon studied. Similar results were found in Sun's study (2016) where the CET test score is used for obtaining city residence in China.

With the expansion of washback research, scholars have started examining washback from a wider and more in-depth range of perspectives: for instance, students' motivation and characteristics of the students in relation to test preparation. Ma (2019) used Green's (2007) washback model to understand how students' characteristics were related to test preparation characteristics and students' performance on the College English Test Band 4 (CET4). The findings suggested strong washback, which was due to the significant relationships identified between students' characteristics and CET4 test scores. As an example, in Ma's study, students from cities had better scores than students from remote regions because the former group of students had access to more test preparation resources. This was also one of the key findings in Sultana's (2019a) study; more access to resources may intensify the washback effect. Ma's study articulated the ethical concerns of test preparation programs and concluded with a call for further research to investigate if test preparation ensures higher scores.

More specifically, there remains a question over whether or not test constructs play a significant role in students' test preparation. A recent study conducted by Kim (2017) on washback of TOEFL writing on students' learning raised the question of whether students learn academic writing by preparing for the test. Based on data collected from over 150 pages of anonymous online forum postings made by Korean TOEFL test takers from two websites, Kim found that students had improved their test scores by memorizing the templates provided by test preparation centers, which created a threat to the validity of the test itself. Kim's study indirectly raised concerns about the relationship between test design,

student performance, and interpretations of that performance (see also Ma and Cheng, 2016 on TOEFL test preparation). Shih's (2007) study found that students were keener on practicing listening and reading skills in GEPT in Taiwan as these skills were predominantly tested in the first level of the examination. This led Shih to conclude that "the nature of tested skill and the structure of the test affect the degree of washback on students' learning" (p. 143), which complemented one of the recent washback studies conducted in the context of an entrance examination in Japan by Sato (2019). Sato found that students' test preparation was closely based on the test constructs and how they perceived those constructs.

While exploring test preparation, Sato (2019) also identified a specific learning behavior of the students associated with their self-efficacy. Sato revealed that students with higher levels of self-efficacy created customized test preparation strategies which suited their individual learning needs. These findings were echoed by Pan's (2016) and Pan and Newfields's (2013) studies in Taiwan, and Sun's (2016) study in China, investigating learner factors of motivation. Sun explored the CET4's powerful influence in creating extrinsic motivation to prepare for the test, which turned out to be one of the unintended consequences of the test. These studies significantly pointed out that "tests are not a panacea that will always succeed in promoting student learning" (Pan and Newfields, 2013: 12). It would, therefore, appear that tests probably influence test preparation, but not motivation toward learning. Sun (2016) pointed out the possibility of better quality learning by improving test design. These studies make us aware that tests should be designed more carefully if we want to achieve positive test washback on learning.

Emerging conceptual and theoretical frameworks

Washback researchers have been challenged by the multi-layered and multi-dimensional nature of washback. The adoption of strong conceptual and theoretical frameworks is essential for washback researchers to unpack the nature and scope of washback. Tsagari and Cheng (2017) indicated that washback researchers can replicate previously used frameworks to examine a different test in a different educational context. For example, almost a decade after Cheng's (2005) application of an exploratory research design (a design employed when there are a limited number of studies available on an existing research problem) in examining the washback effect of an English examination in Hong Kong, Umashankar (2017) employed a similar exploratory framework in his doctoral study investigating an English assessment in Sri Lanka. Among washback researchers, the use of exploratory frameworks has been rare, even though this type of framework allows researchers to identify the existing gaps in an educational system. Another example of adaptation of a previously used framework is Vyn's (2019) study. Following Wall's (2005) washback research, which used innovation theory, Vyn examined the washback effect of the American Council on the Teaching of Foreign Languages (ACTFL) Assessment of Performance toward Proficiency in Languages (AAPPL), drawing on the theories of educational innovation. Vyn's investigation was primarily situated in the area of curricular innovation: that is, "a managed process of development whose principal products are teaching (and/or testing) materials, methodological skills, and pedagogical values that are perceived as new by potential adopters" (Markee, 1997: 172). This approach defined innovation as the agent to bring the desired change. Based on online surveys, semi-structured interviews, and test takers' scores, this mixed-method study contributes to understanding the use of external language assessments to improve K–12 contexts. Relatedly, Anand (2018) employed diffusion of innovation theory (Fullan, 2015; Henrichsen, 1989; Markee, 1997) in researching washback resulting

from the implementation of a new test within an EAP program at a Canadian university. She borrowed some of the key ideas of innovation theories, such as charting the process and meaning of change, gathering different perspectives, strategies and approaches in the process of the change, and understanding the context and participants in the change process. Traditionally, in washback studies, innovation theories have been in use when researchers investigate a newly implemented examination to explore the changes in the system.

Sultana's (2019a) doctoral study examined washback from a curriculum alignment perspective: that is, determining to what extent the curriculum, the textbook, and the national English examination in Bangladesh were (mis)aligned and if that (mis)alignment produced any washback on teaching and learning. Sultana used insights from the argument and evidence-based validation framework (Pellegrino *et al*., 2016) to analyse the data she obtained from document analysis, interviews with teachers and the test developer, classroom observations, and focus groups with the test candidates. Pellegrino and his colleagues developed the argument and evidence-based framework in the context of a classroom-based mathematics assessment. In adopting this framework, Sultana's study is an example of the interdisciplinary approaches that washback researchers have used to strengthen their studies recently. What gives Sultana's washback research a unique perspective is the application of a validation framework to study washback from a curriculum perspective in English classrooms. By contrast, the doctoral study by Abdulhamid (2018) explored the relationship between washback and alignment by employing Webb's (1997) alignment model to investigate the connection between the content standards and curriculum. These two studies (Sultana and Abdulhamid) examining the relationship between curriculum alignment and washback took place in two unique contexts, adopting two types of frameworks, but both suggest the utility of a more systematic application of theory-based models in explaining washback.

As noted earlier, a range of recent studies show a trend of using, specifically, validation frameworks to guide the study of washback (Abdulhamid, 2018; Anand, 2018; Booth, 2012; Sultana, 2019a; Sun, 2016). Sun (2016) employed Kane's (2006, 2013) interpretation/use argument (IUA) framework for examining the washback effects of the CET-4 in China. This is one of the few washback studies which have used an argument-based validation framework to study washback of the CET-4 examination, a test taken by around ten million students per year. By linking the validation interpretations with the observed washback and the uses of test scores, Sun argued how and why washback existed in the educational context of China.

In contrast with the argument-based test validation approaches, Ahmed (2018) applied Weir's (2005) socio-cognitive framework of test validation in examining the washback effect of the Secondary Education Certificate Examination (SECE) on English language teachers' and students' perceptions and motivations in the context of Libya. Ahmed argued that "the socio-cognitive framework considers the abilities to engage with language users' mental processing (the cognitive dimension) and their abilities to engage with language use in performing tasks as a social practice but not as a pure linguistic practice" (p. 77). The socio-cognitive framework has six major components: test taker characteristics, context validity, cognitive validity, score validity, consequential validity, and criterion-related validity. Ahmed's study was mainly situated in the consequential validity component of the framework. Test takers' characteristics and data on related factors influencing those characteristics were collected. Evidence of washback on teaching and learning was also collected in regard to contextual and cognitive validity, and finally, washback on learning outcomes was determined with reference to the scoring validity of the SECE English examination. The findings revealed that the absence of alignment between the focus of the English SECE and the objectives of the curriculum had a detrimental effect on the perceptions, aims, values,

and motivation of the Libyan ELT teachers, inspectors, and students. To achieve the desired scores on the examinations, students bypassed the pedagogical objectives by focusing on the test-taking strategies or even by means of cheating, which threatens the score validity of the examination.

Main research design and methods

Tsagari and Cheng (2017: 10) argued for the importance of a washback research design to be multiphase, multimethod, and longitudinal. The researchers suggested:

> researchers need to vary their methods including mixed method explanatory, exploratory, and concurrent design. Also, more sophisticated data collection and analysis methods, e.g., those linking directly with test-takers' characteristics, perceptions of assessments, learning processes, and their learning outcomes (test performance) need to be employed beyond classroom observations and survey methods (interviews and questionnaires).

Prior to Tsagari and Cheng (2017), key researchers (Bailey, 1999; Green, 2007; Spratt, 2005; Wall and Alderson, 1993) also advised using multiple methods to explore more than one factor to tease out the complexities that washback research implies. If we observe the recent trends in doctoral dissertations, many of them utilise multiphased and multimethod research designs, and a few studies are also longitudinal. For example, Umashankar's (2017) doctoral study in Sri Lanka explored its washback at three levels of the educational system: the decision-making, intervening, and implementing levels. A mixed-methods design was employed in this study using semi-structured interviews with the decision-making and intervening participants (e.g., teacher trainers and in-service advisors), questionnaires with the implementing participants (e.g., teachers and students), and finally classroom observations and follow-up interviews with the teachers. Sultana (2019a) investigated the interplay of policy, classroom instruction, and student learning within the education system by using a multimethod qualitative research design. Sultana used all qualitative approaches, such as document analyses, interviews with policymakers and teachers, classroom observations, and focus groups with the students. Both Umashankar and Sultana used a triangulation approach to interpret their research findings, which was earlier advised by Bailey (1996). These multimethod and especially multi-stakeholder studies reduce the "blind man and the elephant syndrome" (Cheng, 2008; Tsagari and Cheng, 2017) by exploring the same phenomenon using various instruments and collecting data from different stakeholders, thus having a better chance to observe the testing and education system holistically. The collection of test scores is one of the other instruments which Sun (2016) used, along with interviews and questionnaires, to investigate washback in relation to students' learning outcomes. Uses of test scores in washback research essentially connect washback studies with test validity.

However, before deciding on a research design, researchers need to identify the type of washback research they are going to conduct. Wall and Horák (2007) identified two types of washback studies: one, studies which examine the influence of the tests which are already in operation and second, studies which explore the effects of a newly introduced test. To investigate the washback of a newly introduced test, Wall (2012: 88) listed the following steps:

- identify the sorts of washback the test designers originally intended, and the means they planned to use to ensure that this washback actually appeared;

- analyze the new test in detail, to see how it differs from the test it is replacing;
- predict what washback may occur (intended and unintended), using information from the first two steps;
- design a baseline study, to find out what teaching and/or learning look like before participants are aware of the new test demands;
- carry out the baseline study, to document the educational setting and practices before the test is introduced;
- disseminate the results of the study, to inform the test developers of any features in the setting or established practices that might negatively affect the participants' understanding and ability to accommodate the new test demands.

To conduct washback research on an examination which has been in operation for a long time, researchers need to explore the available baseline studies and studies conducted in the same context to build their research. In this case, it is essential to trace the major changes which have taken place when the examination is in operation, and the researcher needs to create a strong rationale for carrying out washback research. The task is more organized and systematic when the examination is already established and well informed (e.g., Green, 2007, who looked at the effects of the IELTS test). However, when a researcher investigates a localised examination that is not as organised and well documented as the international tests, the researcher needs to use methods that will provide richer data about the context. Many local examinations (see Sultana, 2019a) do not have published test manuals or research materials to rely on. In such situations, researchers need to provide thick and nuanced descriptions of the contexts, the participants, and the education system where the test is situated to find out how and why test washback exists within that particular context. The essential steps are as follows:

- study the available background materials to understand the problems, perceptions, and influences of the test;
- study the test in detail from test validation perspectives;
- study the test in detail as to which stage of education the test is situated: elementary, secondary education, or university EAP context. Tests have differential importance to students at differential levels of education;
- find out what the expectations were when the test was originally designed and why the test was introduced. It is important to understand the policymakers' visions;
- carry out the study with methods that can provide more fine-grained understanding about the gaps and conflicts that have hindered the visions of the policymakers;
- explore the stakeholders' perceptions and understanding and if their visions are different than those of the policymakers;
- disseminate the findings to inform the policymakers and stakeholders to improve the test and the overall teaching and learning.

Traditionally, washback studies have used interviews, focus groups, classroom observations, questionnaires, diaries, and document analyses as data collection methods. All these methods help the researchers understand washback from a neutral point of view. However, so far, washback studies have not explored any innovative methods to delve into the cognition of the stakeholders. We have only explored methods by which we have observed washback to take place or stakeholders who have self-reported their experiences; however, it is likely that methods related to cognition will help us see how those stakeholders experience

washback effects and how those experiences create desired or undesired teaching and learning processes and outcomes.

Recommendations and future directions

Wall (2012) concluded the first edition of this chapter with a sense of frustration "to see how many directions the research has gone off into and how little connection there seems to be between researchers working on the same types of questions" (p. 89). This present review of recent washback literature shows a slight shift in the washback scholarship from a cause-and-effect perspective, and from replicating previous survey-based studies, to the adoption of innovative research questions and stronger conceptual and theoretical frameworks. Such an adoption makes a conscious attempt to tie washback closely to the test and testing system and, at the same time, to tie to curriculum alignment and student learning at the education system level. Nevertheless, we recognise the importance of replication of previous studies because it "would allow researchers to build on what we have learned theoretically, conceptually, and methodologically over the years and further our understanding of this phenomenon" (Tsagari and Cheng, 2017: 10). However, Wall's one concern is still valid: that "it is rare to see research that truly builds on work that has been done previously" (p. 89). For example, Cheng's (2005) research concludes on an important note: to examine washback with a focus on teacher education programs. However, we have seen little action from washback researchers responding to this call, apart from a recent doctoral dissertation (Oo, 2019), although this dissertation focused on assessment *for* learning literacy in pre-service teacher education rather than washback specifically. Washback researchers have not systematically explored how teacher education programs can work as an ally in creating washback on teaching and learning and in achieving a higher degree of alignment. So one of the main challenges in washback research seems to be an absence of continuation within and across contexts and a lack of collaboration with other key stakeholders in the education system.

In the array of washback studies, assessment literacy has also not been investigated from a washback perspective. From the pioneering washback studies to the more recent ones, it is noted that often the underlying key tenets of the examinations were not understood by the teachers. Is it possible that the teachers did not have adequate assessment literacy to implement those principles in their instruction? Koh *et al*. (2018) indicated that assessment-literate teachers were able to create more learning opportunities for the students by improving their own teaching. Sultana (2019b) mentioned, "owing to inadequate assessment literacy, language teachers are unable to implement the curriculum goals . . . [and] fail to grasp the essential connections among the curriculum, classroom instruction, and examination" (p. 12). Vogt *et al*.'s (2020) study on assessment literacy revealed that because of the changing examination policy, teachers attended in-service training focusing on examination, which created positive washback in their teaching. Apparently, so far, neither assessment literacy researchers nor washback researchers have attempted to establish a relationship between these two fields. Future washback researchers are encouraged to explore this missing link between teachers' assessment literacy and washback. While this will add further complexity to washback research, it will certainly provide more clarity about the link between testing policies and teacher understanding.

More recently, there has been a promising trend to examine washback on learning, yet few washback studies have attempted to trace the role that cognition plays in student learning. Existing washback studies have explored what students do in test preparation or what they do as a result of the examination; nonetheless, we are yet to know how learning takes

place in their cognitive repertoires. In addressing this gap, washback researchers can use the cognitive validity component as illustrated by Pellegrino *et al*. (2016) in addition to the two aspects of instructional validity and inference validity. A focus on cognitive validity (see Weir, 2005) could guide us in exploring the learning process in relation to test validity and washback. In the next stage of washback research, learning processes should dominate the scholarship. In doing so, innovative research methodologies, if needed from other disciplines, should be borrowed to investigate washback on learning.

Last, but not least, there has been a discussion on the equity and fairness of testing (Shohamy, 1997, 1998, 2001; see also Walters, this volume). Crocker (2006) states that "[n]o activity in educational assessment raises more instructional, ethical and validity issues than test preparation for large-scale, high-stakes tests" (p. 115). Nevertheless, conventionally, washback researchers seem to be less interested in examining these issues. Washback studies are rarely associated with issues of fairness and morality in education. Sultana (2019a) concluded in her study that washback studies certainly dig into the morality of the education system and of key stakeholders in countries like Bangladesh, where resources are very limited. It will be worthwhile to problematise washback effects in regard to the issues of equity, fairness, and morality in relation to the politics and geopolitics of language testing. Washback researchers need to explore more about the political angle of tests, the testing industry, and the political context in understanding the complex multi-layered and multi-dimensional nature of washback.

Further reading

Cheng, L., Sun, Y. and Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching* 48: 436–70. doi:10.1017/S0261444815000233
This is one of the most up-to-date systematic reviews of washback studies in the field so far. A systematic search of the pertinent literature between 1993 and 2013 identified a total of 123 publications consisting of 36 review articles and 87 empirical studies. This review focuses on the empirical studies only. A further breakdown of these empirical studies reveals 11 books and monographs, 27 doctoral dissertations, 40 journal articles, and 9 book chapters. This intensity of research activity underscores the timeliness and importance of this research topic and highlights its maturity.

Tsagari, D. and Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. G. Or and S. May (eds.), *Language Testing and Assessment*, 3rd edn. New York, NY: Springer, 359–72.
Tsagari and Cheng elaborated the terms *washback*, *impact*, and *consequences* by presenting the major empirical washback studies conducted over the past 30 years. This is a second edition of this chapter (see Cheng, 2008, as the first edition). The complexity of washback research has been discussed by pointing out the problems and difficulties that washback studies face. The key strength of this chapter lies in providing detailed future directions of washback studies, which are valuable for both novice and experienced researchers.

References

Abdulhamid, N. (2018). *What Is the Relationship Between Alignment and Washback? A Mixed-Methods Study of the Libyan EFL Context*. Doctoral dissertation, Carleton University, Ottawa, Canada. Retrieved from https://curve.carleton.ca/2fe0aa90-9afd-498d-8e6d-78f96f055ff4.

Ahmed, A. A. M. (2018). *Washback: Examining English Language Teaching and Learning in Libyan Secondary School Education*. Doctoral dissertation, The University of Huddersfield, Huddersfield, UK. Retrieved from http://eprints.hud.ac.uk/id/eprint/34799/.

Alderson, J. C. and Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing* 13: 280–97. https://doi.org/10.1177/026553229601300304.

Alderson, J. C. and Wall, D. (1993). Does washback exist? *Applied Linguistics* 14: 115–29. https://doi.org/10.1093/applin/14.2.115.

Anand, P. (2018). *Testing Regime Change as Innovation: Washback Potential Over Time*. Doctoral dissertation, Carleton University, Ottawa, Canada. Retrieved from https://curve.carleton.ca/system/files/etd/22e85455-2126-4496-a5f2-1b17277a5cb1/etd_pdf/48581ca9698e9159a3f830896ba92cc0/anand-testingregimechangeasinnovationwashbackpotential.pdf.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2: 1–34. https://doi.org/10.1207/s15434311laq0201_1.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*, vol. 1. New York, NY: Oxford University Press.

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing* 13: 257–79. https://doi.org/10.1177/026553229601300303.

———. (1999). *Washback in Language Testing*. Princeton, NJ: Educational Testing Service.

Bhattarai, Y. B. (2019). *The Assessment System Reform of the School Leaving Certificate Exam in Nepal: A Grounded Theory of the Reform Process*. Doctoral dissertation, Ottawa, Ontario, Canada. Retrieved from https://ruor.uottawa.ca/handle/10393/39002.

Biggs, J. (1995). Assumptions underlying new approaches to educational assessment: Implications for Hong Kong. *Curriculum Forum* 4: 1–22.

Booth, D. K. (2018). *The sociocultural activity of high stakes standardised language testing*, vol. 12. Cham: Springer. https://doi.org/10.1007/978-3-319-70446-3_4.

———. (2012). *Exploring the Washback of the TOEIC in South Korea: A Sociocultural Perspective on Student Test Activity*. Doctoral dissertation, The University of Auckland, Auckland, New Zealand. Retrieved from http://researchspace.auckland.ac.nz.

Brunfaut, T. (2014). A lifetime of language testing: An interview with J. Charles Alderson. *Language Assessment Quarterly* 11: 103–19. https://doi.org/10.1080/15434303.2013.869818.

Chapelle, C. A. (2020). An introduction to language testing's first virtual special issue: Investigating consequences of language test use. *Language Testing* 37: 638–45. https://doi.org/10.1177/0265532220928533.

Chen, L. (2002). *Taiwanese Junior High School English Teachers' Perceptions of the Washback Effect of the Basic Competence Test in English*. Doctoral dissertation, Ohio State University, Columbus, OH. Retrieved from https://etd.ohiolink.edu/.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education* 11: 38–54. https://doi.org/10.1080/09500789708666717.

———. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation* 24: 279–301. https://doi.org/10.1016/S0191-491X(98)00018-2.

———. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and Teacher Education* 15: 253–71. https://doi.org/10.1016/S0742-051X(98)00046-8.

———. (2005). *Changing Language Teaching Through Language Testing: A Washback Study*, vol. 21. Cambridge, UK: Cambridge University Press.

———. (2008). Washback, impact and consequences. In E. Shohamy and N. H. Hornberger (eds.), *Encyclopedia of Language and Education*. New York, NY: Springer, 2479–94.

———. (2014). Consequences, impact, and washback. In A. J. Kunnan (ed.), *The Companion to Language Assessment*. Hoboken, NJ: John Wiley & Sons, 1–14.

———. (2018). Geopolitics of assessment. In S. Abrar-ul-Hassan (ed.), *TESOL Encyclopedia of English Language Teaching: Teaching English as an International Language*. Hoboken, NJ: John Wiley & Sons.

Cheng, L., Sun, Y. and Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching* 48: 436–70. https://doi.org/10.1017/S0261444815000233.

Cheng, L., Watanabe, Y. and Curtis, A. (eds.). (2004). *Washback in Language Testing: Research Contexts and Methods*. Mahwah, NJ: Lawrence Erlbaum.

Crocker, L. (2006). Preparing examinees for test taking: Guidelines for test developers and test users. In S. Downing and T. Haladyna (eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum, 115–28.

Cronbach, L. (1971). Test validation. In R. L. Thorndike (ed.), *Educational Measurement*, 2nd edn. Washington, DC: American Council on Education, 443–507.

Fox, J. and Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario secondary school literacy test by first and second language test-takers. *Assessment in Education* 14: 9–26. https://doi.org/10.1080/09695940701272773.

Froehlich, V. (2016). Washback of an oral exam on teaching and learning in German middle schools. In D. Tsagari (ed.), *Classroom-Based Assessment in L2 Contexts*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 161–85.

Froetscher, D. (2016). A new national exam: A case of washback. In D. Tsagari and J. Banerjee (eds.), *Contemporary Second Language Assessment*. Oxford: Bloomsbury Academic, 61–81.

Fullan, M. (2015). *The New Meaning of Educational Change*. New York, NY: Routledge.

Green, A. (2007). *IELTS Washback in Context: Preparation for Academic Writing in Higher Education*. Cambridge, UK: Cambridge University Press.

Henrichsen, L. E. (1989). *Diffusion of Innovations in English Language Teaching: The ELEC Effort in Japan, 1956–1968*. New York, NY: Greenwood Press.

Hoque, M. E. (2011). *Washback of the Public Examination on Teaching and Learning English as a Foreign Language (eFL) at the Higher Secondary Level in Bangladesh*. Unpublished doctoral dissertation. Jahangirnagar University, Dhaka, Bangladesh.

Hughes, A. (1993). *Backwash and TOEFL 2000*. University of Reading, England, Unpublished manuscript.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50: 1–73. https://doi.org/10.1111/jedm.12000.

———. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurement*, 4th edn. Washington, DC: American Council on Education, Praeger, 17–64.

Khaniya, T. R. (1990). *Examinations as Instruments for Educational Change: Investigating the Washback Effect of the Nepalese English Exams*. Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK.

Kim, E. Y. J. (2017). The TOEFL iBT writing Korean students' perceptions of the TOEFL iBT writing test. *Assessing Writing* 33: 1–11. https://doi.org/10.1016/j.asw.2017.02.001.

Koh, K., Burke, L. E. C. A., Luke, A., Gong, W. and Tan, C. (2018). Developing the assessment literacy of teachers in Chinese language classrooms: A focus on assessment task design. *Language Teaching Research* 22: 264–88. https://doi.org/10.1177/1362168816684366.

Kunnan, A. J. (2009). Politics and legislation in citizenship testing in the U.S. *Annual Review of Applied Linguistics* 29: 37–48. https://doi.org/10.1017/S0267190509090047.

Latham, H. (1877). *On the Action of Examinations Considered as a Means of Selection*. Cambridge, UK: Deighton, Bell and Company.

Ma, J. (2019). Did test preparation practices for the college English test work? A study from Chinese students' perspective. In S. Papageorgiou and K. Bailey (eds.), *Global Perspective on Language Assessment*. New York, NY: Routledge, 169–82.

Ma, J. and Cheng, L. (2016). Chinese student' perceptions of the value of test preparation courses for the TOEFL iBT: Merit, worth and significance. *TESL Canada Journal* 33: 58–79. https://doi.org/10.18806/tesl.v33i1.1227.

Markee, N. (1997). *Managing Curricular Innovation*. Cambridge: Cambridge University Press.

Marshall, B. (2017). The politics of testing. *English in Education* 51: 27–43. https://doi.org/10.1111/eie.12110.

Mathew, R. (2012). Understanding washback: A case study of a new exam in India. In C. Tribble (ed.), *Managing Change in English Language Teaching: Lessons from Experience*. London: British Council, 195–202.

McNamara, T. (2010). The use of language tests in the service of policy: Issues of validity. *Revue française de linguistique appliquée xv*: 7–23. https://doi.org/10.3917/rfla.151.0007.

Memon, N. (2015). *'If I Just Get One IELTS Certificate, I Can Get Anything': An Impact Study of IELTS in Pakistan*. Doctoral dissertation, The University of Edinburgh, The University of Edinburgh. Retrieved from https://era.ed.ac.uk/handle/1842/21122.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: Ace and Macmillan, 13–103.

———. (1996). Validity and washback in language testing. *Language Testing* 13: 241–56. https://doi.org/10.1177/026553229601300302.

Onaiba, A. M. E. (2013). *Investigating the Washback Effect of a Revised EFL Public Examination on Teachers' Instructional Practices, Materials and Curriculum*. Doctoral dissertation, University of Leicester, UK. Retrieved from https://pdfs.semanticscholar.org/8e6c/c63f0213d4f965e1e9b19dd0b8d39d5bd3f7.pdf.

Oo, C. Z. (2019). *Assessment for Learning Literacy and Pre-Service Teacher Education: Perspectives from Myanmar*. Unpublished doctoral dissertation. The University of New South Wales, Australia.

Pan, Y. C. (2016). Learners' perspectives of factors influencing gains in standardized English test scores. *TEFLIN Journal* 27: 63–81. https://doi.org/10.15639/teflinjournal.v27i1/63-81.

Pan, Y. C. and Newfields, T. (2013). Student washback from tertiary standardized English proficiency exit requirements in Taiwan. *Journal of Teaching and Learning* 9: 1–16. https://doi.org/10.22329/JTL.V9I1.3540.

Pellegrino, J. W., DiBello, L. V. and Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist* 51: 59–81. https://doi.org/10.1080/00461520.2016.1145550.

Popham, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappan* 68: 679–82.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing* 22: 142–73. https://doi.org/10.1191/0265532205lt300oa.

Rind, I. A. and Mari, M. A. (2019). Analysing the impact of external examination on teaching and learning of English at the secondary level education. *Cogent Education* 6: 1–14. https://doi.org/10.1080/2331186X.2019.1574947.

Sato, T. (2019). An investigation of factors involved in Japanese students' English learning behavior during test preparation. *Papers in Language Testing and Assessment* 8: 69–95.

Shepard, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement* 9: 15–22. https://doi.org/10.1111/j.1745-3992.1990.tb00374.x.

Shih, C. (2007). A new washback model of students' learning. *The Canadian Modern Language Review* 64: 135–62. https://doi.org/10.3138/cmlr.64.1.135.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing* 14: 340–49. https://doi.org/10.1177/026553229701400310.

———. (1998). Critical language testing and beyond. *Studies in Educational Evaluation* 24: 331–45. https://doi.org/10.1016/S0191-491X(98)00020-0.

———. (2001). Democratic assessment as an alternative. *Language Testing* 18: 373–91. https://doi.org/10.1177/026553220101800404.

Shohamy, E., Donitsa-Schmidt, S. and Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing* 13: 298–317. https://doi.org/10.1177/026553229601300305.

Spratt, M. (2005). Washback and the classroom: The implications for teaching and learning of studies of washback from exams. *Language Teaching Research* 9: 5–29. https://doi.org/10.1191/1362168805lr152oa.

Sultana, N. (2018). A brief review of washback studies in the South Asian countries. *The Educational Review, USA* 2: 468–74. https://doi.org/10.26855/er.2018.09.002.

———. (2019a). *Exploring the Alignment of the Secondary School Certificate English Examination with Curriculum and Classroom Instruction: A Washback Study in Bangladesh*. Doctoral dissertation, Queen's University, Kingston, Ontario, Canada. Retrieved from https://qspace.library.queensu.ca/handle/1974/26482.

———. (2019b). Language assessment literacy: An uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia* 9: 1–14. https://doi.org/10.1186/s40468-019-0077-8.

Sun, Y. (2016). *Context, Construct, and Consequences: Washback of the College English Test in China*. Doctoral dissertation, Queen's University, Kingston, Canada. Retrieved from https://qspace.library.queensu.ca/handle/1974/13985.

Tan, M. and Turner, C. E. (2015). The impact of communication and collaboration between test developers and teachers on a high-stakes ESL exam: Aligning external assessment and classroom practices. *Language Assessment Quarterly* 12: 29–49. https://doi.org/10.1080/15434303.2014.1003301.

Tsagari, D. and Cheng, L. (2017). Washback, impact, and consequences revisited. In E. Shohamy, I. G. Or and S. May (eds.), *Language Testing and Assessment*, 3rd edn. New York, NY: Springer, 359–72.

Umashankar, S. (2017). *Washback Effects of Speaking Assessment of Teaching English in Sri Lankan Schools*. Doctoral dissertation, University of Bedfordshire, Luton, England. Retrieved from https://core.ac.uk/download/pdf/151440224.pdf.

Vogt, K., Tsagari, D. and Spanoudis, G. (2020). What do teachers think they want? A comparative study of in-service language teachers' beliefs on LAL training needs. *Language Assessment Quarterly* 17: 386–409. https://doi.org/10.1080/15434303.2020.1781128.

Volante, L., DeLuca, C., Adie, L., Baker, E., Harju-Luukkainen, H., Heritage, M., Schneider, C., Stobart, G., Tan, K. and Wyatt-Smith, C. (2020). Synergy and tension between large-scale and classroom assessment: International trends. *Educational Measurement: Issues and Practice* 39: 21–29. https://doi.org/10.1111/emip.12382.

Vyn, R. (2019). *Promoting Curricular Innovation Through Language Performance Assessment: Leveraging AAPPL Washback in a K–12 World Languages Program*. Doctoral dissertation, University of Iowa, Iowa City, Iowa, the USA. Retrieved from https://ir.uiowa.edu/etd/6872/.

Wall, D. (2005). *The Impact of High-Stakes Examinations on Classroom Teaching: A Case Study Using Insights from Testing and Innovation Theory*, vol. 22. Cambridge, UK: Cambridge University.

———. (2012). Washback. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. London, UK: Routledge, 79–92.

Wall, D. and Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing* 10: 41–69. https://doi.org/10.1177/026553229301000103.

Wall, D. and Horák, T. (2007). Using baseline studies in the investigation of test impact. *Assessment in Education* 14: 99–116. https://doi.org/10.1080/09695940701272922.

Wang, S. (2018). *Investigating the Consequential Validity of the Hanyu Shuiping Kaoshi (Chinese Proficiency Test) by Using an Argument-Based Framework*. Doctoral Dissertation, McGill University, Montreal, Quebec, Canada. Retrieved from https://escholarship.mcgill.ca/concern/theses/4q77ft93g.

Watanabe, Y. (1996). Does grammar translation come from the entrance examination? Preliminary findings from classroom-based research. *Language Testing* 13: 318–33. https://doi.org/10.1177/026553229601300306.

Webb, N. L. (1997). *Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education*. NISE Research Monograph No. 6. Madison: University of Wisconsin-Madison, National Institute for Science Education.

———. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education* 20: 7–25. https://doi.org/10.1080/08957340709336728.

Weir, C. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Xie, Q. (2010). *Test Design and Use, Preparation, and Performance: A Structural Equation Modeling Study of Consequential Validity*. Unpublished doctoral dissertation. University of Hong Kong, Hong Kong.

9

Assessing young learners

Yuko Goto Butler

Introduction

Young language learners (YLLs) are defined as children ages 3–12 who are learning one or more additional or foreign languages. These learners have many unique characteristics related to both their age and their environment. Unlike adult additional/foreign (L2) language learners, YLLs are still developing their first language (L1) in addition to developing an L2. They are also in the midst of cognitive, social-cognitive, and affective development. Importantly, there are substantial individual variabilities in these linguistic- and age-related developments both within and across chronological age groups. Environmental factors influencing YLLs' language development include their goals for learning the additional language, the resources and support available in their school and communities, and adults' attitudes toward the target and home languages. Educators and researchers must account for this wide range of developmental and environmental factors when assessing YLLs, including defining assessment constructs, designing assessment formats and procedures, and interpreting the results.

After outlining some historical perspectives, this chapter discusses the central issues at play when assessing YLLs in the context of increasing globalization and the advancement of technology. Topics covered in this chapter include the age appropriateness of assessment development and practice, assessment for learning, and reconciling children's diverse needs with demands for standardization. The chapter concludes with implications for practice as well as future directions for research.

Historical perspectives

Despite the critical role that assessment plays in YLLs' language education, empirical research on assessing YLLs did not receive sustained attention until the 1990s, and it was not until the turn of the twenty-first century that the field blossomed. Pioneering works – including McKay (1995), Shohamy (1997), Teachers of English to Speakers of Other Languages (TESOL, 1998), and Torrance and Pryor (1998), as well as a special issue of *Language Testing* (Rea-Dickins, 2000) – paved the way for succeeding inquiries into the

DOI: 10.4324/9781003220756-13

assessment for YLLs. A number of books (Jang, 2014; Nikolov, 2016; Papp and Rixon, 2018; Prošić-Santovac and Rixon, 2019; Schissel, 2019; Wolf and Butler, 2017) and various journal articles on assessing YLLs appeared in the 2010s, covering a wide range of regions and diverse groups of YLLs. *Language Teaching for Young Learners*, an international academic journal that accepts studies on assessment for YLLs, was launched in 2019. Special issues focusing on YLLs are underway in two major language assessment journals: *Language Assessment Quarterly* (Huang and Butler, 2020) and *Language Testing* (Timpe-Laughlin and Cho, 2021). A number of international collaborative projects, such as the Early Language Learning in Europe Study (the ELLiE Study; Enever, 2011), were established to conduct comparative investigations on assessment for YLLs. Researchers have also actively formed various professional networks both regionally and internationally; these include the Research Group for Early Language Learning in the Association Internationale de Linguistique Appliquée (International Association of Applied Linguistics, AILA, https://aila.info/) and the Special Interest Group of Assessment of Young Learners at the International Language Testing Association (ILTA, www.iltaonline.com/general/custom.asp?page=SIGs).

Recent active research on YLL assessment seems to be influenced by drastic changes since around the turn of the century, in the environments where YLLs' language learning takes place. Two macro-level factors that have been particularly influential are increasing globalization and advances in digital technology (Butler, 2019). First, greater mobility of people across regions and countries has yielded an increase in the number of children who receive instruction through a language other than their home language(s). These YLLs are highly diverse in terms of their language backgrounds, schooling experience, background knowledge, and other developmental and individual factors. They usually live in multilingual contexts, where multiple language resources are dynamically used and practiced as a norm (Schissel *et al*., 2019). They need to learn the content of subjects such as math and science in addition to the language of schooling. Globalization has also changed the profiles of young foreign language learners. Foreign language education is no longer limited to children with privileged backgrounds. Diverse groups of children now learn a foreign language, English in particular, largely due to the inclusion of foreign language in early school curricula (Enever, 2018). Moreover, children are starting to learn foreign languages at much younger ages – a change that can be attributed at least in part to the widely held belief that "the younger the better" for successful language learning; indeed, foreign language learning at the pre-primary level is gaining in popularity (e.g., Mourão and Lourenço, 2015). Teachers of YLLs are increasingly responsible for conducting age-appropriate assessments for children with diverse needs (including very young children and children in multilingual contexts) and using the assessment results formatively to assist children's learning (Butler, 2019; McKay, 2006; Nikolov, 2016; Nikolov and Timpe-Laughlin, 2020).

Second, advances in digital technology, such as personal computers, tablets, and smartphones, have changed the way we teach and learn languages. As educators and YLLs themselves rely more heavily on digital technology for language learning, a likely natural consequence is that they will also rely on technology for assessment (i.e., computer-based language assessment). Increasing reliance on digital technology has also changed target language abilities for YLLs who grow up with such technology from birth; consequently, we may need to modify or adjust our construct definition for the age of digital technology. For example, as reading online is increasingly common among young learners, reading may no longer merely entail constructing meaning out of texts but also involve processing visual, audio, and sensory information. Such multimodal processing in assessment may more authentically reflect YLLs' real-world language use. Some digital technology (e.g.,

social networking services [SNS] and chats) obscures the traditional boundaries between oral and written language that are used in language assessment (i.e., assessment in the four skill domains). Technology can also assist test developers as they work to improve the quality of assessments. Assessment practices relying on technology (e.g., machine raters) and validation studies using technology (e.g., eye-tracking as in Lee and Winke, 2018) seem to be on the rise.

Critical issues and topics

Bearing in mind the macro-level environmental changes discussed earlier, this section addresses the following three major critical issues: (1) developing and implementing age-appropriate assessment; (2) the interconnectedness of teaching, learning, and assessment; and (3) reconciling children's diverse needs with demands for standardization.

Age-appropriate assessment development and practice

Age appropriateness is a central concern when deciding on the purpose, content, and procedure of any language assessment as well as when interpreting assessment results. Children's information-processing speed, memory and attention span, analytical reasoning, and metacognitive and self-reflective abilities show drastic improvement over the span of their school-age years. Similarly, they develop socio-cognitive interactive abilities such as maintaining joint attention, making use of nonverbal information, and understanding other people's intentions, and they also become better at collaborating with others in tasks. As children mature and accumulate world experiences, they also develop various cognitive and socio-cognitive strategies when engaging in tasks (Butler, 2019; Papp, 2018). It is important to note, however, that these changes do not necessarily happen in a linear fashion and that various domains (e.g., cognitive and social domains) do not necessarily develop in tandem. There are also substantial individual differences depending on children's personal traits and sociocultural environments (Butler, 2019).

When deciding tasks for assessing YLLs, therefore, it is critical to consider the cognitive and socio-cognitive demands that the tasks entail. For example, when asking children to tell a story based on pictures, cognitive demands can be manipulated by changing the number and order of the pictures, using simpler or more complicated story plots, adjusting the amount of story-planning time that they are allowed, and so forth (Pinter, 2015). Tasks that children find slightly cognitively challenging can motivate them (Jang, 2014), but cognitive demands that exceed children's developmental capacity are a potential threat to assessment validity and may negatively influence children's confidence in, and motivation for, learning (Mihaljević Djigunović, 2019).

YLLs' socio-cognitive development should also be taken into account when deciding on the format and procedure of the assessment. Consider, for example, one-on-one oral interviews between a teacher and student. This popular assessment format for YLLs has some merits in that the teacher can tailor questions to suit the child's developmental and linguistic levels and even strategically stretch the child's abilities. This format may work particularly well for younger children, children with special needs, and children with lower proficiency and confidence levels. Despite the affordances of this individualized format, however, it may elicit a limited type of response (e.g., answering questions and agreeing with the teacher's remarks) due to the power imbalances inherent in the child-teacher relationship. Assessments in which children work among themselves, either in pairs or larger

groups, may elicit a greater variety of discourse types (e.g., asking questions, confirming questions, negotiating meanings, disagreeing with others' opinions, providing alternative opinions, etc. (Butler and Zeng, 2014, 2015)). But even these pair- and group-assessment formats may not work well if there are uneven power dynamics or substantial variations in the children's socio-cognitive development levels (e.g., Butler and Zeng, 2014, 2015; García Mayo and Agirre, 2016; also see Oliver and Azkarai, 2019 for a review of child interaction research).

Developmental factors play a critical role in how YLLs' target language abilities – namely, language abilities for constructing and negotiating meaning appropriately and successfully in various social and academic contexts – are identified and conceptualized. Target language-use domains vary based on YLLs' learning objectives and age. As noted at the beginning of this chapter, the fact that YLLs, unlike adults, are still in the process of developing both their L1 and their additional/foreign language uniquely influences their L1 and L2 learning (e.g., Bailey, 2017; Łockiewicz *et al*., 2018). Conceptualizing YLLs' language abilities is particularly challenging if one takes into account their plurilinguistic practices (Schissel *et al*., 2019).

Roughly speaking, the field of language assessment has taken two major approaches to deciding the target language abilities for YLL assessment: a top-down, theory-driven approach and a bottom-up approach driven by the judgment of experts. With respect to the theory-driven approach, a number of theoretical models for language abilities have been proposed; some models assume that language abilities are a composite of traits and abilities at the individual level, while other models treat language abilities as co-constructed among participants through interaction in a specific context (Purpura, 2016). Applying these theoretical models to YLLs' assessment presents a serious challenge. On one hand, if we accept some compositionality of traits or abilities, we have limited information about the interrelations among the components, how they differ from those of adults, or how such relationships change over the course of development. For example, Gu and Hsieh (2019) suggested that grammar accuracy in oral proficiency – an important ability contributing to adults' oral proficiency – may not be straightforward for YLLs because children's oral speech is often short and fragmented. Their study found that different ability features of oral performance characterize different oral proficiencies. Some features, such as the use of cohesive devices for connecting information, were confounded with cognitive maturity; younger YLLs had more trouble using cohesive devices. Hsieh and Wang (2019) also found that, unlike in adults, disfluencies as well as the frequency and duration of pauses were associated with oral proficiency among YLLs. Wolf *et al*. (2017) examined kindergarten through second grade students' oral performance in the United States and reported that certain types of grammatical errors (e.g., verb forms, tense, etc.) were observed for both native English speakers and English learners alike, irrespective of the grade levels that they examined. If we accept an emic and fluid conceptualization of language abilities, however, constructs should be defined specifically in the given context and should involve "diligent documentation of how particular language interactions take place within specific sociocultural settings from the perspectives of those involved" (Chalhoub-Deville, 2019: 476). In multilingual contexts, it has been suggested that construct definitions require the incorporation of various multilingual features while acknowledging students' use of multiple linguistic resources, such as translanguaging (Chalhoub-Deville, 2019). What kinds of multilingual linguistic repertoires one should expect to see for different age groups is an open question, however.

The other approach for identifying the target language abilities is a bottom-up approach that often relies on the judgment of teachers and experts. One example of this approach is the Common European Framework of Reference for Language (CEFR) (Council of

Europe, 2001, a revised version appeared in 2018). The CEFR has had a tremendous influence on foreign language education teaching and assessment, including for YLLs, not only in Europe but also in other regions in the world. The CEFR's competence descriptors are composed of six proficiency levels (A1, A2, B1, B2, C1, and C2, where C2 is the highest level) specifying what learners are supposed to be able to do using the target language at each level. Many countries claim to set the goal for their foreign language education at the end of primary school at either A1 or A2 (Rixon, 2019). However, the CEFR was not specifically designed for YLLs, and applying it to this age group requires substantial age-related modification for its valid use. Effort has been made to create developmentally more appropriate CEFR-aligned descriptors for YLLs by creating a pre-A1 level, expanding the A1 and A2 levels, and so forth (e.g., Benigno and de Jong, 2016; North, 2014; Pearson Education, 2018). Preliminary example descriptors developed specifically for YLLs age 7–10 and 11–15 are available from the Council of Europe for further validation checks (Goodier and Szabo, 2018).

Finally, researchers have responded to the surge in popularity of teaching additional/foreign languages to very young children (ages three to five, or preschoolers and kindergarteners) by conducting a number of assessment studies focusing on this age group (e.g., Huang and Konold, 2014; Guzman-Orth *et al*., 2019; Łockiewicz *et al*., 2018; Prošić-Santovac and Rixon, 2019). It has been suggested that assessment for such a very young group should not focus on mastering language objectives but rather on fostering further learning. Formative assessment should be embedded in regular language activities by employing multiple means of assessment, including teachers' observations, task performance based on rubrics or criteria, drawing and other nonverbal means (e.g., gestures), show and tell, portfolios, and so forth. If summative assessments are to be carried out, they should reflect activities with which children are familiar; additionally, careful consideration should be given to children's attention and memory spans as well as to minimizing potential negative influences on children's affects (e.g., anxiety and stress) (Ghosn, 2019). Preschool children are different from older learners in that their learning is primarily kinesthetic and implicit, and it is deeply tied to particular scenes and activities associated with language use (Alexiou *et al*., 2019). Such characteristics, therefore, should be reflected in assessment (e.g., Prošić-Santovac and Navratil, 2019, for use of puppet shows as assessment). Research has also revealed that, with sufficient and appropriate support, children as young as four can be self-reflective and autonomous. Assessments that foster these abilities can be usefully implemented among very young children as well as older children (Ellis and Rixon, 2019).

Interconnectedness of teaching, learning, and assessment: assessment for learning

Assessment is indispensable in that it not only has the potential to provide useful information on teaching and learning but can also directly enhance the learners' learning. The idea of *assessment for learning* (Black and Wiliam, 1998), or assessment primarily used for formative and diagnostic purposes, is important for all learners (see Fox *et al*., this volume). It is particularly critical for YLLs, however, because they are at the beginning of a lifelong journey of learning, and their experiences with assessment (both positive and negative) may have a long-lasting impact on their successive learning. Because self-reflection and autonomy can ultimately lead to better learning, assessments, such as self-assessments, that aim to foster these abilities have become popular.

Rooted in sociocultural theory of mind (Vygotsky, 1978), *dynamic assessment* (DA) incorporates various types of assistance during assessment in order to support students' learning. Usually structured as a one-to-one interaction between a teacher and a student, DA offers assistance that is responsive and adjustable to the learner's needs in a particular moment in interaction. Based on the notion of *zone of proximal development* (Vygotsky, 1978), DA aims to identify not only what the learner can do individually but also what they can achieve with the help of the teacher through mediation (see Anton and Garcia, this volume; also, Lantolf and Poehner, 2004; Poehner *et al*., 2017). Similarly, *cognitive diagnostic assessment* (CDA) has been implemented with YLLs to uncover their cognitive and metacognitive processing and strategies while they engage in tasks such as comprehending a text (see Jang and Sinclair, this volume). By offering ongoing diagnostic feedback on learners' strengths and weaknesses, CDA assists learners with completing tasks and, ultimately, improves their learning. CDA yields positive interventional effects among young learners in particular as it provides diagnostic feedback on young learners' cognitive and metacognitive developmental factors, which can have potentially large effects on their language development (Jang, 2008; Jang *et al*., 2013, 2017). In their implementation of CDA to YLLs, Jang *et al*. (2017) also highlighted the substantial role that emotion (e.g., interest and anxiety) plays in children's cognitive and metacognitive processing and their task performance. Emotion is a factor that has not received sufficient attention in assessment for YLLs.

In the spirit of *assessment for learning*, efforts have been made to incorporate ongoing feedback even in computer-based and standardized assessment contexts (Choi *et al*., 2019; Jang, 2008; Wolf *et al*., 2016). Choi *et al*. (2019), for example, offered in the form of scaffolding in a storytelling task (i.e., supports provided to YLLs to perform the assessment task) among third-grade students learning English as an additional language in the United States as part a large-scale, computer-based standardized test. They found that step-by-step scaffolding helped the children perform better in the post-scaffolding task. It is important to note that the improvement as a result of scaffolding was not uniformly observed in all test takers in Choi *et al*.'s study. Researchers are still determining how best to offer scaffolding to YLLs in assessment contexts. For scaffolding to be successfully implemented in a standardized format during assessment, we must identify the interaction between the quality and timing of scaffolding and individual children's characteristics.

Self-assessment has become widely recognized as a way to foster YLLs' self-reflective abilities and autonomy, which in turn are expected to lead to greater learning (Oscarson, 1989). Self-assessment is also considered to be a relatively low-stakes approach and thus less stressful for young learners (Butler, 2016). Various kinds of self-assessment items and descriptors are available for teachers of YLLs, including the self-assessment grid as part of the European Language Portfolio for Young Learners (Council of Europe, 2019). Despite efforts to promote the use of self-assessment for YLLs, teachers of YLLs may not always effectively use self-assessment in their own practice (e.g., Becker, 2015). Teachers have expressed concern about the extent to which young learners are capable of assessing their own language abilities, and they are often not sure how best to implement self-assessment in their classrooms (Butler, 2016). Research has shown that there are some developmental effects in the accuracy of YLLs' self-assessment responses; upper primary grade students' responses, compared with those among lower primary graders, are better correlated with external measures such as objective tests and teachers' judgments (e.g., Butler and Lee, 2006; Keane and Griffin, 2018). However, as mentioned already, younger children can show sophisticated and accurate self-appraisal abilities when they receive sufficient guidance and

experience with self-assessment and are asked to evaluate their performance in familiar tasks or tasks with relatively low cognitive demands (Butler, 2016; Ellis and Rixon, 2019).

It has been suggested that successful implementation of self-assessment for YLLs might require more contextualized items and greater immediacy to language activities (Butler, 2016). For example, instead of asking YLLs to respond to "I can follow a short story in English," we should ask them to respond to "I could follow the short story that the teacher read to us in class today" – and to do so immediately after the story-listening activity in class. It is important to note that the mental processing required to answer self-assessment items is very complicated and has complex interactions with children's age and experience as well as their social environments (Butler, 2018a, 2018b). The contextualized self-assessment approach assists YLLs in seeing the relationship between their self-efficacy and actual performance. It is less influenced by individual factors such as YLLs' motivation, personalities, confidence levels, and so forth (Butler and Lee, 2006).

The accuracy of self-assessment may not be critically important, however, when considering the potential merits of self-assessment for enhancing learners' self-regulation, autonomy, and, ultimately, learning. We have some evidence that the repeated implementation of contextualized self-assessment does enhance YLLs' confidence, self-regulation, and learning of the target language (Butler and Lee, 2010; Wong and Mak, 2019). From an assessment-for-learning point of view, self-assessment should be able to help children set their own goals, self-reflect on their performance and progress in relation to their goals, and identify actions to be taken as the next step to achieve those goals. Teachers play a critical role in this process. Research has shown that the effect of self-assessment on YLLs' learning can be limited when teachers do not see the value of directly using assessment for learning purposes (Butler and Lee, 2010). It has been suggested that it is important for teachers to explicitly discuss with their students the purpose and potential benefits of conducting self-assessment, to assist them with self-constructing self-assessment items, to create positive learning environments where students feel comfortable with self-reflecting, and to provide individualized and encouraging feedback (e.g., Dann, 2002; Wong and Mak, 2019). How best to undertake these activities with YLLs remains unclear, however. Implementing peer assessment along with self-assessment can be helpful (Hung *et al*., 2016), perhaps because peer assessment can offer YLLs an opportunity to better understand assessment criteria when reflecting on their own performance.

Reconciling children's diverse needs with demands for standardization

Assessment for YLLs should reflect the wide range of language development and academic achievement experienced by this age group. Any administration and interpretation of standardized tests for YLLs need to be carried out with great care and caution.

Reflecting the global popularity of learning English, a number of standardized international English proficiency tests are available for YLLs, including the Cambridge English Qualifications Designed for Young Learners, TOEFL Primary® and TOEFL Junior® tests, and Pearson's PTE Young Learners.[1] Recent efforts to validate these tests provide useful information on test development and interpretation of the test results. Winke *et al*. (2018), for example, examined the cognitive validity of the Cambridge Michigan Language Assessment through drawings and interviews obtained from children ages seven to nine. They identified some construct-irrelevant variables contributing to the children's incorrect answers, including their unfamiliarity with assessment tasks, mismatched background knowledge,

and age-inappropriate cognitive demands. Based on the results, Winke *et al*. addressed the importance of rigidly examining cognitive validity by interviewing children – and doing so even for a psychometrically established test. Many of the standardized proficiency tests are aligned with international standards such as CEFR. It is important to note, however, that test users "should not misinterpret any type of linking [with international standards] as a sufficient indicator of the overall quality of an assessment or as confirmation of the validity of its scores for their intended use," as Papageorgiou and Baron (2017: 148) warned. Moreover, it seems that these standardized tests are frequently administered to younger children (younger than the tests specify); it is therefore important to interpret the scores of younger children with substantial caution.

Using standardized test results for YLLs who receive their academic studies in their additional language requires even more careful consideration because of potentially substantial impacts on their academic studies as well as their language learning. In some regions, such as the United States and the United Kingdom, where educational policies push standardization of curriculum and assessment with high accountability requirements, various problems have been articulated in making decisions about YLLs based on standardized tests (e.g., Carroll and Bailey, 2016; Menken *et al*., 2014; Schissel *et al*., 2019; Takanishi and Le Menestrel, 2017). Students' performance in their L1 (or home language) can predict their later academic performance, which in turn can have significant consequences, such as reclassification of their language-learning status. Information on L1 performance can also provide teachers with valuable information about the kinds of resources that YLLs bring to the classroom (Takanishi and Le Menestrel, 2017). Despite the importance of assessing YLLs' language abilities in both their L1 and their target language, however, the majority of YLLs are assessed solely in their target language. Unfortunately, with just a few exceptions, psychometrically compatible tests in their home language(s) are unavailable to YLLs. Translating a test into a different language does not necessarily ensure the same level of validity and reliability as the original test. Moreover, even if tests in the students' L1 are available, the "equivalent" assessments in non-target languages are often normed against monolingual speakers of the given languages; the norming is performed based on different groups of students, which can produce potentially misleading results. Indeed, test misuse or use of inappropriate L1 assessments has resulted in children being misclassified as "nonproficient" in their L1; when these children were also classified as nonproficient in their L2, they were at risk of being placed in special education by mistake (Artiles *et al*., 2005).

Various test accommodations have been suggested and implemented to remove or minimize measurement-irrelevant sources of difficulty that YLLs may encounter when taking monolingual, standardized, academic achievement tests for accountability purposes (Abedi *et al*., 2004). Test accommodations can be achieved by modifying the test presentation/response methods (e.g., using plainer language without changing the content of the test) or the procedures (e.g., allowing students extra time for the test). However, such modifications can be cause for concern. In the United States, for example, accommodations are often employed inconsistently and without any basis in empirical evidence, and the "effectiveness" of the accommodation appears to be largely learner and context dependent (Abedi *et al*., 2004; Rivera and Collum, 2006).

More critically, however, the very assumption underlying test accommodation for YLLs – that the test should be able to measure the same ability for YLLs and their monolingual counterparts as long as the potential sources of difficulties for YLLs are removed – can be challenged. If one accepts the view that bilinguals'/multilinguals' language abilities are *qualitatively* different from those of monolinguals (e.g., Cook, 1992), it would be meaningless

to administer tests to bilinguals/multilinguals that are based on monolingual constructs and normed accordingly to monolinguals. Similarly, as mentioned already, researchers such as Chalhoub-Deville (2019), who advocate a dynamic and fluid view of multilingual competence, have argued for assessment practice resting on *multilingual constructs* that take all plurilingual resources into account. While such an approach may require "a disruptive innovation" (Chalhoub-Deville, 2019: 277) in assessment, implementation of this approach remains largely experimental (e.g., Guzman-Orth *et al*., 2019).

The relationship between language and academic content is complicated, and researchers continue to grapple with the concept of *academic language* – or language needed for academic studies. In addition to the sociocultural approaches to language competence mentioned earlier, cognitive and linguistic approaches also have been prevalent. Cummins's (1979) distinction between *basic interpersonal communication skills* (BICS) and *cognitive academic language proficiency* (CALP) is an influential cognitive approach that has been widely accepted in education. But this approach has also been heavily criticized for its dichotomous conceptualization of social and academic languages and its anticipated cognitive consequences based on the distinction. Linguistic approaches, including Scarcella (2003) and Schleppegrell (2004, based on systemic functional linguistics), that describe academic languages as a series of domain- or register-specific linguistic features have also been impactful; test developers can use the types of linguistic features that should be expected in academic contexts as a guideline. Boals and his colleagues (2015) observed, however, that the current inventories of linguistic features identified by linguists are not sufficient and have suggested the need for "a comprehensive description addressing modalities of language use, age of students, and contexts of use" (p. 133). Some other researchers, instead of theorizing academic language, have taken a practical approach in which they have examined language demands realized in standards, curriculum, textbooks, and other material (e.g., Wolf *et al*., 2008). In any event, it is important to note that different conceptualizations of academic language lead to different operationalizations in assessment practice, and this leads to different decisions (e.g., reclassification). As YLLs' learning environments and needs have become increasingly diversified, more *contextualized* research is necessary to better understand the relationship between language and academic content in curriculum and assessment practices. For example, the linguistic and academic demands for children in Content Language Integrated Learning (CLIL) programs in Europe may be different from those in immersion programs in North America. It is also worth noting that current research on academic language is almost exclusively conducted on English, and little information is available for other languages.

Finally, when it comes to identifying YLLs with disabilities, using assessments normed against monolingual children or conducting assessments solely in YLLs' target language can have serious consequences. Specifically, such use results in frequent *overrepresentation* (YLLs classified in special education when they do not need to be) and *underrepresentation* (YLLs not classified in special education when they should be) of YLLs in special education (e.g., Ballantyne, 2013, for a case in the United States). Both cases are problematic because they leave students without adequate supports. When identifying YLLs' disabilities, Kangas argued for the importance of using multiple data sources and considering individual students' specific linguistic and academic experiences (Kangas, 2019).

Among the data sources that educators can rely on to diagnose disabilities in YLLs, a particularly important source is YLLs' L1 (or home language) performance data. If difficulties are observed in a child's L1 *and* L2, it is reasonable to assume that the difficulties may come from disabilities. It is important to note that just because an assessment does not

signal a disability does not rule out the possibility of one (Kormos, 2017). We have limited knowledge about how L1 and L2 difficulties overlap. Moreover, in reality, valid and reliable diagnostic measures in YLLs' L1 are not always available. The challenge of making an accurate diagnosis also influences decisions related to initial classification and reclassification, although research is limited on reclassification among YLLs with disabilities. In the United States, for example, there is no evidence-based uniform criteria for reclassifying these students (Kangas, 2019). Burke *et al*. (2016) reported a case in Indiana in which YLLs with disabilities were 59 percent less likely to be reclassified than YLLs without disabilities. It is important to keep in mind that YLLs with disabilities are highly heterogenous; their difficulties can be cognitive, social, physical, emotional, or any combination of these. Therefore, individualized assistance is crucial. Researchers have just begun to explore various types of individualized assessment in assisting YLLs with disabilities, including dynamic assessment (Peterson and Gillam, 2013) and cognitive assessment (Smith, 2013).

Recommendations for practice

To conduct age-appropriate, learning-focused, and flexible assessments that meet the diverse needs of YLLs, teachers must obtain sufficient understanding of how assessment works, how to develop and administer assessment validly and reliably, and how to interpret the results in order to best assist individual student learning. Such knowledge and skills, or *language assessment literacy*, are considered a critical element in teacher training (Brindley, 2001). A number of models of language assessment literacy have been proposed; they differ in terms of their epistemological traditions as well as how they configure and weight various elements, but they are all multifaceted. The models consider it to be indispensable for teachers not only to obtain knowledge about assessment principles but also to develop practical skills for conducting assessment and to understand the social consequences of assessment (Lee and Butler, 2020; Harding and Kremmel, 2016; Téllez and Mosqueda, 2015).

Among the different language assessment literacy models, Edelenbos and Kubanek-German's (2004) concept of teachers' *diagnostic competence* is of particular relevance to assessing YLLs. Based on observations of teachers' practices in primary school foreign language classrooms in the Netherlands and Germany, as well as on interviews with those teachers, the authors proposed that teachers need to develop strong "abilities to interpret students' foreign language growth, to skillfully deal with assessment material and to provide students with appropriate help in response to this diagnosis" (p. 260). Edelenbos and Kubanek-German identified a series of diagnostic competencies for teachers, including detecting the meaning of children's silence and facial expressions, articulating children's language growth with concrete examples, administering both quantifiable and nonquantifiable assessments while understanding the limitations of each, grouping children based on their individual needs, and analyzing assessment materials (p. 279). More research is needed to validate these competencies. However, considering that YLLs' teachers tend not to make sufficient use of formative assessment in their classrooms (e.g., Becker, 2015; Rixon, 2018; Rixon and Papp, 2018), it is reasonable to emphasize teachers' abilities to diagnose YLLs' needs and detect their progress. It is important for teachers to adjust their attitudes toward assessment (i.e., from assessment *of* learning to assessment *for* learning) and to create classroom atmospheres that promote assessment for learning.

To avoid or minimize the misuse of assessment results, particularly results based on standardized assessments, teachers must learn to interpret assessment results critically and with a high degree of sophistication and to consider the social consequences of such results.

As exemplified in the case of classification and reclassification of YLLs with disabilities in the United States, mentioned earlier, teachers play a critical role in various decision-making processes; any high-stakes decisions should be made based on multiple assessment sources and in consultation with various stakeholders, including YLLs themselves. Teachers also need to be mindful of potential washback effects of any assessment that they use. In teaching contexts in which standardized assessment-based accountability is required by policies, it is not uncommon for teachers to focus on the content and response methods of those assessments (e.g., a multiple-choice response method) in their lessons. Even when such practices are unintended, they send a strong message to students about what needs to be learned and how to go about learning it (Rixon, 2019; see also Cheng and Sultana, this volume). Assessment results and adults' attitudes toward assessment can make a strong impression on YLLs. Negative experiences with assessment, especially at an early stage of their learning, often lead them to attribute their performance to ability rather than effort, which may have long-lasting negative impacts on their motivation, self-esteem, and confidence as well as later learning (e.g., Perry and Van de Kamp, 2000).

Future directions

As mentioned before, research on YLLs' assessment is blooming. There are many exciting directions for future research. Considering the rapid advancement of digital technology and its potentially huge impact on young learners' use and learning of language, I focus on the role of digital technology in assessment in this section (see also Sawaki, this volume).

Many children – including YLLs – are immersed in environments rich in digital media and technology. In the United States, as of 2015, 35 percent of babies before the age of one had touched tablet screens, 15 percent had used applications, and 12 percent had played games. By the age of two, 26 percent of children had used digital devices, including tablets, smart phones, PCs, and laptops; that number shot up to 38 percent by the time children were four years old (Sifferlin, 2015). A large-scale multinational study in Europe revealed similar levels of early exposure to digital devices for European children (Chaudron *et al*., 2017). Children ages zero to eight quickly learned how to interact with digital technology and used it for information gathering, learning, and communication as well as entertainment and creation.

It is reasonable to assume that this early exposure to and use of digital technology has a substantial influence on YLLs cognitive, social, and affective development, which in turn has implications for assessment practice. Thus, there is a great need for basic research on technology's effect on children's language-related interaction.

Cognitively, we already have empirical evidence indicating that the use of digital technology can enhance children's cognitive skills and brain functions (Firth *et al*., 2019; Giacomo *et al*., 2017). The extent to which and in what circumstances (e.g., the timing of use) technology influences children's cognitive processing and strategies (e.g., memory and attention) remain unclear, however. We have limited information on YLLs' cognitive processing across age groups when they multitask using digital technology. When developing assessment tasks and conducting cognitive validity research, we should go beyond the age factor alone to consider the interaction between age and technology. For example, a certain type of technology may be more suitable to a certain age group than other age groups. We should also attend to individual differences in cognitive processing and strategies during computer-based assessment. For instance, Lee and Winke's (2018) eye-tracking study revealed that YLLs, compared with their monolingual counterparts, tended to fixate

longer and more frequently on a countdown timer, which is a distractor, instead of looking at onscreen pictures meant to assist their performance. Computer literacy skills certainly play a substantial role in YLLs' performance in computer-based assessment as well (Papp and Walczak, 2016).

Digital technology has also changed how we communicate. Communication with interactive technology (e.g., communicating with artificial intelligence, AI) is no longer uncommon, at least in some educational contexts. When implementing avatars or AI in assessment, we need to have a better understanding about how YLLs interact with AI and engage in tasks that involve AI. We know that attractive avatars change adult's interactive behaviors and social-linguistic competence (e.g., Yee and Bailenson, 2007), but we know little about how children interact with avatars and other AI, such as robots. Vogt *et al*. (2017) recommended that, in order to assist second language learning among preschool children, robots should act as a knowledgeable peer who can provide children with scaffolding in an adult-like manner and should build trust and common ground with children. Such evidence-based information has implications for using AI as an assessment tool.

YLLs' affects toward digital-based communication – their communication through technology and with technology – is also an important topic with applications for assessment. Research to date has generally reported that YLLs have positive attitudes toward computer-based assessment (Courtney and Graham, 2019; Forsyth *et al*., 2019; Kormos *et al*., 2020; Papp and Walczak, 2016). For example, Courtney and Graham (2019), a large-scale study with 3,437 YLLs (ages 8–13), reported that the children overall enjoyed a digital game–based assessment. However, we need to keep in mind that children are sensitive to the pragmatic oddness that computer-based assessment may create. It has been reported, for instance, that YLLs feel anxious when they are asked to speak to a computer and that their anxiety is due to the absence of both an interlocutor and spontaneous feedback (Kormos *et al*., 2020; Lee and Winke, 2018). YLLs' motivation and affects, and how such mental factors influence their performance in computer-based assessment, appear to be complex (Kormos *et al*., 2020) and warrant more investigation.

Finally, we must remember that technology does not replace the role of teachers. As technology allows more individualized assessment and instant feedback to both teachers and YLLs (e.g., Poehner *et al*., 2017 for a computerized dynamic assessment), research is needed concerning how teachers can best use such individualized information to support student learning. Moreover, technology is good at evaluating certain types of language knowledge and skills (e.g., pronunciation and use of vocabulary) but not others (discourse structures). A potentially fruitful area of future research is how best to incorporate assessment information obtained through digital technology into human-based assessment information, with the goal of obtaining a comprehensive understanding of students' learning both in and outside of the classroom.

Further reading

Nikolov, M. (ed.). (2016). *Assessing Young learners of English: Global and Local Perspectives*. New York, NY: Springer.

Wolf, M. and Butler, Y. G. (eds.). (2017). *English Language Proficiency for Young Learners*. New York, NY: Routledge.

Prošić-Santovac, D. and Rixon, S. (eds.). (2019). *Integrating Assessment into Early Language Learning and Teaching*. Bristol, UK: Multilingual Matters.

The three edited books gather recent research on assessment for YLLs, but each book has a slightly different focus. Nikolov (2016) considers different assessment practices around the world and discusses the unique challenges associated with assessing YLLs. Wolf and Butler (2017) offer theoretical and empirical accounts for developing and validating assessment for YLLs, mainly focusing on large-scale assessments in English. Prošić-Santovac and Rixon (2019) turn their attention to integrating assessment with pedagogy in the context of foreign language education in Europe.

Papp, S. and Rixon, S. (eds.). (2018). *Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners*. Cambridge, UK: Cambridge University Press.
As part of the Studies in Language Testing Series of Cambridge English Assessment, this volume focuses on YLLs (ages 6–16) and surveys how language assessment constructs for YLLs have been conceptualized and operationalized in language tests. This is a highly comprehensive and readable book, particularly for testing development professionals.

Nikolov, M. and Timpe-Laughlin, V. (2020). Assessing young learners' foreign language abilities. *Language Teaching*. https://doi.org/10.1017/s0261444820000294.
This is a state-of-the-art review article on assessment of young learners that focuses on foreign language contexts. The review offers a critical and yet accessible review on the major issues concerning assessments for young foreign language learners. The paper concludes with ten questions for future directions, which can be an excellent guide for both researchers and practitioners.

Note

1 Detailed information about these tests is available on their websites: Cambridge English Qualifications Designed for Young Learners (https://www.cambridgeenglish.org/exams-and-tests/qualifications/schools/), TOEFL Primary® and Junior® tests (www.ets.org/toefl_primary, www.ets.org/toefl_junior, respectively), and Pearson's PTE Young Learners (https://qualifications.pearson.com/en/qualifications/international-certificate/young-learners.html).

References

Abedi, J., Hofstetter, C. H. and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research* 74: 1–28. https://doi.org/10.3102/00346543074001001.

Alexiou, T., Roghani, S. and Milton, J. (2019). Assessing the vocabulary knowledge of preschool language learners. In D. Prošić-Santovac and S. Rixon (eds.), *Integrating Assessment into Early Language Learning and Teaching*. Bristol, UK: Multilingual Matters, 207–20.

Artiles, A. J., Rueda, R., Salazar, J. J. and Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children* 71: 283–300. https://doi.org/10.1177/001440290507100305.

Bailey, A. L. (2017). Progressions of a new language: Characterizing explanation development for assessment with young language learners. *Annual Review of Applied Linguistics* 37: 241–63. https://doi.org/10.1017/s0267190517000113.

Ballantyne, K. G. (2013). Disproportional representation of English learners among students identified with disabilities: Equity and the federal accountability system. In D. Tsagari and G. Spanoudis (eds.), *Assessing L2 Students with Learning and Other Disabilities*. Newcastle upon Tyne, UK: Cambridge Scholars, 3–25.

Becker, C. (2015). Assessment and portfolios. In J. Bland (ed.), *Teaching English to Young Learners: Critical Issues in Language Teaching with 3–12 Year Olds*. London, UK: Bloomsbury, 261–78.

Benigno, V. and de Jong, J. (2016). The "global scale of English learning objectives for young learners": A CEFR-based inventory of descriptors. In M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*. New York, NY: Springer, 43–64.

Black, P. J. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education* 5: 7–74. https://doi.org/10.1080/0969595980050102.

Boals, T., Kenyon, D. M., Blair, A., Cranley, M. E., Wilmes, C. and Wright, L. J. (2015). Transformation in K-12 English language proficiency assessment: Changing contexts, changing constructs. *Review of Research in Education* 39: 122–64. https://doi.org/10.3102/0091732x14556072/.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara and K. O'Loughlin (eds.), *Experimenting with Uncertainty: Essays in Honour of Alan Davies*. Cambridge, UK: Cambridge University Press, 126–36.

Burke, A. M., Morita-Mullaney, T. and Smith, M. (2016). Indiana emerging bilingual student time to reclassification: A survey analysis. *American Educational Research Journal* 53: 1310–42. https://doi.org/10.3102/0002831216667481.

Butler, Y. G. (2016). Self-assessment of and for young learners' foreign language learning. In M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*. New York, NY: Springer, 291–315.

———. (2018a). The role of context in young learners' processes for responding to self-assessment items. *The Modern Language Journal* 102: 242–61. https://doi.org/10.1111/modl.12459.

———. (2018b). Young learners' processes and rationales for responding to self-assessment items: Cases for generic can-do and five-point Likert-type formats. In J. Davis *et al*. (eds.), *Useful Assessment and Evaluation in Language Education*. Washington, DC: Georgetown University Press, 21–39.

———. (2019). Assessment of young English learners in instructional settings. In A. Gao (ed.), *Second Handbook of English Language Teaching*. New York, NY: Springer, 477–95.

Butler, Y. G. and Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *The Modern Language Journal* 90: 506–18. https://doi.org/10.1111/j.1540-4781.2006.00463.x.

———. (2010). The effect of self-assessment among young learners. *Language Testing* 27: 5–31. https://doi.org/10.1177/0265532209346370.

Butler, Y. G. and Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessment. *Language Assessment Quarterly* 11: 45–75. https://doi.org/10.1080/15434303.2013.869814.

———. (2015). Young learners' interactional development in task-based paired-assessment in their first and foreign languages: A case of English learners in China. *Education 3–13* 43: 292–321. https://doi.org/10.1080/03004279.2013.813955.

Carroll, P. E. and Bailey, A. L. (2016). Do decision rules matter? A descriptive study of English language proficiency assessment classifications for English-language learners and native English speakers in fifth grade. *Language Testing* 33: 23–52. https://doi.org/10.1177/0265532215576380.

Chalhoub-Deville, M. B. (2019). Multilingual testing constructs: Theoretical foundations. *Language Assessment Quarterly* 16: 472–80. https://doi.org/10.1080/15434303.2019.1671391.

Chaudron, S., Di Gioia, R. and Gemo, M. (2017). *Young Children (0–8) and Digital Technology: A Qualitative Study Across Europe*. Retrieved from https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/young-children-0-8-and-digital-technology-qualitative-study-across-europe.

Choi, I., Wolf, M. K., Pooler, E., Sova, L. and Faulkner-Bond, M. (2019). Investigating the benefits of scaffolding in assessments of young English learners: A case for scaffolding retell tasks. *Language Assessment Quarterly* 16: 161–79. https://doi.org/10.1080/15434303.2019.1619180/.

Cook, V. J. (1992). Evidence for multicompetence. *Language Learning* 42: 557–91. https://doi.org/10.1111/j.1467-1770.1992.tb01044.x.

Council of Europe. (2001/2018). *The Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Retrieved from www.coe.int/en/web/common-european-framework-reference-languages/home.

———. (2019). *ELP Check Lists for Young Learners*. Retrieved from www.coe.int/en/web/portfolio/the-language-passport.

Courtney, L. and Graham, S. (2019). "It's like having a test but in a fun way": Young learners' perceptions of a digital game-based assessment on early language learning. *Language Teaching for Young Learners* 1: 161–86. https://doi.org/10.1075/ltyl.18009.cou.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism* 19: 121–29. Retrieved from https://files.eric.ed.gov/fulltext/ED184334.pdf.

Dann, R. (2002). *Promoting Assessment as Learning: Improving the Learning Process*. New York, NY: Routledge.

Edelenbos, P. and Kubanek-German, A. (2004). Teacher assessment: The concept of "diagnostic competence." *Language Testing* 21: 259–83. https://doi.org/10.1191/0265532204lt284oa.

Ellis, G. and Rixon, S. (2019). Assessment for learning with younger learners: Is thinking about their learning a step too far? In D. Prošić-Santovac and S. Rixon (eds.), *Integrating Assessment into Early Language Learning and Teaching*. Bristol, UK: Multilingual Matters, 87–104.

Enever, J. (ed.). (2011). *ELLiE: Early Language Learning in Europe*. London, UK: The British Council.

———. (2018). *Policy and Politics in Global Primary English*. Oxford, UK: Oxford University Press.

Firth, J., Torous, J., Stubbs, B., Firth, J. A., Steiner, G. Z., . . . Sarris, J. (2019). The "online brain": How the Internet may be changing our cognition. *World Psychiatry* 18: 119–29. https://doi.org/10.1002/wps.20617.

Forsyth, C. M., Luce, C., Zapata-Rivera, D., Jackson, G. T., Evanini, K. and So, Y. (2019). Evaluating English language learners' conversations: Man vs. machine. *Computer Assisted Language Learning* 32: 398–417. https://doi.org/10.1080/09588221.2018.1517126.

García Mayo, M. P. and Agirre, A. I. (2016). Task repetition and its impact on EFL children's negotiation of meaning strategies and pair dynamics: An exploratory study. *The Language Learning Journal* 44: 451–66. https://doi.org/10.1080/09571736.2016.1185799.

Giacomo, D. D., Ranieri, J. and Lacasa, P. (2017). Digital learning as enhanced learning processing? Cognitive evidence for new insights of smart learning. *Frontiers in Psychology* 8: 1329. https://doi.org/10.3389/fpsyg.2017.01329.

Ghosn, I. K. (2019). Integrating developmentally appropriate assessment with instruction in the young learner classroom. In D. Prošić-Santovac and S. Rixon (eds.), *Integrating Assessment into Early Language Learning and Teaching*. Bristol, UK: Multilingual Matters, 52–68.

Goodier, T. and Szabo, T. (2018). *Collated Representative Samples of Descriptors of Language Competences Developed for Young Learners (Aged 7–10 and 11–15 Years)*. Retrieved from www.coe.int/en/web/common-european-framework-reference-languages/bank-of-supplementary-descriptors.

Gu, L. and Hsieh, C. N. (2019). Distinguishing features of young English language learners' oral performance. *Language Assessment Quarterly* 16: 180–95. https://doi.org/10.1080/15434303.2019.1605518.

Guzman-Orth, D. A., Lopez, A. A. and Tolentino, F. (2019). Exploring the use of a dual language assessment task to assess young English learners. *Language Assessment Quarterly* 16: 447–63. https://doi.org/10.1080/15434303.2019.1674314.

Harding, L. and Kremmel, B. (2016). Teacher assessment literacy and professional development. In D. Tsagari and J. Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin, Germany: De Gruyter, 413–28.

Hsieh, C. N. and Wang, Y. (2019). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing* 36: 27–50. https://doi.org/10.1177/0265532217734240.

Huang, B. and Butler, Y. G. (2020). Validity considerations for assessing language proficiency for young language minority students. *Language Assessment Quarterly*. https://doi.org/10.1080/15434303.2020.1826486.

Huang, F. L. and Konold, T. R. (2014). A latent variable investigation of the phonological awareness literacy screening-kindergarten assessment: Construct identification and multigroup comparisons between Spanish-speaking English-language learners (ELLs) and non-ELL students. *Language Testing* 31: 205–21. https://doi.org/10.1177/0265532213496773.

Hung, Y., Samuelson, B. L. and Chen, S. (2016). Relationship between peer- and self-assessment and teacher assessment of young EFL learners' oral presentations. In M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*. New York, NY: Springer, 317–38.

Jang, E. E. (2008). A framework for cognitive diagnostic assessment. In C. A. Chapelle, Y. R. Chung and J. Xu (eds.), *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment*. Ames, IA: Iowa State University, 117–31.

———. (2014). *Focus on Assessment*. Oxford, UK: Oxford University Press.

Jang, E. E., Dunlop, M., Wagner, M., Kim, Y.-H. and Gu, Z. (2013). Elementary school ELLs' reading skill profiles using cognitive diagnosis modeling: Roles of length of residence and home language environment. *Language Learning* 63: 400–36. https://doi.org/10.1111/lang.12016.

Jang, E. E., Vincett, J. M., van der Boom, E. H., Lau, C. and Yang, Y. (2017). Considering young learners' characteristics in developing diagnostic assessment intervention. In M. K. Wolf and Y. G. Butler (eds.), *English Language Proficiency Assessments for Young Learners*. New York, NY: Routledge, 193–213.

Kangas, S. E. N. (2019). English learners with disabilities: Linguistic development and educational equity in jeopardy. In X. Gao (ed.), *Second Handbook of English Language Teaching*. New York, NY: Springer, 919–37.

Keane, L. and Griffin, C. P. (2018). Assessing self-assessment: Can age and prior literacy attainment predict the accuracy of children's self-assessment in literacy? *Irish Educational Studies* 37: 127–47. https://doi.org/10.1080/03323315.2018.1449001.

Kormos, J. (2017). The effects of specific learning difficulties on processes of multilingual language development. *Annual Review of Applied Linguistics* 37: 30–44. https://doi.org/10.1017/s026719051700006x.

Kormos, J., Brunfaut, T. and Michel, M. (2020). Motivation factors in computer-administered integrated skills tasks: A study of young learners. *Language Assessment Quarterly* 17: 43–59. https://doi.org/10.1080/15434303.2019.1664551.

Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics* 1: 49–72. https://doi.org/10.1558/japl.1.1.49.55872.

Lee, J. and Butler, Y. G. (2020). Reconceptualizing language assessment literacy: Where are language learners? *TESOL Quarterly* 54: 1098–111. https://doi.org/10.1002/tesq.576.

Lee, S. and Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing* 35: 239–69. https://doi.org/10.1177/0265532217704009.

Łockiewicz, M., Sarzała-Przybylska, Z. and Lipowsk, M. (2018). Early predictors of learning a foreign language in pre-school: Polish as a first language, English as a foreign language. *Frontiers in Psychology* 9: 1813. https://doi.org/10.3389/fpsyg.2018.01813.

McKay, P. (1995). Developing ESL proficiency descriptions for the school context. In G. Brindley (ed.), *Language Assessment in Action*. Sydney: National Center for English Language Teaching and Research, 31–63.

———. (2006). *Assessing Young Language Learners*. Cambridge, UK: Cambridge University Press.

Menken, K., Hudson, T. and Leung, C. (2014). Symposium: Language assessment in standards-based education reform. *TESOL Quarterly* 48: 586–614. https://doi.org/10.1002/tesq.180.

Mihaljević Djigunović, J. (2019). Affect and assessment in teaching L2 to young learners. In D. Prošić-Santovac and S. Rixon (eds.), *Integrating Assessment into Early Language Learning and Teaching*. Bristol, UK: Multilingual Matters, 19–33.

Mourão, S. and Lourenço, M. (2015). *Early Years Second Language Education: International Perspectives on Theory and Practice*. New York, NY: Routledge.

Nikolov, M. (2016). *Assessing Young Learners of English: Global and Local Perspectives*. New York, NY: Springer.

Nikolov, M. and Timpe-Laughlin, V. (2020). Assessing young learners' foreign language abilities. *Language Teaching*. https://doi.org/10.1017/s0261444820000294.

North, B. (2014). *The CEFR in Practice*. Cambridge, UK: Cambridge University Press.

Oliver, R. and Azkarai, A. (2019). Patterns of interaction and young ESL learners: What is the impact of proficiency and task type? *Language Teaching for Young Learners* 1: 82–102. https://doi.org/10.1075/ltyl.00006.oli.

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing* 6: 1–13. https://doi.org/10.1177/026553228900600103.

Papageorgiou, S. and Baron, P. (2017). Using the common European framework of reference to facilitate score interpretations for young learners' English language proficiency assessments. In M. K. Wolf and Y. G. Butler (eds.), *English Language Proficiency Assessments for Young Learners*. New York, NY: Routledge, 136–52.

Papp, S. (2018). Test taker characteristics. In S. Papp and S. Rixon (eds.), *Examining Young Learners: Research and Practice in Assessing the English of School-Age Children*. Cambridge, UK: Cambridge University Press, 70–127.

Papp, S. and Rixon, S. (eds.). (2018). *Examining Young Learners: Research and Practice in Assessing the English of School-Age Learners*. Cambridge, UK: Cambridge University Press.

Papp, S. and Walczak, A. (2016). The development and validation of a computer-based test of English for young learners: Cambridge English young learners. In M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*. New York, NY: Springer, 139–90.

Pearson Education. (2018). *The Global Scale of English Learning Objectives for Young Learners*. Harlow, UK: Pearson.

Perry, N. E. and Van de Kamp, K. J. O. (2000). Creating classroom contexts that support young children's development of self-regulated learning. *International Journal of Educational Research* 33: 821–43. https://doi.org/10.1016/s0883-0355(00)00052-5.

Peterson, D. B. and Gillam, R. B. (2013). Predicting reading ability for bilingual Latino children dynamic assessment. *Journal of Learning Disabilities* 48: 3–21. https://doi.org/10.1177/0022219413486930.

Pinter, A. (2015). Task-based learning with children. In J. Bland (ed.), *Teaching English to Young Learners: Critical Issues in Language Teaching with 3–12 Year Olds*. London, UK: Bloomsbury, 113–27.

Poehner, M. E., Zhang, J. and Lu, X. (2017). Computerized dynamic assessment for young language learners. In M. K. Wolf and Y. G. Butler (eds.), *English Language Proficiency Assessments for Young Learners*. New York, NY: Routledge, 214–33.

Prošić-Santovac, D. and Navratil, A. (2019). Assessment of very young learners of English and the joy of puppetry: A multiple case study. In D. Prošić-Santovac and S. Rixon (eds.), *Integrating Assessment into Early Language Learning and Teaching*. Bristol, UK: Multilingual Matters, 170–87.

Prošić-Santovac, D. and Rixon, S. (eds.). (2019). *Integrating Assessment into Early Language Learning and Teaching*. Bristol, UK: Multilingual Matters.

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal 100*(supplement): 190–208. https://doi.org/10.1111/modl.12308.

Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing* 17: 115–22. https://doi.org/10.1177/026553220001700201.

Rivera, C. and Collum, E. (2006). *State Assessment Policy and Practice for English Language Learners: A National Perspective*. New York, NY: Routledge.

Rixon, S. (2018). Consequential validity of tests of English for young learners: The impact of assessment on young learners. In S. Papp and S. Rixon (eds.), *Examining Young Learners: Research and Practice in Assessing the English of School-Age Children*. Cambridge, UK: Cambridge University Press, 547–87.

———. (2019). Developing language curricula for young language learners. In X. Gao (ed.), *Second Handbook of English Language Teaching*. New York, NY: Springer, 277–95.

Rixon, S. and Papp, S. (2018). The educational contexts of the teaching of English to young learners and the roles of assessment. In S. Papp and S. Rixon (eds.), *Examining Young Learners: Research and Practice in Assessing the English of School-Age Children*. Cambridge, UK: Cambridge University Press, 18–69.

Scarcella, R. (2003). *Academic English: A Conceptual Framework* (Technical Report No. 2003-1). Irvine, CA: University of California Linguistic Minority Research Institute.

Schissel, J. L. (2019). *Social Consequences of Testing for Language-Minoritized Bilinguals in the United States*. Clevedon, UK: Multilingual Matters.

Schissel, J. L., Leung, C. and Chalhoub-Deville, M. (2019). The construct of multilingualism in language testing. *Language Assessment Quarterly* 16: 373–78. https://doi.org/10.1080/15434303.2019.1680679.

Schleppegrell, M. J. (2004). *The Language of Schooling: A Functional Linguistics Perspective*. New York, NY: Lawrence Erlbaum.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? Are they fair? *Language Testing* 14: 340–49. https://doi.org/10.1177/026553229701400310.

Sifferlin, A. (2015, April 25). 6-month-old babies are now using tablets and smartphones. *TIME*. Retrieved October 11, 2019, from https://time.com/3834978/babies-use-devices/.

Smith, A. M. (2013). Developing cognitive assessments for multilingual learners. In D. Tsagari and G. Spanoudis (eds.), *Assessing L2 Students with Learning and Other Disabilities*. Newcastle upon Tyne, UK: Cambridge Scholars, 151–67.

Takanishi, R. and Le Menestrel, S. (2017). *Promoting the Educational Success of Children and Youth Learning English: Promising Futures*. Washington, DC: The National Academic Press.

Teachers of English to Speakers of Other Languages (TESOL). (1998). *Managing the Assessment Process: A Framework for Measuring Student Attainment of the ESL Standards*. Alexandria, VA: TESOL.

Téllez, K. and Mosqueda, E. (2015). Developing teachers' knowledge and skills at the intersection of English language learners and language assessment. *Review of Research in Education* 39: 87–121. https://doi.org/10.3102/0091732x14554552.

Timpe-Laughlin, V. and Cho, Y. (2021). Assessing young foreign language learners in global and local contexts. *Language Testing* (special issue).

Torrance, H. and Pryor, J. (1998). *Investigating Formative Assessment*. Milton Keynes, UK: Open University Press.

Vogt, P., de Haas, M., de Jong, C., Baxter, P. and Krahmer, E. (2017). Child-robot interactions for second language tutoring to preschool children. *Frontiers in Human Neuroscience* 11. https://doi.org/10.3389/fnhum.2017.00073.

Vygotsky, L. S. (1978). *Mind and Society: The Development of Higher Mental Processes*. Cambridge, MA: Harvard University Press.

Winke, P., Lee, S., Ahn, J. I., Choi, I., Gui, Y. and Yoon, H. (2018). The cognitive validity of child English language tests: What young language learners and their native-speaking peers can reveal. *TESOL Quarterly* 52: 274–303. https://doi.org/10.1002/tesq.396.

Wolf, M. K. and Butler, Y. G. (eds.). (2017). *English Language Proficiency for Young Learners*. New York, NY: Routledge.

Wolf, M. K., Guzman-Orth, D., Lopez, A., Castellano, K., Himelfarb, I. and Tsutagawa, F. S. (2016). Integrating scaffolding strategies into technology-enhanced assessments of English learners: Task types and measurement models. *Educational Assessment* 21: 157–75. https://doi.org/10.1080/10627197.2016.1202107.

Wolf, M. K., Kao, J. C., Herman, J., Bachman, L. F., Bailey, A. L. Bachman, P. L. . . . Chang, S. M. (2008). *Issues in Assessing English Language Learners: English Language Proficiency Measures and Accommodation Uses: Literature Review* (CRESST Report No. 731). Los Angeles: CRESST.

Wolf, M. K., Lopez, A., Oh, S. and Tsutagawa, F. S. (2017). Comparing the performance of young English language learners and native English speakers on speaking assessment tasks. In M. K. Wolf and Y. G. Butler (eds.), *English Language Proficiency Assessments for Young Learners*. New York, NY: Routledge, 171–87.

Wong, K. M. and Mak, P. (2019). Self-assessment in the primary L2 writing classroom. *The Canadian Modern Language Review/La revue Canadienne des langues vivantes* 75: 183–96. https://doi.org/10.3138/cmlr.2018-0197.

Yee, N. and Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research* 33: 271–90. https://doi.org/10.1111/j.1468-2958.2007.00299.x.

10

Dynamic assessment

Marta Antón and Próspero N. García

Introduction

Dynamic assessment (DA) integrates assessment and instruction with the aim of gauging and advancing the learning potential of individual learners and devising appropriate educational strategies. Conceptually, it is based on sociocultural theory, particularly on Vygotsky's notion of zone of proximal development (ZPD), as well as on European and American psychology of the 1920s and 1930s (Grigorenko, 2009). This chapter intends to familiarize the reader with DA, particularly with its application to the assessment of second language abilities. We will discuss first the theoretical roots of DA and its early development in cognitive psychology. Then we will present some key characteristics and approaches to DA. We will consider some criticisms and challenges as well as differences with similar types of assessment. Much of the chapter is devoted to applications of DA to second language contexts, particularly the role of mediated language experiences in assessment, the use of DA as a diagnostic tool, program and peer/group assessment, and DA delivered through technology. Finally, we delineate some recommendations for practice and future developments for DA.

Theoretical background and historical perspective

DA practices are rooted in Vygotsky's theory of cognitive development and his notion of ZPD (Vygotsky, 1978; Minick, 1987) as the place where learning occurs. Vygotsky's proposals for the use of ZPD in assessment are based on his conception of ZPD, which he defines as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers" (Vygotsky, 1978: 86). The ZPD reflects development because "It is not what one is but what one can become; it is not what has developed but what is developing" (Grigorenko and Sternberg, 1998: 78). It comes into existence through social interaction with others and the self (Negueruela *et al*., 2015). Based on these assumptions, Vygotsky called for a change from *symptomatic assessment*, which focuses on present behavior of a particular

DOI: 10.4324/9781003220756-14

developmental stage, to *diagnostic assessment*, which focuses on future behavior and on developing recommendations to foster development (Minick, 1987). The fundamental link between the notion of ZPD and DA is the central idea that assessment focusing only on what the child is able to do independently ignores important differences in mental functioning that come to light when the child's interaction with an expert is analyzed (Minick, 1987: 120). For Vygotsky, the application of ZPD to assessment allows for a more complete picture of the learner's actual stage of development and of potential development. He was mostly concerned with qualitative assessment of psychological processes and the dynamics of their development.

In the early 1950s, Feuerstein and his colleagues in Israel articulated and developed DA procedures to estimate the learning potential of low-performing children (Feuerstein *et al*., 1979). Since then, different models of DA have been applied to the assessment of cognitive development potential by psychologists. While some followed Feuerstein's interactionist model of DA (Haywood and Lidz, 2007; Lidz, 1987, 1991; Tzuriel, 1997), other interventionist quantitative procedures were developed (Budoff, 1987; Campione and Brown, 1987; Güthke, 1982).

Critical issues and topics

Grigorenko (2009) highlights three basic assumptions that underlie the various methodologies known as DA. First, traditional testing may fail to capture the level of cognitive development of learners from dissimilar cultural backgrounds. Second, the focus of assessment should be not only what learners can do now but also what they may be able to do tomorrow. Third, assessment should be integrated with intervention and should have selecting or modifying intervention as a goal. A major focus of the proponents of DA is the ability to identify those students who are likely to experience difficulties and to provide rich descriptions of the abilities of these students so that remedial programs may be developed (Campione and Brown, 1987).

Lidz (1987) defined DA as "an interaction between an examiner-as-intervener and a learner-as-active participant, which seeks to estimate the degree of modifiability of the learner and the means by which positive changes in cognitive functioning can be induced and maintained" (p. 4). The goal of DA is to measure, intervene, and modify behaviors by concentrating on the process of learning. *Activity* on the part of examiner and learner and *modifiability* of behavior through *mediation* are crucial characteristics of DA (Lidz, 1991). Mediation in DA refers to the intervention of the assessor in order to select, amplify, and interpret objects and processes to the learner during the assessment.

Approaches to DA

The methods employed in DA educational psychology studies and second language learning studies are presented here along the interactionist-versus-interventionist paradigm (Lantolf and Poehner, 2004). Interventionist studies follow a test-intervene-retest format and thus tend to focus on quantitative results of the intervention while interactionist studies rely on qualitative analysis and interpretation of key features of the interaction blending learning and assessment.

Procedurally, interventionist approaches are carried out following what has been called *the sandwich format* or *the cake format*. The first one has three stages. First, the learner is tested in order to establish the level of independent performance. Second, there is an

intervention in the form of teaching what is included in the test. In the final stage, the learner is tested again to assess the degree and the nature of the change. Intervention (second stage) follows a predetermined script, and thus the procedure is standardized. In this approach, there is a tendency to devise quantitative formulae for calculating scores for learning potential derived from gains from the pre-test to the post-test as a result of the intervention stage. In the cake model, assessment and mediation are interspersed in layers, one depending on how the other evolves.

Interactionist approaches, on the other hand, make use of unscripted mediation and depend on the mediator's judgment as well as on the learner's response to the mediation during the assessment. Among the first to employ interactionist DA was Feuerstein, who with his colleagues developed the mediated learning experience (MLE) and learning propensity assessment device (LPAD). In the MLE, there is a "reciprocal, emotional, affective and motivational aspect of the interaction that melds the activity into a meaningful and structural whole, leading to self-awareness, structural change and cognitive development" (Feuerstein, Feuerstein *et al*., 2002: 75).

Finding DA's place in assessment

Research related to DA models in educational psychology has concluded that mediation is associated with improved performance, that verbalization and elaborated feedback seem to be the most powerful components of mediation, that low-performing children benefit from mediated intervention more than other groups, and that DA contributes significantly to variability in achievement (Lidz, 1991). However, concerns about validity and reliability of DA techniques have been raised (Grigorenko and Sternberg, 1998; Sternberg and Grigorenko, 2002) regarding a lack of standardization in the methodology (particularly in interactionist approaches) and a lack of empirical validation of results. Results from one DA study may not be generalized to other situations since the effect of mediation is dependent upon the context and quality of the interaction between mediator and student.

Poehner (2008b) argued extensively against psychometric criticisms of DA, claiming that non-DA constructs are not suitable to the interactionist model of DA because DA represents a very different perspective on human mind functioning. The dialectic relation between mind and culture makes DA incompatible with other approaches that try to observe mental functioning independently from the environment. However, Poehner (2008b) adds that this difference does not make DA an inappropriate means of assessment. It simply evidences that DA researchers need to delineate their own methods and criteria for understanding individuals, not measuring them (Lantolf and Poehner, 2013).

In L2 contexts, recent attention to DA has coincided with a heightened awareness of the role of discourse and social context in learning and development; dissatisfaction with traditional testing for its lack of attention to co-construction of interaction, contextual variability, and social constraints on assessment (McNamara and Roever, 2006); and the practitioners' desire for "useful" assessment (Davis, 2018). Following this trend, DA and other assessment procedures of relevance to teachers and students such as formative assessment (Black and Williams, 1998), assessment for learning (Davison and Leung, 2009), classroom-based assessment (Fox, Abdulhamid and Turner, this volume), learning-oriented assessment (Turner and Purpura, 2015), and diagnostic assessment (Jang and Sinclair, this volume) have been prominently featured among assessment scholars and practitioners. In the case of DA, its basic tenets and procedures are linked to a sociocultural understanding of language development, not as an individual, isolated, cognitive process but as the result of complex

interactions between an individual, the environment, and socio-historical means available in the culture. Scholarly attention to DA has been abundant, as evidenced by published monographs (Poehner, 2008b) and inclusion of DA in handbooks and edited collections related to language assessment, language education, or sociocultural theory in second language teaching (Fulcher and Davidson, 2012; Hornberger, 2007; Lantolf *et al*., 2018). Thus, although use of DA is still not widespread in L2 educational contexts, it clearly has found its place as an established form of assessment with a focus on learner development.

DA and similar approaches

Although other forms of assessment, such as formative assessment, assessment for learning, and diagnostic assessment, resonate with dynamic assessment in their emphasis on formative uses of the assessment, there are marked differences between these approaches. Incidental formative assessment, which takes place during instructional conversations in the classroom and provides feedback to the learner during the activity, is mostly constrained to informal unsystematic procedures. Poehner and Lantolf (2005) maintain that while FA focuses on the completion of the task, DA aims at cognitive development and transfer of skills to future tasks through intentional and systematic mediation. Importantly, Lantolf and Thorne (2006) add that DA is derived from a well-developed theory of learning. DA also has common threads with diagnostic assessment (Antón, 2018) in its focus on learning and the importance of diagnosis in planning future action. However, in DA the ultimate goal is not assisted performance, current-task completion, or evidence of actual ability but fostering potential development with the assistance of mediational tools. As a result, development may be signaled not simply by performing a language task appropriately but also by other means of demonstrating conceptual growth. The motivation to conduct DA, the theoretical principles guiding its assessment protocols, and the interpretation of the results of the assessment reveal fundamental philosophical differences with other seemingly similar approaches (Poehner and Lantolf, 2005).

Current contributions and research in L2 DA

DA has been applied to several areas in language teaching and learning such as teaching language concepts (García, 2012, 2014), sociopragmatics (van Compernolle and Kinginger, 2013), and language skills: writing, reading, listening, speaking (Ableeva, 2008; Hill and Sabet, 2009; Rahimi *et al*., 2015). Its use has been found to be positive in face-to-face as well as in computer-mediated assessments. Although much of the early work involved individual sessions (Antón, 2009; Poehner, 2007), DA can also be implemented during peer and group interaction (Lund, 2008), with an entire class (Davin and Donato, 2013; Poehner, 2009), and through automated computer assessment (Poehner and Lantolf, 2013; Poehner *et al*., 2015; Zhang and Lu, 2019). A selective summary of advances in these areas is presented in the following sections.

Learners with L2 learning problems and at-risk learners

One of the first applications of DA to L2 contexts was related to the instruction of students experiencing L2 learning problems. Schneider and Ganschow (2000) encouraged teachers to teach and assess students' awareness of metalinguistic skills through dialogue. They recommended guiding students through self-discovery with assisting questions and other verbal and nonverbal hints in the use of such strategies as recognizing similarities/

differences within language patterns, developing mnemonic devices, etc. Kozulin and Garb (2002) conducted a study of text comprehension by at-risk EFL adult students in Israel. Following an interventionist model (i.e., test-teach-retest), they assessed the students' ability to learn and use reading comprehension strategies. After students completed the pre-test individually, learners and teachers reviewed it together while focusing on the reading strategies called forth by each item and how to transfer these strategies from one task to the next. The post-test that followed revealed that many students' scores had improved, but not all. Post hoc, the authors devised a formula that differentiated between high– and low–learning potential students. They noted that some students with high and low learning potential obtained the same scores in the pre-test, showing that DA adds important information that remains hidden in static testing. Despite some limitations regarding the reliability of learning potential assessment, the results have instructional value in identifying students in need of further instruction and developing action plans according to individual students' needs. While this topic appeared to be one of great interest in early L2-DA research, few recent studies have applied DA to the assessment of learners with language differences and language impairment. Peña and Greene (2018) report on several studies that illustrate the value of a test-teach-retest approach to DA with mediated language experiences for children in these learner populations. In the reported studies, assessment of each learner's modifiability in several language tasks (word learning, narration, etc.) was highly accurate. However, the authors acknowledge that DA interventions with individual learners, especially in these contexts, require time and training.

Mediated L2 learning experiences

Several studies have focused on the role of interaction in mediated language learning experiences (MLE), highlighting the importance of the quality of interaction and its role in probing and provoking learner development (Poehner, 2018; Poehner and Infante, 2016), learner reciprocity (Ableeva, 2018; Davin, 2018; Poehner, 2008a), and agency (García, 2014; Levi and Poehner, 2018). In an early study on mediation and transcendence (i.e., the learner's ability to move beyond a task and transfer what has been learned to a new activity of increased complexity), Poehner (2007) explored how both these notions are realized in dialogue with advanced learners of French during individual tutoring sessions focused on past tense and aspect in narrations. At the end of the sessions, the initial assessment was given to the learners followed by two transcendence assessments that required learners to transfer and to apply the concepts learned to different tasks. Mediation during transcendence sessions demonstrated differences between learners that were not apparent in assessment sessions, thus providing a richer picture of the level of development attained by the learners in the study. As illustrated next, one of the learners ("Donna") had internalized her conceptual understanding of tense/aspect through previous interactions, to a degree that allowed her to function almost independently in the transcendence session. Transcendence occurred not only in the performance of different tasks but also from one linguistic feature to another. In this particular case, the learner was able to transfer internalization of tense/aspect differences between *passé composé* and *imparfait* to the past perfect.

1 D: . . . elle devenait uh elle avait elle devenait fâchée elle devenait elle a été
. . . she was becoming uh she was having she was becoming she was
2 elle était en colère quelle était la mieux?
she was angry which was the better one?

3 M: well, uh—
4 D: she became angry
5 M: she well uh do you want to use *imparfait* or *passé composé* how do you
6 want to do it?
7 D: she became angry she was being angry she became angry that's what I
8 want to say
9 M: right well um you could use the verb *se fâcher* [but would it change *to be angry*]
10 sort of how you
11 D: (to self) it's a verb
12 M: you know what you're emphasizing if you're using *imparfait* or *passé*
13 *composé* like um if you were saying just here a second ago she got angry
14 D: there was a definite point where she became angry so that would be
15 passé composé
16 M: yeah
17 D: elle s'est fâchée? Elle s'est fâchée et uh juste après ça . . . *she got angry?*
she got angry and uh just after that . . .
18 D: il savait bien qu'il y a quelqu'un qu'il y avait quelqu'un qu'il y avait *he knew well*
that there is someone that there was someone that there was
19 quelqu'un dans l'atelier mais le soldat ne peut* trouver donc tout à fait – *someone*
in the attic but the soldier can't find therefore completely
20 M: il savait bien qu'il y avait quelqu'un dans l'atelier mais il? he knew well that there
was someone in the attic but he?
21 D: il ne pouvait pas trouver il ne pouvait pas le trouver, c'est mieux que il *he*
couldn't find he couldn't find him, that's better than he
22 n'a pas pu le trouver?
c*ouldn't find him?*
23 M: I guess it depends on the meaning right? il ne pouvait pas trouver or il
24 n'a pas pu trouvé, either is grammatical . . .
25 D: je peux faire l'imparfait je crois
I can do the imperfect I believe M: all right
26 D: il ne pouvait pas trouver – *he couldn't find –*
27 M: you see the difference in meaning between the two?
28 D: well he couldn't find him and then he stopped looking for him would
29 be the *passé composé*, *l'imparfait* would be he couldn't find him but
30 there's no it doesn't imply a time when the soldier stopped looking for
31 him
32 M: right so it kind of like depends I think on what you follow it up with

(Poehner, 2007: 329–30)

Drawing from MLE principles, Poehner and Infante (2016) proposed mediated development (MD): a pedagogical model that incorporates a dynamic process of joint thinking in which mediator and learner engage in instructional tasks as dialectical activity. Infante and Poehner (2019) indicate that DA and mediated development (MD) are complementary pedagogical models in which DA focuses on diagnosing learners' emerging abilities, and MD intends to foster development through instruction aimed at the learner's ZPD. In their study, Infante and Poehner (2019) explored the use of DA as in the development of aspectual distinctions in an ESL writing context. They suggested an agentive shift in the assessment process, which is no longer being administered to learners but undertaken by them. Indeed,

the interactions that emerged during the implementation of MD in this study fostered the learner's ability to create alternative meanings in her L2 writing, indicating that DA and MD can be implemented together to generate a more coherent and systematic L2 pedagogical model anchored in sociocultural principles.

Group dynamic assessment and peer assessment

The studies discussed in this section have pushed the boundaries of our understanding of DA as an individual enterprise to explore the possibilities and limitations of realizing DA within groups or as a whole class and using peers instead of teachers as assessors.

The interrelation between teaching and assessment during teacher-led interaction in the classroom has been explored in a number of studies. Poehner (2009) stressed the role of teacher intentionality and group cohesion in whole-class discussions in which the role of group DA during teacher-student exchanges in whole-class interaction is proposed as a way to make interaction "more attuned to learners' emergent abilities" (p. 488). In Poehner's study, group and individual DA followed the same procedures, but in group DA, mediation was geared toward the group ZPD. Central to this idea is the view of the class not as an aggregate of individuals but as a socially constructed entity. In group DA, teachers' questions and feedback, informed by an understanding of the role of interaction in development, followed a structured, predetermined pattern of increasing explicitness according to learners' needs. In this study, as well as in other classroom DA studies (Davin, 2016; Davin and Donato, 2013; Davin and Gómez-Pereira, 2019), the teachers showed intentionality by planning their feedback and keeping track of the degree of explicitness provided to learners during whole-class interaction (see Table 10.1). In this way, teachers were able to work in a systematic way within the group ZPD, and, as a result, it is argued that learners required less explicit feedback, increased participation, and gained control of the language problem at hand.

Lund (2008) argued in favor of shifting "the locus of assessment from the individual to the intersection of the individual and the collective, mediated by cultural tools" (p. 32). The interaction of learning, instruction, and assessment in a collective environment, Lund maintains, provides new opportunities for learning, thus advancing the ZPD. This point is illustrated by two qualitative accounts of collective assessment. In the first one, a group of EFL student teachers collectively graded exam papers. In the second, one peer assessment was performed among young learners. The analysis of overlapping, complementary, and competing contributions to the discourse, together with the cultural artifacts that mediate assessment, such as criteria, guidelines, and policies, provide a view of the inner workings of assessment. Data provided by logs posted by student teacher groups responding to each

Table 10.1 Pre-scripted prompts for classroom DA.

Level of Explicitness	*Prompt*
1 (least explicit)	Pause with questioning look
2	Repetition of entire phrase by teacher with inflection on source of error
3	Repetition of specific site of error
4	Forced choice option
5 (most explicit)	Correct response and explanation provided

Source: Adapted from Davin (2016: 818)

other's evaluation of student papers revealed that guidelines and criteria were appropriated by the participants in different ways, according to their different experiences and historical insights on the process. Negotiation played a major role in reaching a shared understanding of the assessment. In a second episode of collective assessment, videotaped activity while learners prepared for an oral exam and responses to questionnaires by those same learners engaged in group work and peer assessment illustrated the interaction and integration between teaching, learning, and assessment. By involving students in the assessment process, they came to understand criteria and guidelines for evaluation. Collaborative assessment increases the validity of grades by making criteria explicit, and therefore, from a pedagogical standpoint, the opportunities for collective assessment among teachers should be increased.

Diagnostic assessment

DA procedures have also been used with explicit diagnostic purposes. In DA, diagnosis and feedback blend in the same activity. Poehner and van Compernolle (2011) describe the process as diagnosis through intervention since diagnosis/assessment and intervention/teaching are inter-related. DA follows a systematic, well-established approach to diagnosis, grounded in sound empirical constructs. An assessor moves methodically from less implicit to more implicit mediation as deemed necessary in order to determine a level of support appropriate to the learner's needs. An assessment of the degree of mediation needed is essential in establishing an appropriate diagnosis and determining next steps. Antón (2009) illustrated the application of DA in assessment of writing and speaking abilities of advanced Spanish language learners. In the writing assessment, learners could make revisions under three conditions. First, following Aljaafreh and Lantolf (1994), learners revised their compositions individually albeit in the presence of the evaluator. Further revisions were allowed by consulting cultural artifacts, such as a dictionary and a grammar reference. Finally, learners were able to ask the evaluator specific questions about their compositions. During the oral assessment, the evaluator took an agentive role, intervening with assisting questions and providing learners with opportunities to make corrections and to add details to a story narration. Qualitative analysis of the interaction between the examiner and the learner during the narrative task revealed important differences among learners and the importance of interaction in establishing an appropriate diagnosis of the learners' linguistic ability.

Technology has also been helpful in facilitating diagnosis through interventionist DA. A series of studies illustrate the diagnostic potential of computerized DA. These will be discussed in the next section.

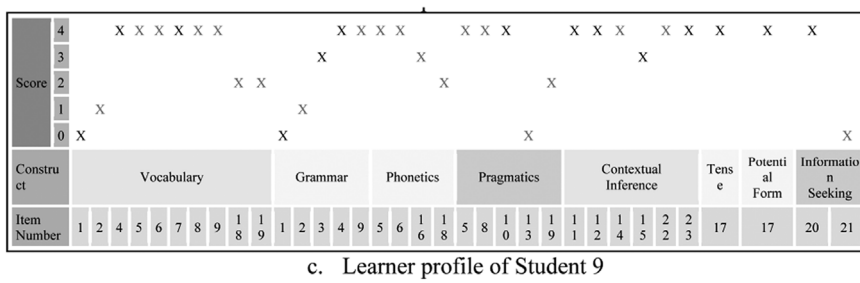
Computerized dynamic assessment (CDA)

The field of educational psychology provides models of interventionist approaches to computerized DA that may be adapted to second language contexts. A series of studies (Lantolf and Poehner, 2013; Poehner *et al*., 2015; Zhang and Lu, 2019) report on the development and implementation of a computer-based test of L2 listening and reading comprehension across several languages that provides independent and mediated performance results. Following an interventionist approach to DA, test items included four prompts with different degrees of explicitness (i.e., from the most implicit to the most explicit). This CDA approach generated actual and potential scores for learners, as well as individual profiles reporting independent or assisted performance on specific items or clusters of items in the test (see Table 10.2).

Table 10.2 Functions of the scores and learner profile automatically generated by the C-DA tests

Score/Profile	*Function*
Actual score	A measure of independent, unassisted performance. For each test item, a learner receives either maximum points (4) or no point, depending on whether the first response is correct.
Mediated score	A measure of mediated, assisted performance. For each test item, a learner receives 0–4 points, depending on whether and how many prompts are used, with 1 point deducted for each prompt used.
Learning potential score	A measure of learning potential. Calculated as (2 * Mediated Score – Actual Score)/Maximum Score (Kozulin and Garb, 2002: 121).
Transfer score	A measure of knowledge transfer within the test. Reports a learner's score on items at the end of the test designed to retest previously tested constructs in a more challenging way.
Learner profile	Details the level of support needed for each item and for clusters of items focusing on specific constructs (e.g., all items on grammatical knowledge).

Source: Zhang and Lu, 2019: 96



c. Learner profile of Student 9

Figure 10.1 Sample CDA learner profile

Source: Zhang and Lu, 2019: 104

Zhang and Lu (2019) used this model of CDA to measure initial differences among L2 learners of Chinese, which formed the basis for curricular design (See Figure 10.1 for a sample CDA learner profile). They showed through student feedback that students used CDA as a learning tool, retaking the test several times as needed. Overall, feedback from students was very positive for the use of DA (they found it less stressful and viewed it as an opportunity to concentrate and revise information). The variety of interventions in their study (CDA, teacher interaction, individualized assignments, etc.) and the use of CDA results to guide instructional design constitute a powerful pedagogical model.

More recently, other technologies have been used as tools to deliver DA. Rassaei (2020), for instance, discusses dialogic mediation in reading comprehension and vocabulary learning via mobile technology in a pedagogical intervention inspired by DA principles. Andujar (2020), following interventionist DA procedures, used a mobile instant messaging application (WhatsApp) to mediate EFL learners' grammatical and lexical development over a five-month period. Learners in Andujar's study engaged in daily WhatsApp conversations in which the researcher would provide participants with mediation through a pre-designed inventory of prompts ranging from most implicit to most explicit. Results indicated that

the dialogic interaction and mediation allowed learners to reflect on their linguistic performance and perform with increasingly less explicit feedback. Despite important advances and experimentation in this area, the application of technology to DA in L2 contexts has not been developed to its full potential yet.

Program assessment

Erben *et al*. (2008) reported on the use of DA among faculty and pre-service teachers in the ESOL program at a university college of education. The progressive introduction of DA techniques, starting with one quiz in one class at the same time that the students were being instructed about the theoretical underpinnings of DA and its pedagogical advantages, gained increasing support to the point of being deployed voluntarily by the majority of the faculty for high-stakes exams (midterm and final). The exam protocol introduced DA by having students complete the exam one page at a time, providing individual, pair, and group responses. The pair and group responses involved negotiation among learners. Students received a combined score for their individual and negotiated responses with their pair and group. Open-ended interviews with students and faculty provided insights into participants' perceptions of the use of DA techniques. Students reported more confidence in content knowledge than students who had not participated in DA and viewed DA as a learning opportunity. For example, one student said, "Doing it DA style, well, I guess I feel I know more because the stuff we learned during the DA exam stuck" (p. 102). Instructors had a positive view of DA. However, there were remaining concerns regarding the ability to assess state standards at the individual level, time restrictions regarding individual assessment, and the mediating sessions themselves. The authors recommended more faculty training in order to diffuse educational change. To this end, they highlight the importance of social and institutional support for faculty to decide whether to adopt an innovation such as DA.

Teacher education

Program-wide implementation of the kind reported by Erben *et al*. (2008) is dependent on teacher experiences with DA and training programs that prepare teachers to adopt this type of assessment. Several practical guides and resources are available to aid teachers in L2 DA training (Lantolf and Poehner, 2011; Poehner, 2019). Some studies report on successful programs to familiarize and train teachers in DA use. For example, Davin *et al*. (2016) describe changes in teachers' discursive practices as they attempt to implement DA as a result of one professional development workshop. Analysis of classroom discourse and teachers' reflections revealed a progressive move away from corrective feedback practices toward mediation focused on development, although teachers were able to adopt DA to varying degrees. They struggled with creating graduated prompts contingent on the learners' needs, and they seemed to need more mediation to appropriate the new concepts. The authors suggest that an interventionist approach to DA, with pre-scripted prompts, may be more appropriate for teachers new to DA.

Inclusion of DA in the mainstream curriculum in teacher education programs is still rare. García (2019) combined concept-based instruction (CBI) with interactionist DA in a case study that took place in a teacher training program. He explored conceptual development of verbal aspect in Spanish as evidenced by narrations, verbalizations, and reflections collected during dyadic DA sessions with an in-service teacher of Spanish who was also a native

speaker of the language. The concepts were visualized with representations used as a guide to create meaning during communicative activities. Subsequent mediated "verbalizations" provided evidence of the learner's shift in her understanding of the grammatical concept. By having teachers experience the process of conscious conceptual manipulation and the process of assessing conceptual development dynamically, the study could transform teaching and assessment practices in the teachers' own classrooms.

Efforts in the area of teacher development in DA practices are still limited. Beyond occasional training opportunities, wide dissemination of DA practices would require better integration in the curriculum of teacher preparation programs.

Recommendations for practice

DA promotes an individualized approach to assessment that is both its major strength and its major challenge. Its strength lies in its ability to provide rich and useful information to teachers and educators for the development of pedagogical interventions that will address the needs of particular learners and for learners to engage in a model of assessment that allows them to exercise their agency while co-constructing knowledge and uncovering their potential for linguistic development. The challenges are several. Internally, DA approaches are not homogeneous. There has been much experimentation in the implementation of DA to new contexts in L2 teaching and learning. While practical applications to L2 settings have helped shape and expand DA, it is important to ensure fidelity to basic DA principles and alignment with the overall SCT orientation toward language learning.

Teacher development is a key aspect of effective implementation of DA. Interactionist approaches require a high level of training, skills, and expertise by the assessor in order to judge in the moment the level of assistance required during the DA. Furthermore, if the goal of DA is to promote development (rather than learning), assessment cannot be disassociated from a sociocultural understanding of development and how development is achieved. Thus, DA practice should be coherently embedded in pedagogical models that support the aims of this type of assessment, such as concept-based instruction, mediated language experiences, etc. DA can be most effective when it deploys cognitive as well as metacognitive mediational tools as part of a coherent orientation to pedagogical practice. Using an analogy, IPA (integrated performance assessment) only make sense in communicative-friendly instructional approaches; otherwise, content as well as construct validity is lost.

Teachers and administrators who wish to adopt DA will have to prepare for the practical challenges outlined in this chapter. As illustrated by some of the studies discussed here, the creative use of technology can help overcome time and human constraints in imparting DA to large groups of learners. Also, the use of group and peer assessment may provide a way to deal with the time challenge. Whether DA is used as a complement to other methods of assessment or as the only means of assessment, appropriate teacher training and institutional acceptance and support for DA are key in its successful implementation.

Future directions

Research on DA has been fruitful in recent years. DA, along with other learning-oriented forms of assessments that empower teachers and learners in the learning process, has received increasing attention from the field of assessment in general and from language

learning and teaching in particular. There is increasing interaction and potential collaboration across assessment paradigms, particularly with diagnostic assessment and formative assessment, with the aim of emphasizing similarities and potentiating synergies that may aid learners (Lee, 2015; Leung, 2007).

Interesting work has been conducted using DA with groups of learners (classroom centered or peer assessment), delivering DA via technology, understanding the process of mediation and transcendence in relation to teaching/assessment during DA sessions, and understanding the role of the learners in DA. Although many innovative single-case studies have been reported in recent years, more small- and large-scale studies are needed to explore the potential of DA for diagnosis and development in language learning. The field would benefit from studies reporting systematic and wide implementation of DA within programs of study, including systematic efforts to include DA in the overall plan of instruction and assessment for language programs.

We see strong potential in DA to help language programs address the needs of learners with special needs, including diverse populations. Although DA, from its origins, has proven effective with learners experiencing learning difficulties, L2 DA has not yet explored in depth its potential contribution to pedagogical programs in order to enhance learning for all. Fertile ground for more research continues to be the expanded use of computer and peer-mediated assessment, analysis of mediator/learner interaction and learner reciprocity, and, particularly, the development of standardized procedures in DA implementation and teacher training so that learners' development may be systematically documented.

The resolution of practical obstacles to large-scale implementation, the use of DA as a means of assessment on its own right, and the acceptance of DA as a valid means of assessment with fundamental philosophical differences with mainstream testing remain as workable challenges.

Further reading

Lantolf, J. P. and Poehner, M. E. (2011). *Dynamic Assessment in the Foreign Language Classroom: A Teacher's Guide*, 2nd edn. University Park, PA: Calper.

This text is an introduction to the implementation of DA in language contexts. It is aimed at language teachers who wish to bring DA to their own classrooms and provides a brief theoretical background, multiples examples from actual practice, and practical activities for teacher and student development alike.

Lidz, C. (1991). *Practitioner's Guide to Dynamic Assessment*. New York, NY: Guilford Press.

This is a good introduction to practical aspects of interactionist DA.

The book summarizes different models of DA in clinical and educational settings, but, most importantly, it provides a step-by-step guide of procedures and tools for conducting DA assessment. The two manuals included in the book provide a good starting point for anyone wishing to start implementing DA.

Poehner, M. (2008). *Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting L2 Development*. Boston, MA: Springer Science.

This is the first comprehensive review of theoretical and practical aspects of DA in second language learning. The first part of this book discusses the origins of DA in Vygotsky's sociocultural theory of mind and subsequent development of different approaches to DA. The second part of the book presents current research in DA applications to L2 settings and offers a model for implementing DA in second language classrooms.

References

Ableeva, R. (2008). The effects of dynamic assessment on L2 listening comprehension. In J. Lantolf and M. Poehner (eds.), *Sociocultural Theory and the Teaching of Second Languages*. London, UK: Equinox Pub, 57–86. http://doi.org/10.1558/equinox.29293.

———. (2018). Understanding Learner L2 Development Through Reciprocity. In J. Lantolf, M. Poehner and M. Swain (eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development*. London: Routledge, 266–81. https://doi.org/10.4324/9781315624747.

Aljaafreh, A. and Lantolf, J. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *Modern Language Journal* 78: 465–83. https://doi.org/10.1111/j.1540-4781.1994.tb02064.x.

Andujar, A. (2020). Mobile-mediated dynamic assessment: A new perspective for second language development. *ReCALL*: 1–17. https://doi.org/10.1017/S0958344019000247.

Antón, M. (2009). Dynamic assessment of advanced foreign language learners. *Foreign Language Annals* 42: 576–98. https://doi.org/10.1111/j.1944-9720.2009.01030.x.

———. (2018). Dynamic diagnosis of second language abilities. In J. Lantolf, M. Poehner and M. Swain (eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development*. London, UK: Routledge, 310–23. https://doi.org/10.4324/9781315624747.

Black, P. J. and Williams, D. (1998). Assessment and classroom learning. *Assessment in Education* 5: 7–74. https://doi.org/10.1080/0969595980050102.

Budoff, M. (1987). The validity of learning potential assessment. In C. Lidz (ed.), *Dynamic Assessment: An Interactive Approach to Evaluating Learning Potential*. New York, NY: Guilford Press, 52–81.

Campione, J. and Brown, A. (1987). Linking dynamic assessment with school achievement. In C. Lidz (ed.), *Dynamic Assessment*. New York, NY: Guilford Press, 82–115.

Davin, K. (2016). Classroom dynamic assessment: A critical examination of constructs and practices. *Modern Language Journal* 100: 813–29. https://doi.org/10.1111/modl.12352.

———. (2018). Mediator and learner engagement in co-regulated inter-psychological activity. In J. Lantolf, M. Poehner and M. Swain (eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development*. London, UK: Routledge, 282–94. https://doi.org/10.4324/9781315624747.

Davin, K. J. and Donato, R. (2013). Student collaboration and teacher-directed classroom dynamic assessment: A complementary pairing. *Foreign Language Annals* 46: 5–22. https://doi.org/10.1111/flan.12012.

Davin, K. J. and Gomez-Pereira, D. (2019). Evaluating instruction through classroom dynamic assessment: A sandwich approach. *Language and Sociocultural Theory* 6: 6–31. https://doi.org/10.1558/lst.38914.

Davin, K. J., Herazo, J. D. and Sagre, A. (2016). Learning to mediate: Teacher appropriation of dynamic assessment. *Language Teaching Research* 21: 1–20. https://doi.org/10.1177%2F1362168816654309.

Davis, J. M. (2018). Preface. In J. M. Davis, J. M. Norris, M. E. Malone, T. H. McKay and Y. Son (eds.), *Useful Assessment and Evaluation in Language Education*. Washington, DC: Georgetown University Press, vii–ix. https://doi.org/10.2307/j.ctvvngrq.

Davison, C. and Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly* 43: 393–415. https://doi.org/10.1002/j.1545-7249.2009.tb00242.x.

Erben, T., Ban, R. and Summers, R. (2008). Changing examination structures within a college of education: The application of dynamic assessment in pre-service ESOL endorsement courses in Florida. In J. Lantolf and M. Poehner (eds.), *Sociocultural Theory and the Teaching of Second Language*. London, UK: Equinox Pub, 87–114. https://doi.org/10.1558/equinox.29294.

Feuerstein, R., Feuerstein, R. S., Falik, L. H. and Rand, Y. (2002). *The Dynamic Assessment of Cognitive Modifiability: The Learning Propensity Assessment Device: Theory, Instruments and Techniques*. Jerusalem, Israel: International Center for the Enhancement of Learning Potential.

Feuerstein, R., Rand, Y. and Hoffman, M. (1979). *The Dynamic Assessment of Retarded Performers: The Learning Potential Assessment Device, Theory, Instruments, and Techniques*. Baltimore, MD: University Park Press.

Fulcher, G. and Davidson, F. (eds.) (2012). *The Routledge Handbook of Language Testing*. London, UK: Routledge. https://doi.org/10.4324/9780203181287.

García, P. N. (2012). *Verbalizing in the Second Language Classroom: The Development of the Grammatical Concept of Aspect*. Unpublished PhD thesis. University of Massachusetts, Amherst, MA, USA. https://doi.org/10.7275/acvx-px32.

———. (2014). Verbalizing in the second language classroom: Exploring the role of agency in the internalization of grammatical categories. In P. Deters, X. Gao, E. R. Miller and G. Vitanova (eds.), *Theorizing and Analyzing Agency in Second Language Learning: Interdisciplinary Approaches*. Bristol, UK: Multilingual Matters, 213–31. https://doi.org/10.21832/9781783092901-014.

———. (2019). Dynamic assessment: Promoting in-service teachers' conceptual development and pedagogical beliefs in the L2 classroom. *Language and Sociocultural Theory* 6: 32–62. https://doi.org/10.1558/lst.38915.

Grigorenko, E. L. (2009). Dynamic assessment and response to intervention. *Journal of Learning Disabilities* 42: 111–32.

Grigorenko, E. L. and Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin* 124: 75–111. https://doi.org/10.1177/0022219408326207.

Güthke, J. (1982). The learning test concept – an alternative to the traditional static intelligence test. *The German Journal of Psychology* 6: 306–24.

Haywood, H. and Lidz, C. (2007). *Dynamic Assessment in Practice: Clinical and Educational Applications*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511607516.

Hornberger, N. (ed.). (2007). *The Encyclopedia of Language and Education, vol. 7: Language Testing and Assessment* (E. Shohamy, ed.). Cambridge, UK: Cambridge University Press.

Hill, K. and Sabet, M. (2009). Dynamic speaking assessments. *TESOL Quarterly* 43: 537–45. https://doi.org/10.1002/j.1545-7249.2009.tb00251.x.

Infante, P. and Poehner, M. (2019). Realizing the ZPD in second language education: The complementary contributions of dynamic assessment and mediated development. *Language and Sociocultural Theory* 6: 63–91. https://doi.org/10.1558/lst.38916.

Kozulin, A. and Garb, E. (2002). Dynamic assessment of EFL text comprehension of at-risk students. *School Psychology International* 23: 112–27. https://doi.org/10.1177%2F0143034302023001733.

Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment: Bringing the past into the future. *Journal of Applied Linguistics* 1: 49–74. http://doi.org/10.1558/japl.v1i1.647.

———. (2011). *Dynamic Assessment in the Foreign Language Classroom: A Teacher's Guide*, 2nd edn. University Park, PA: Calper.

———. (2013). The unfairness of equal treatment: Objectivity in L2 testing and dynamic assessment. *Educational Research and Evaluation* 19: 141–57. https://doi.org/10.1080/13803611.2013.767616.

Lantolf, J. P., Poehner, M. E. and Swain, M. (eds.). (2018). *Handbook of Sociocultural Theory and Second Language Learning*. London, UK: Routledge. https://doi.org/10.4324/9781315624747.

Lantolf, J. and Thorne, S. (2006). *Sociocultural Theory and the Genesis of Second Language Development*. Oxford, UK: Oxford University Press.

Lee, Y. (2015). Future of diagnostic language assessment. *Language Testing* 32: 295–98. https://doi.org/10.1177/0265532214565385.

Leung, C. (2007). Dynamic assessment: Assessment for and as teaching? *Language Assessment Quarterly* 4: 257–78. https://doi.org/10.1080/15434300701481127.

Levi, T. and Poehner, M. (2018). Employing dynamic assessment to enhance agency among L2 learners. In J. Lantolf, M. Poehner and M. Swain (eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development*. London: Routledge, 295–309. https://doi.org/10.4324/9781315624747.

Lidz, C. (1987). *Dynamic Assessment*. New York, NY: Guilford Press.

———. (1991). *Practitioner's Guide to Dynamic Assessment*. New York, NY: Guilford Press.

Lund, A. (2008). Assessment made visible: Individual and collective practices. *Mind, Culture, and Activity* 15: 32–51. https://doi.org/10.1080/10749030701798623.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Oxford, UK: Blackwell Publishing.

Minick, N. (1987). Implications of Vygotsky's theories for dynamic assessment. In C. Lidz (ed.), *Dynamic Assessment*. New York, NY: Guilford Press, 116–40.

Negueruela-Azarola, E., García, P. N. and Buescher, K. (2015). From inter-action to intra-action: The internalization of talk, gesture, and concepts in the second language classroom. In N. Markee (ed.), *The Handbook of Classroom Interaction*. Malden, MA: Wiley-Blackwell, 233–49. https://doi.org/10.1002/9781118531242.ch14.

Peña, E. and Greene, K. (2018). Dynamic assessment of children learning a second language. In J. Lantolf, M. Poehner and M. Swain (eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development*. London, UK: Routledge, 324–40. https://doi.org/10.4324/9781315624747.

Poehner, M. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal* 91: 323–40. https://doi.org/10.1111/j.1540-4781.2007.00583.x.

———. (2008a). Both sides of the conversation: The interplay between mediation and learner reciprocity in dynamic assessment. In J. Lantolf and M. Poehner (eds.), *Sociocultural Theory and the Teaching of Second Languages*. London, UK: Equinox Pub, 33–56. http://doi.org/10.1558/equinox.29292.

———. (2008b). *Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting l2 Development*. Boston, MA: Springer Science.

———. (2009). Group dynamic assessment: Mediation for the L2 classroom. *TESOL Quarterly* 43: 471–91. https://doi.org/10.1002/j.1545-7249.2009.tb00245.x.

———. (2018). Probing and provoking L2 development: The object of mediation in dynamic assessment and mediated development. In J. Lantolf, M. Poehner and M. Swain (eds.), *The Routledge Handbook of Sociocultural Theory and Second Language Development*. London, UK: Routledge, 249–65. https://doi.org/10.4324/9781315624747.

———. (2019). *A Casebook of Dynamic Assessment in Foreign Language Education*. University Park, PA: Calper.

Poehner, M. and Infante, P. (2016). Mediated development: A Vygotskian approach to transforming learner L2 abilities. *TESOL Quarterly* 51: 332–57. https://doi.org/10.1002/tesq.308.

Poehner, M. and Lantolf, J. (2005). Dynamic assessment in the language classroom. *Language Teaching Research* 9: 1–33. https://doi.org/10.1191%2F1362168805lr166oa.

———. (2013). Bringing the ZPD into the equation: Capturing L2 development during computerized dynamic assessment. *Language Teaching Research* 17: 323–42. https://doi.org/10.1177%2F1362168813482935.

Poehner, M. and van Compernolle, R. A. (2011). Frames of interaction in dynamic assessment: Developmental diagnoses of second language learning. *Assessment in Education: Principles, Policy and Practice* 18: 183–98. https://doi.org/10.1080/0969594X.2011.567116.

Poehner, M., Zhang, J. and Lu, X. (2015). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing* 32: 337–57. https://doi.org/10.1177%2F0265532214560390.

Rahimi, M., Kushki, A. and Nassaji, H. (2015). Diagnostic and developmental potentials of dynamic assessment for L2 writing. *Language and Sociocultural Theory* 2: 185–208. https://doi.org/10.1558/lst.v2i2.25956.

Rassaei, E. (2020). Effects of mobile-mediated dynamic and nondynamic glosses on L2 vocabulary learning: A sociocultural Perspective. *Modern Language Journal* 104: 284–303 https://doi.org/10.1111/modl.12629.

Schneider, E. and Ganschow, L. (2000). Dynamic assessment and instructional strategies for learners who struggle to learn a foreign language. *Dyslexia* 6: 72–82. https://doi.org/10.1002/(SICI)1099-0909(200001/03)6:1%3C72::AID-DYS162%3E3.0.CO;2-B.

Sternberg, R. J. and Grigorenko, E. L. (2002). *Dynamic Testing: The Nature and Measurement of Learning Potential*. Cambridge, UK: Cambridge University Press.

Turner, C. E. and Purpura, J. E. (2015). Learning-oriented assessment in the classroom. In *Handbook of Second Language Assessment*. Berlin, Germany and Boston, MA: DeGruyter Mouton, 255–74.

Tzuriel, D. (1997). The relation between parent-child MLE interactions and children's cognitive modifiability. In A. Kozulin (ed.), *The Ontogeny of Cognitive Modifiability*. Jerusalem, Israel: International Center for the Enhancement of Cognitive Modifiability, 157–80.

van Compernolle, R. and Kinginger, C. (2013). Promoting metapragmatic development through assessment in the zone of proximal development. *Language Teaching Research* 17: 282–302. https://doi.org/10.1177%2F1362168813482917.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes* (M. Cole, V. John-Steiner, S. Scribner and E. Souberman, eds.). Cambridge, MA: Harvard University Press.

Zhang, J. and Lu, X. (2019). Measuring and supporting second language development using computerized dynamic assessment. *Language and Sociocultural Theory* 6: 92–115. https://doi.org/10.1558/lst.31710.

11

Diagnostic assessment in language classrooms

Eunice Eunhee Jang and Jeanne Sinclair

Introduction

Testing has become the primary means to drive educational reforms and measure the resulting outcomes for accountability purposes (Jang and Ryan, 2003). However, in classrooms, teachers report little interest in the prototypical information provided by large-scale standardized tests but seek instead detailed information about individual students' strengths and weaknesses in specific skills taught (Brindley, 1998; Huff and Goodman, 2007). The type of information that teachers seek is *diagnostic*.

Eighteen years ago, Spolsky (1992) envisioned that "diagnostic testing is an art, albeit one informed by the science of testing – an art that has learned from science ways to ensure fairness and reliability in subjective measures that will lead to responsible language testing" (p. 29). Since then, much progress has been made. Several books and journal special issues have been written expressly on diagnostic assessment in language and educational assessment (Alderson, 2005; Frederiksen *et al*., 1990; Nichols *et al*., 1995; Leighton and Gierl, 2007; Rupp *et al*., 2010). Furthermore, the diagnostic potential of various performance assessments is evident in scaled proficiency descriptors, such as the Common European Framework of Reference for Languages (CEFR), the Canadian Language Benchmarks (CLB), the American Council of the Teaching of Foreign Languages (ACTFL), the Steps to English Proficiency (STEP) used by mainstream classroom teachers in Ontario public schools, K–12 English language proficiency standards adopted throughout the United States under the Every Student Succeeds Act (ESSA), and the World-Class Instructional Design and Assessment (WIDA) English language proficiency standards. The performance descriptors included in those scales are often translated into learners' self-assessment, as shown in the DIALANG system (Alderson, 2007).

Another significant advancement in diagnostic assessment is computer-assisted assessment systems that provide a dynamic interface for integrating language assessment into teaching and learning (Chapelle *et al*., 2008). These interfaces, afforded by technological advances, make possible an interactive diagnostic feedback loop that can support differentiated instruction and learning.

DOI: 10.4324/9781003220756-15

Despite the appeal of diagnostic assessment, misconceptions remain. One idea is that *diagnostic* focuses on student knowledge *prior to* instruction, whereas *formative* is an ongoing assessment of students' learning progress. This distinction appears to undermine the potential of the former. Another misconception is that diagnostic assessment is interchangeable with formative assessment and distinct from summative assessment (Leighton, 2009). However, not all formative assessments are diagnostic, as discussed next.

Characteristics of diagnostic assessment

Purposes of diagnostic assessment

The diagnostic language assessment goals are to *make diagnostic inferences* about learners' strengths and weaknesses in skills of interest and to *utilize* them for positive changes in learning. This pedagogical desire resonates well with formative assessment, in that both can be used to plan differentiated instruction tailored to students' learning needs (Black and Wiliam, 1998; Nichols *et al*., 2009). Nichols *et al*. (2009) distinguish formative assessment from other assessment purposes by highlighting test score use as a main validity claim over valid score interpretation. That is, the effectiveness of the assessment is not determined by its accuracy as a snapshot of a student's ability but by the extent to which it can be used to improve student learning. This pedagogical effect is an important validity criterion for diagnostic and formative assessments alike because both are oriented toward improving teaching and learning. Still, a fundamental difference between the two is that while diagnostic language assessment requires a profound understanding of developmental pathways to language proficiency and evidence for diagnostic inferences and use, formative assessment alone lacks such attention to theoretical construct representations (Leighton, 2009; Nichols *et al*., 2009).

Diagnostic language assessment also differs from diagnosis performed in clinical and psycho-educational fields. The former presupposes no prescribed algorithm for estimating language proficiency so that it can be easily measurable and classifiable; the latter involves prescribed classifications based on decisive causal determinations. It is difficult to establish causal inferences about what causes weaknesses in language proficiency, as language development is a latent construct that is only visible when evidence of proficiency is elicited through measures of reading, writing, speaking, and listening. Another defining feature for differentiating between the two is *use*. Action is not necessarily a primary concern for a diagnostician in the latter. Alderson *et al*. (2015) support this distinction that unlike medical diagnosis, whose primary goal is to identify a problem, educational diagnosis is intended to take action for problem solving.

Diagnostic language assessment takes a mastery criterion-referenced approach that determines learners' mastery of skills toward user-specified learning expectations. This is differentiated from norm-referenced high-stakes proficiency tests (Kunnan and Jang, 2009), which are primarily intended to rank, select, and discriminate among students against population norms established by psychometric principles, such as a Gaussian distribution. These tests often attempt to multitask by serving diagnostic and placement purposes in addition to their primary selection purpose, despite their inability to measure developmental growth (Fox, 2009). For such tests, the pedagogical effect is not a primary validity concern.

Sources of diagnostic inferences

Identifying sources of discrepancies from expected performances or characterizing patterns of errors can be diagnostically useful (Bejar, 1984). One common source of diagnosis is a *vertical discrepancy* in performance between a current and an expected knowledge state for a given skill. Diagnosis based on a vertical discrepancy is grounded in developmental theories of language learning. The primary criteria for determining a student's proficiency state are specific learning expectations. Thus, diagnostic feedback can direct students to skill-specific learning activities. Such vertical discrepancy assessment models can be found in mastery-oriented scales of language proficiency, such as the CEFR common reference global scales (Little, 2006).

Content experts, including teachers, are often involved in a standard-setting procedure to determine criterion levels based on mastery expectations for target domains (Cizek, 2001). The use of standards makes learning expectations explicit (Cumming, 2009). The resulting performance categories are expressed on the mastery-oriented proficiency scales. For example, the CEFR includes six reference levels such as Basic (Breakthrough, Waystage), Independent (Threshold, Vantage) and Proficient (Effective Operational Proficiency, Mastery). In fact, it is the detailed performance-level descriptors (PLDs) that are diagnostic, rather than the performance category labels. PLDs describe students' knowledge and skills within each performance level (Cizek *et al*., 2004) and facilitate teachers' *diagnostic judgments* about students' mastery status. In classrooms, a mastery criterion level can be set at any point of instruction where teachers need diagnostic information about students' mastery status, instead of relying on end-of-term performance (Hudson and Lynch, 1984).

However, setting mastery criteria or standards is neither straightforward nor fully knowable (Cizek *et al*., 2004). Determining the level of specificity for diagnosis is equally challenging – which grain size of feedback is most useful for teachers and students? Furthermore, the development of language standards and proficiency descriptors relies on content experts' (e.g., teachers) intuitive judgments and ad hoc statistical scaling (Alderson, 2005, 2007; Cumming, 2009; North and Schneider, 1998). Cizek *et al*. (2004) respond to this criticism by characterizing standard setting as a procedure involving "artistic, political, and cultural" elements (p. 33) and further note:

> It is now a widely accepted tenet of measurement theory that standard-setting panels' work is not to search for a knowable boundary between categories that exist. Instead, standard-setting procedures enable participants to bring to bear their judgments in such a way as to translate policy decisions (often, as operationalized in performance level descriptors) into locations that create the effective performance categories. This translation and creation are seldom, if ever, purely statistical, impartial, apolitical, or ideologically neutral activities.
>
> *(Cizek et al., 2004: 33)*

A lack of theoretical foundations underlying existing vertical scales seems a significant obstacle in seeking validity evidence for diagnostic inferences. Not all competing theoretical frameworks are equally relevant to diagnostic assessment, and there is no single best-fitting theoretical framework for all learning contexts. It is neither desirable nor practical to assume that a single diagnostic assessment can assess the entire spectrum of language

development. The viability of any theoretical framework for diagnosis rests on evaluative judgments about its relevance to a specific learning context and assessment purposes.

Another source of discrepancy can be *horizontal*. This occurs when achievement in one content or skill area is in discord with achievement in other content or skills. Proficiency tests based on multimodal construct definitions (i.e., reading, writing, oral) without hierarchical or developmental linkage are often used to determine horizontal discrepancies for diagnostic purposes. Mousavi (2002) notes that proficiency tests, such as the Test of English as a Foreign Language (TOEFL), can be used for diagnosing students' strengths and weaknesses based on subtest (reading, writing, speaking, and listening) scores, concluding that any language test has diagnostic potential. Subtest scores can be used to diagnose students' abilities; however, the level of specificity is often too broad to direct students' learning. Therefore, to conclude that any language test has some degree of diagnostic potential is misleading without considering the appropriate level of specificity in diagnosis as determined by those most affected by the effects of diagnosis – namely, students (Fulcher and Davidson, 2009).

The expected knowledge state can be emergent or adjacent to the current state of knowledge, instead of being oriented toward distant mastery learning goals. Indeed, this is the main characteristic of dynamic assessment (Lantolf and Poehner, 2004; see also Antón and García, this volume). In this view, the source of discrepancy is the gap between students' actual and potential development, and the goal of assessment is to gather diagnostic information through careful mediation. In this scenario, the assessor assesses the current state of development (to determine the present state of a learner's skills) and provides mediation to understand the learner's adeptness at taking up instruction during the mediation process – which is indispensable for a learner's future development. Lantolf and Poehner (2004) point out that "assessing without mediation is problematic because it leaves out part of the picture – the future – and it is difficult to imagine an assessment context that is *not* interested in the future" (p. 251).

Yet another source of discrepancy can arise from students' self-assessment of their own abilities. Here, the students' *performed* competencies are compared against their *perceived* competencies. Students are change agents for their own learning; therefore, students' self-assessment is pivotal for ensuring the maximal use of diagnostic information. The DIALANG assessment system (Alderson and Huhta, 2005) links the assessment results to a test taker's self-assessment results and provides diagnostic feedback focusing on the discrepancy between them. Jang (2005) notes that students tend to pay attention to the discrepancy when evaluating the validity of diagnostic feedback.

Standardized diagnostic tests

A close review of a few K–12 diagnostic test tools helps us understand the status quo of contemporary diagnostic assessments available for classroom teachers (see also Fox, Abdulhamid, and Turner, this volume). For example, two diagnostic tests that are being widely implemented in elementary schools across Ontario, Canada, are the Developmental Reading Assessment (DRA; Beaver, 2001) for kindergarten to grade 3 (DRA2 for grades 4 to 8) and the Comprehension, Attitude, Strategies, Interest (CASI; Gibson, 2007) for grades 4 through 8. Both are commercial, standardized reading assessments. Both claim to serve multiple purposes, including diagnosing students' strengths and learning needs; planning instructional interventions regarding the provincial reading expectations; tracking student growth on core skills and strategies that effective readers would use; determining students'

reading achievements relative to their grade-specific norm; and supporting decision-making procedures taking place among divisions, schools, and parents. These purposes are comprehensive and ambitious, a point we will revisit later in this section.

The DRA is administered individually through a one-on-one conference. During the individual conference, the teacher administers a running record (Clay, 2000) by introducing a text and recording the numbers of errors and self-corrections while the child is reading the text. The assessor asks a series of comprehension questions. Depending on the text level, the student may independently finish the reading and answer the questions. The DRA results are used to determine the child's independent reading level and profile students in terms of their reading levels (below, at, or above the grade-level proficiency).

In contrast to the DRA, CASI is administered to a group of students. It includes several subtests, including reading interviews, a reading attitude survey, a reading interests inventory, a student self-assessment, and a reading comprehension test. The reading comprehension test includes ten reading passages (five fiction and five non-fiction each) per grade, each of which includes eight multiple-choice questions. The results are reported in four levels of performance related to the tested elements. A positive aspect of the CASI reading assessment is its inclusion of measures that assess students' attitudes and values toward reading and their interest in reading.

Though those assessments are widely used in classrooms for diagnostic and tracking purposes in Ontario, surprisingly, there is little publicly available research evidence supporting their validity. Significant correlations, $r = .62$, $n = 57$, $p < .01$, have been demonstrated between a grade 6 provincial standardized literacy exam and DRA scores in grades 7 and 8 (Stewart, 2016), but there is a dearth of such information for younger students, who are the primary focus of the DRA. The DRA has been shown to be susceptible to inter-rater reliability issues (Burgin and Hughes, 2009). A review of the DRA (McCarthy and Christ, 2010) found evidence for face validity and construct validity with regard to identifying reading difficulties; however, these reviewers found a lack of objectivity in the scoring and classification scheme (i.e., intervention, instructional, independent, and advanced), as well as a lack of sufficient information about how this classification could be used to guide instruction. As for the CASI, very limited information is available with regard to its quality, likely due in part to its use only in Canadian classrooms.

Based on interviews with teachers in grades 3 through 5 who use the DRA and CASI, Jang (2010) reports that teachers' use of DRA test results is limited to individual students' overall performance levels or a scaled or percentile score. The following interview excerpt with a grade 3 teacher who uses the DRA in her classroom illustrates its limited use.

> Many students in my class speak other languages than English at home. They are proficient in English when they speak, but jeez, they just can't write. Not only that, they are not motivated; they simply don't care. For example, Jason's reading level is only six on DRA. It took two months to figure out what's wrong. . . . It's the fifth year in my teaching, and I'm getting better with this, but it's so hard to know exactly what they are struggling with and support them with individualized instruction.

These standardized commercially available diagnostic tests are based on several subtests, each of which measures interrelated skills. Scores are reported using proportion-correct scores, percentiles, or scores scaled by item response theory. They often fail to serve classroom teachers because they are not properly aligned with what and how they teach in the classroom. We need to critically examine such assessment products by asking two questions:

What is the level of specificity? And are assessment tasks cognitively intelligent enough to elicit the targeted skills?

How do we determine the level of specificity?

An important step in designing a diagnostic assessment is defining and specifying the primary diagnostic criterion specific to a learning context. Diagnostic assessment tends to assess a target range of skills instead of covering a wide range of skills (Alderson, 2005). However, a too narrowly specified skill specification may pose a problem resulting from "a gross simplification of the reality of cognition" (Mislevy, 1995: 55); that is, an assessment with too narrow a skill specification may lack relevance to teaching, learning, and real-world application. How do we then strike a balance? Mislevy (1995) notes that the level of detail at which students' competencies are differentiated needs to be determined based on pedagogical considerations – that is, can the specified detail direct students' learning? There is no absolutely true level of specificity for diagnostic criteria (Davidson and Lynch, 2002). We need to ask whether the level of specificity serves the diagnostic purpose and instructional goals in a given learning context in evaluating existing diagnostic tests used in the classroom.

Are assessment tasks sufficiently cognitively intelligent to elicit target skills?

Diagnostic assessment tasks should investigate a unique part of language learning to gather useful information about students' language profiles. Static taxonomies of content specifications are not sufficient for creating diagnostic tasks that reflect "the dynamic and sequential nature of diagnostic assessment" (Bejar, 1984: 175). Diagnostic tasks require a multi-layered, micro-analysis of learning behaviour, level of cognitive difficulty, and learner background characteristics.

Various task types should be considered. Selected response task types can be cognitively designed to measure partial knowledge and misconceptions. For example, one can gather diagnostic information by analysing students' responses to multiple distracters. Gu and Jang (2009) examined how distracters included in grade 6 curriculum-based reading comprehension assessment items function across different ability levels and whether they can provide additional diagnostic information. They concluded that approximately half (21 out of 40) exhibited *optimal* or *partially rich* distracter functioning in differentiating among students at different ability levels. The analysis of test items and distractor functioning can provide useful information for evaluating the diagnostic potential of items. Generally speaking, items that are too difficult or too easy may have little diagnostic capacity for differentiating among students with different skill mastery profiles (Jang, 2005). A diagnostic task should be more useful for differentiating among different skill profiles of students rather than for discriminating among students (Mislevy, 1995).

Verbal protocol analysis offers insights into why moderate task difficulty is ideal for providing diagnostically useful test information. Leighton (2004: 11) notes:

> Tasks of moderate difficulty are ideal because these tasks require a student to engage in controlled cognitive processing; that is, cognitive processing in which knowledge and strategies are selected consciously, thus allowing the student to be aware of his or her thinking as he or she solves a problem. In contrast, tasks that are too easy for students

to solve are not good candidates for eliciting verbal reports because these tasks do not require controlled processing for their solution, eliciting rapid recall instead. This rapid recall or automatic cognitive processing leaves the student unaware of how the problem was solved.

If tasks with extreme difficulty levels fail to elicit students' cognitive processing, the effect of diagnostic feedback may remain limited because the learner may be unconscious of the problem-solving process or because difficult tasks overload working memory (Robinson, 2001). Therefore, tasks with moderate difficulty that are novel enough to evoke conscious deliberation may be good candidates for designing a diagnostic assessment.

Further issues in diagnostic assessment

What is cognitive diagnostic assessment?

In recent years, cognitive diagnostic assessment (CDA) and its psychometric modeling have been applied widely in the field of language assessment. CDA utilizes latent class modeling approaches either by linking students' test item responses to a set of user-determined cognitive skills (Roussos *et al*., 2009) or by using diagnostic criteria that summarise response patterns and erroneous misconceptions for statistical classifications (Buck and Tatsuoka, 1998; Tatsuoka, 1983). The resulting diagnostic profiles describe individual students' mastery status for each skill in detail (Jang, 2009; Rupp *et al*., 2010).

In the early years of this application, Jang (2005, 2009) applied CDA to the TOEFL and determined that the CDA modeling approach generated reliable diagnostic feedback in the form of skill profiles that both teachers and students found useful. Further, the CDA approach identified assessment items that lacked discriminatory power, which supports ongoing test maintenance. Subsequently, entire issues of language testing journals have been dedicated to CDA and its psychometric modeling (e.g., Lee and Sawaki, 2009), and CDA has seen wide application in research and practice in both higher education and K–12 settings. For example, Kim (2015) describes successfully retrofitting CDA to a college-entry L2 reading test used for language support placement; the diagnostic reading profiles can be used by language educators to tailor their program to learners' needs. Such retrofitting applications are not uncommon, as recently demonstrated by Chen and Chen (2016) with the reading test of the Programme for International Student Assessment (PISA) and Toprak *et al*. (2019) with the Cambridge-Michigan Language Assessment listening test. Myriad retrofitting examples have been demonstrated in both foreign and second language settings globally (e.g., Aryadoust, 2018; Effatpanah *et al*., 2019; Yi, 2017a, 2017b). New language assessments have also been developed with the intended purpose of CDA; for example, Ranjbaran and Alavi (2017) describe the development of a reading comprehension assessment for a general English course at the University of Tehran, finding again that the resulting diagnostic feedback was useful for planning and teaching. Applications of CDA to a diverse range of educational contexts internationally is now quite feasible, in part due to the availability of free software such as R, which now has several packages for performing CDA modeling (Ravand, 2016; Ravand and Robotzsch, 2015).

Nevertheless, a dichotomous determination of skill mastery in CDA can be problematic and restrictive. Alternately, a diagnostic standard-setting process can be used to determine the diagnostic rules for classification (Rupp *et al*., 2010). Henson and Templin (2008) applied this process to the development of a diagnostic Algebra 2 assessment for high school

students. An expert panel specified the performance of students who demonstrated mastery of each skill and estimated the percentage of items answered correctly by a student who has mastered the targeted skill, and they then established a set of diagnostic criteria for the classification of different skill mastery profiles. Henson and Templin (2008) concluded that the skill mastery profiles developed through the standard-setting procedure were congruent with teachers' evaluation of students' mastery levels in classrooms.

Yet the technical aspects of CDA only partially tell the story of why it can be effective in language assessment. The strength of this approach is in the information it offers. Jang *et al*. (2017) recount a reading intervention for young learners wherein diagnostic feedback was successfully used as a tool to instigate metacognition and goal setting; in fact, the feedback provided essential data for the participants to self-assess their abilities and thus increase their agency over their learning.

While CDA has become prevalent, pedagogical and language-learning theories must be incorporated for the approaches to be valid. Addressing their concern of a lack of theory supporting CDA in most second and foreign language assessment applications, Alderson *et al*. (2015) researched the notion of diagnosis across fields (i.e., medicine, automotive mechanics, education). Despite the disparate natures of these fields, some semantic similarities exist, and the authors concluded with five principles of diagnostic language assessment practice. First, the test itself does not diagnose; it is the user of the test who diagnoses. Second, measures used for this purpose should be user friendly and effective for all stakeholders. Third, learner self-assessments and other views from diverse stakeholders should be elicited during the diagnostic assessment process. Fourth, the diagnostic process should not be truncated: it should begin with informal observation and development of a hypothesis about students' language strengths and weaknesses. Fifth, diagnosis must lead to positive actions, such as instructional intervention. These principles provide actionable steps for educators, assessment developers, and researchers to improve the validity of CDA applications in language assessment; Harding *et al*. (2015) also provide a worked example as further support. Jang (2017) describes ways in which technological advances can support such principles, such as allowing for less intrusive data gathering and real-time feedback; further, analytical techniques such as machine learning can incorporate more data points into analyses, thus increasing capacity for contextualisation, even in large-scale assessments. These technological advances for diagnostic assessment are discussed later in this chapter.

Considering that the development of a systematic diagnostic assessment requires financial and human resources in order to bring together content experts and assessment specialists, the further development of cognitive diagnostic assessment is likely to be large scale. Given its current stage of development, diagnostic assessment needs to be investigated systematically and empirically both in large- and small-scale applications through collaboration among different expert groups.

Washback effects of diagnostic assessment

Much research has focused on the negative impact of externally mandated tests on pedagogy (Alderson and Wall, 1993; Cheng *et al*., 2004; Cheng and Sultana, this volume). Little research exists, however, on washback issues associated with school-based diagnostic assessments, as Spolsky (1992: 30) points out:

> For those who worked to develop proficiency tests, any influence that the test has on instruction is usually a side effect, which is generally called backwash . . . and

> unfortunately occurs when preparation for the examination causes students to concentrate on unwanted skills, such as memorizing possible essays or practicing multiple-choice techniques. This concern for curriculum-fair testing then deflected attention from the fundamental instructional relevance of language tests.

Fulcher and Davidson (2007, 2009) emphasize an "effect-driven" approach as an important principle to test design and validation. Unlike high-stakes proficiency tests, pedagogical changes through a systematic diagnostic feedback loop are part of *intended effects* in diagnostic assessment instead of the *impact* of externally mandated testing. Because of its strong association with external high-stakes testing, the term *washback* needs to be reconceptualized in light of *effect* when it is discussed for diagnostic assessment in classroom settings.

Diagnostic assessment can also possibly have a negative effect or washback on teaching and learning. For example, whether too fine-grained construct definitions used for diagnostic assessment can narrow the scope of pedagogy is subject to debate (Collins, 1990; Jang, 2005; Leighton, 2009). Collins (1990) highlights two issues: *context* and *purpose*. His emphasis on the context where "diagnosis and remediation are tightly coupled" supports the effect-driven approach to diagnostic assessment deeply grounded in instructional contexts. It is the *purpose* that distinguishes diagnostic assessment from high-stakes tests. The stakes of diagnostic assessment may rise when diagnosis focuses on identifying only deficits, which could have a profoundly negative impact on students' self-esteem (Glazer and Searfoss, 1988). Diagnostic profiles need to highlight both what learners can do and what areas need further improvement.

Use of diagnostic feedback

What follows diagnostic assessment is often overlooked relative to research on assessment design and scoring. However, feedback utilization merits extensive attention because it is a core mechanism supporting the validation of such assessments. Validity hinges largely on the meaningfulness and appropriateness of the inferences generated from scores and, specifically, how these inferences pertain to learners' anticipated performance in the target domain (see Chapelle and Lee, this volume; Ross, this volume). Examining how feedback is perceived and interpreted by learners and teachers can provide specific evidence in this key aspect of assessment validity. In seeking such evidence, one needs to closely examine a mechanism of change through diagnostic feedback by articulating how it mediates students' cognition and metacognition.

Research shows that students' uptake of feedback differs by their self-concepts, such as self-perceived ability or learning orientations (Dweck, 1986). Jang *et al*. (2015) examined how elementary students with different learning orientations and self-perceived ability interpreted and responded to diagnostic feedback about their reading skills. Students with different learning orientations did not differ statistically in their performance, which aligns with existing literature; however, students with a mastery orientation did engage more deeply and critically with the diagnostic feedback than did performance-oriented students.

The study further showed that most students demonstrated discrepancies between their actual and self-assessed reading abilities. Of particular concern were learners with less-developed reading abilities who exhibited inflated perceptions of their own reading abilities. Students' inflated and deflated perceived abilities are shown to be strongly associated with their language and literacy abilities (Butler and Lee, 2010; Jang *et al*., 2015). In particular, poor feedback users tended to inflate their ability, also known as the

Dunning-Kruger effect (1999), suggesting the significant role of metacognition in learning (Trofimovich *et al*., 2016).

Characteristics of mediated diagnostic intervention

Dynamic interactions between students and their social environment heavily influence their learning process and outcomes (Byrne *et al*., 2002; Shepard, 2005; Vygotsky, 1978). One such interaction in the classroom is dialogic mediation provided for instruction and assessment.

> Assessment and instruction are a single activity that seeks to simultaneously diagnose and promote learner development by offering learners mediation, a qualitatively different form of support from feedback. Mediation is provided during the assessment procedure and is intended to bring to light underlying problems and help learners overcome them.
>
> *(Lantolf and Poehner, 2008: 273)*

It has been demonstrated that diagnostic interventions provided to students through dialogic mediation increase students' engagement and control of the reading comprehension process (Woolley, 2010). Individual feedback interventions (Jang *et al*., 2015) over eight sessions showed that as students became familiar with the nature of mediation, there were fewer questions for clarifications and less prompting by mediators. Through a guided process of interpreting and comparing their self-assessment and diagnostic reading skill assessment, learners engaged in critical metacognitive reflection and planning, supporting them to be more efficacious self-assessors and to better understand their areas to focus on for future improvement.

This mediation-based diagnostic intervention approach is similar to interactionist and interventionist dynamic assessment approaches (Lantolf, 2009; Poehner and Lantolf, 2005). In these approaches, mediation plays a key role in engaging the student in scaffolded learning, which is continually adjusted and negotiated with the student based on the initial diagnostic assessment and his or her responsivity. Students' responsivity to mediation is further influenced by their non-cognitive traits, such as interest, motivation, and emotional regulation. Although the learner's cognitive language ability is often the most important pedagogical goal, it does not operate in isolation; instead, it develops at the confluence of affective, metacognitive, and motivational traits. This is well reflected in complexity theory, which posits the learner as a unique learning system (Larsen-Freeman and Cameron, 2008). Nested in a larger social environment, the complex learner system adapts to the social environment through social interactions.

In mediated diagnostic assessment interventions, the underlying dialogic interaction between the mediator and the learner is *scaffolding*. It is an act of teaching that supports the immediate construction of knowledge by the learner and provides the basis of future independent learning of the individual (Holton and Clarke, 2006). Holton and Clarke (2006) describe three types of scaffolding that operate within the domains of conceptual and heuristic scaffolding. Conceptual scaffolding aims to promote concept development, while heuristic scaffolding focuses on investigative learning skills or problem-solving abilities. The first type of scaffolding is expert scaffolding, which is provided by the mediator or a teacher – someone who understands the material to provide both conceptual and heuristic scaffolding. The second type is reciprocal scaffolding, which involves two (or more) students working

collaboratively, through peer-to-peer interactions where each student contributes in the form of scaffolding. The third type is self-scaffolding, in which the student can provide scaffolding for herself to tackle new concepts or skills. The ideal development of learning is for a student to transition from expert scaffolding to self-scaffolding for a particular concept or skill. To maximize the effects of diagnostic feedback on student learning, we need to articulate the mechanism of change by specifying the learning goal (conceptual vs. heuristic), learner traits composing the complex learning system, and mediation/scaffolding types and dosage.

Technological advances in diagnostic assessment

The fields of language education and assessment are rapidly advancing with technological innovations. Technological advances allow us to make assessment activities less intrusive, real time, and interactive. Automated scoring through the application of natural language processing to constructed responses helps overcome overreliance on selected-response assessment items, which tend to assess discrete knowledge over processing skills (see Xi, this volume). Machine learning further supports comprehensive learner modeling based on process-oriented multichannel data involving physiological, neural, affective, behavioral, and metacognitive traces. Real-time data processing and analysis allow for timely and adaptive assessment feedback for individual students.

In assessing students' speaking proficiency, digitized speech and video offer greater authenticity to assessment: automated speech recognition techniques allow students to record their voices while performing a task (Cheng *et al*., 2014; Evanini *et al*., 2015). For example, automated rating programs, such as Educational Testing Service's Criterion, provide detailed feedback based on error analysis (Chapelle and Douglas, 2006). Another example is a novel multifaceted computerized assessment system being developed for elementary teachers and students, called BalanceAI (Jang and Sinclair, 2019; Sinclair *et al*., 2021), which utilizes artificial intelligence to provide detailed feedback about learners' areas of strength and areas to improve. Rather than navigating the typical constraints on language and literacy assessment – the questionable validity of selected-response (e.g., multiple-choice) items and the expense of human scoring of constructed responses – BalanceAI uses natural language processing to extract a multitude of speech and textual data about learners' skills, including semantic content (useful for story retelling, picture description, and writing tasks), grammatical complexity, and lexical sophistication. Natural language processing can produce tens of thousands of data points per student, so traditional statistics (even non-parametric approaches) generally lack sufficient degrees of freedom to converge. Thus, machine-learning analytic techniques are optimal for developing language skill profiles based on such rich speech and textual data sets. Since cognitive diagnostic assessment centers on the development of learners' skill profiles for feedback and instruction, the advent of natural language processing and machine learning present enormous potential for developing skill profiles with speech and textual data beyond selected responses.

In assessing reading comprehension, digital assessment can gather process data, such as the frequency of glossary and electronic dictionary use from computer logs, facial expressions of emotion, physiological heart rates, gestures, speech, and eye-tracking methods (Rayner, 1998; Winke, 2013). An increase in the learner's heart rate during reading, for example, may indicate difficulty with reading, signaling the presence of external factors influencing reading processes. Through triangulation with other data sources, such as lexical complexity indices, physiological responses can help identify the source of reading

difficulty. For another example, eye-tracking methods have been used to observe the quantity and quality of readers' attention to the text and task features in language assessment. Eye-tracking data coupled with think-aloud protocols can pinpoint which features learners fixated on, and verbalized thought processes can further substantiate inferences made from eye-gaze data, such as repeated fixations and saccades between the text and task. Bax (2013) and Brunfaut and McCray (2015) show how eye-gaze data can provide information about differences in cognitive processing between proficient and less proficient readers (see also Brunfaut, this volume). In Bax's study, proficient readers located key information in the text more effectively and fixated less on the task features than less proficient readers, suggesting higher efficiency in information processing. In addition, both studies highlight a varying level of cognitive processing and strategy use across different assessment task types.

Along with eye-tracking methods, facial expression analysis may be used to gather information about students' emotional engagement and regulation during assessment tasks. Pekrun *et al*. (2002) define academic emotions as emotional experiences learners have during academic tasks. Task-related academic emotions can facilitate or interfere with cognitive engagement in academic tasks. Task-related emotions include anxiety, enjoyment, hope, pride, relief, anger, boredom, and shame. Test anxiety in particular has been a major focus for research in the field of language assessment to date. These technological advances will have a profound impact on both what and how to assess, as well as how to provide useful feedback.

Future directions

In 1992, Spolsky stated, "The task is not simple, but if we can succeed, we will surely be on the way to developing responsible postmodern language testing" (36). We have begun to pay full attention to the nature of diagnostic assessment and to experiment with its potential to bring close theories of language learning and assessment. In closing, we call for research in four main areas for diagnostic language assessment to reach its full fruition in the next ten years.

First, learning-oriented assessment approaches shift the focus of assessment toward the learner's cognitive, metacognitive, and affective traits and growth while capitalising on the contextual influence. Considering the learner as a complex dynamic system (Larsen-Freeman and Cameron, 2008; Smith and Thelen, 2003; van Geert and van Dijk, 2002), diagnostic assessment should not be limited to cognitive competence. Instead, it needs to focus on the mechanism of competence development, the influence of intrapersonal traits, and their interactions with contextual factors, such as assessment settings, tasks, and assessors. Intra-personal variability is no longer viewed as a source of unsystematic measurement errors, as classical test theory has argued. Rather, it can be a driving force of development (van Geert and van Dijk, 2002). Future language diagnostic assessment will embrace the opportunity to assess, model, track, and scaffold progressions as a result of such interactions. Interdisciplinary research collaborations will advance the field of language assessment through dynamic learner modeling and scaffolding based on multichannel process data involving physiological, neural, affective, behavioural, and metacognitive traces. Real-time data processing and an interactive feedback loop will allow for timely and adaptive assessment for individual students.

Second, technologies should not be taken as a threat to classroom teachers and their pedagogical and assessment roles. They are not meant to replace teachers but to serve as instructional and assessment aids by enhancing the efficiency of pedagogical documentation.

On a daily basis, teachers use a variety of instructional and assessment methods, such as questioning, observing, conversing, and quizzing students to understand what they know and do not know (Black and Wiliam, 1998; Rea-Dickins, 2001). However, research shows that teachers often do not design and implement such activities with diagnostic assessment in mind, spend a small amount of time assessing individual students, and lack diagnostic assessment competencies largely due to little professional development in assessment (Edelenbos and Kubanek-German, 2004). User-friendly technology-based tools already exist that greatly facilitate such documentation and can provide a bridge to developing assessment competencies.

Teachers often express their desire to develop (diagnostic) assessment competence: the ability to assess students' learning progress, interpret the results from various sources of assessment, and provide pedagogical intervention for students' improvement (Edelenbos and Kubanek-German, 2004; Stiggins, 1997). One way to develop teachers' diagnostic assessment competence is through systematic reflections on learners' performance guided by relevant knowledge in a variety of learning contexts (Cumming, 2009; Fulcher and Davidson, 2007; McNamara, 2000). The use of ongoing diagnostic assessment affords teachers the opportunity to target their students' specific needs and promote curricular innovations.

Along with teachers' assessment competence development, students' responsivity to diagnostic feedback varies depending on their beliefs and learning environments. Students tend to react to diagnostic feedback by comparing it with their own self-assessment and decide whether or not to focus their learning on the areas of improvement (Jang, 2005, 2009). Regardless of whether or not students' self-assessment is reliable and accurate, students' beliefs about their own competencies should be consulted in tailoring instructional intervention and evaluating the effect of diagnostic feedback. Black (2009: 521) notes:

> However, more is involved than a merely cognitive exercise, for the response may be controlled by a desire to protect one's own sense of well-being. For example, a learner may refuse to engage for fear of appearing stupid. . . . One implication is that a focus on marks and grades, on written work or in a regime of frequent testing, can do positive harm by developing obstacles to engagement, effects which can be as harmful to the high achievers as to the low.

The view of students as change agents in diagnostic assessment should consider non-cognitive learner characteristics. Positive learning outcomes attributable to formative feedback are correlated with students' attitudes toward learning, especially mastery learning (Black and Wiliam, 1998; Dweck, 1986). More research is necessary for understanding what non-cognitive variables mediate students' use of diagnostic feedback for improving their own learning.

Third, research on diagnostic assessment will contribute to developing various theoretical models of language ability, which are vital for characterizing what to diagnose. While there are competing theories of language ability, we do not know much about their viability for diagnostic assessment. Fully functional models of language development for diagnosis require micro-level analyses of both cognitive and non-cognitive characteristics of classroom-based language learning. We will benefit from active classroom research that experiments with various theoretical models of language development and provides a thick description of learners' language use and development.

Lastly, principled diagnostic assessment design will contribute to enhancing the quality of diagnostic assessment in language classrooms. Existing, excellent test design frameworks,

such as evidence-centered design (Mislevy *et al*., 2002) and effect-driven test development (Fulcher and Davidson, 2007), can guide research on diagnostic assessment construction and validation. Furthermore, these principles can help us develop criteria for evaluating the quality of diagnostic assessment tasks. Students' deep-thinking processes should be micro-analysed to understand how tasks elicit their processing skills. Diagnostically rich tasks can evoke students' conscious deliberation. Granularity (or specificity) underlying diagnostic assessment is an exciting issue that links a theory of language, assessment purpose, task, scoring, and use. Diagnostic assessment is not a simple task, yet the complexity that it illuminates and the promise that it holds for teachers and students inform compelling arguments for future research.

Further reading

Alderson, J. C., Brunfaut, T. and Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics* 36: 236–60.

This article reconceptualizes the notion of diagnosis in a broader context across multiple professional fields and offers commonalities applicable to language assessment. Based on interviews with professionals from car mechanics, IT systems support, medicine, psychology, and education, the study discusses five principles of diagnostic language assessment.

Cheng, J., D'Antilio, Y. Z., Chen, X. and Bernstein, J. (2014). *Automated Assessment of the Speech of Young English Learners*. Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications. Baltimore, MD: Association for Computational Linguistics, 12–21.

This set of proceedings describes the development of a machine learning model to score multilingual children's responses to a speaking test with 14 different item types. The average correlation between human scores and the algorithm (using a 0–4 holistic rating) was .92. This study explains each step of the model development process, including both acoustic and language content modeling, and explains results and limitations by student age, from kindergarten through high school.

Jang, E. E., Dunlop, M., Park, G. and van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing* 32: 359–83.

This study examines how holistic diagnostic feedback is interpreted and used by young readers with different cognitive and psychological profiles. The holistic diagnostic feedback consists of information about skill mastery levels, students' self-assessed skill proficiency, and goal orientations. The study results show that students' interpretations and use of feedback for future learning differ significantly by their perceptions about ability and goal orientations. The study highlights that the mechanism of diagnostic feedback use must consider learner characteristics holistically in order to maximise its effects.

Larsen-Freeman, D. (2012). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? *Language Teaching* 45: 202–14.

This article introduces complexity theory as a transdisciplinary theme, which holds a significant implication for applied linguistics. The premise is that the language system can be understood as a system of interdependent systems and parts that interact non-linearly. Language researchers should embrace this complexity by studying modes of system change with a focus on interconnected causes, instead of attempting to isolate individual variables. The author describes 12 principles underlying complexity theory and further delves into dynamism, complexity, and context.

Spolsky, B. (1992). The gentle art of diagnostic testing revisited. In E. Shohamy and A. R. Walton (eds.), *Language Assessment for Feedback: Testing and Other Strategies*. Dubuque, IA: Kentall, Hunt, 29–41.

In this chapter, Spolsky attributes a lack of attention to diagnostic assessment to commitments to proficiency testing devoid of curricular matters. He then discusses various approaches to diagnostic

testing from three different perspectives, including pedagogical, linguistic, and processing, as well as conflicts associated with test form and content, the competency-performance divide, and teachers' and students' roles in diagnostic testing.

References

Alderson, J. C. (2005). *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. London, UK: Continuum.

———. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner and C. Doe (eds.), *Language Testing Reconsidered*. Ottawa, Canada: University of Ottawa Press, 21–39. https://doi.org/10.2307/j.ctt1ckpccf.8.

Alderson, J. C., Brunfaut, T. and Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics* 36: 236–60. https://doi.org/10.1093/applin/amt046.

Alderson, J. C. and Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the common European framework. *Language Testing* 22: 301–20. https://doi.org/10.1191/0265532205lt310oa.

Alderson, J. C. and Wall, D. (1993). Does wash back exist? *Applied Linguistics* 14: 115–29. https://doi.org/10.1093/applin/14.2.115.

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore-Cambridge General Certificate of Education O-Level: Application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *International Journal of Listening* 35: 29–52. https://doi.org/10.1080/10904018.2018.1500915.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing* 30: 441–65. https://doi.org/10.1177/0265532212473244.

Beaver, J. (2001). *Developmental Reading Assessment K-3 Teacher Resource Guide*. Parsippany, NJ: Pearson Learning.

Bejar, I. I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement* 21: 175–89. Retrieved from www.jstor.org/stable/1434541.

Black, P. J. (2009). Formative assessment issues across the curriculum: The theory and the practice. *TESOL Quarterly* 43: 519–23. Retrieved from www.jstor.org/stable/27785033.

Black, P. J. and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education* 5: 7–74. https://doi.org/10.1080/0969595980050102.

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing* 15: 45–85. https://doi.org/10.1177/026553229801500103.

Brunfaut, T. and McCray, G. (2015). *Looking into Test-Takers' Cognitive Processes Whilst Completing Reading Tasks: A Mixed-Method Eye-Tracking and Stimulated Recall Study*. (ARAGs Research Reports Online; Vol. AR/2015/001). London, UK: The British Council. Retrieved from www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf.

Buck, G. and Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing* 15: 119–57. https://doi.org/10.1177/026553229801500201.

Burgin, J. and Hughes, G. D. (2009). Credibly assessing reading and writing abilities for both elementary student and program assessment. *Assessing Writing* 14: 25–37.

Butler, Y. and Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing* 27: 5–31. https://doi.org/10.1177/0265532209346370.

Byrne, S., Todd, D., Simpson, B., Woods, C. and Seidel, R. (2002). Inquiring into learning as system. In G. Ragsdell, D. West and J. Wilby (eds.), *Systems Theory and Practice in the Knowledge Age*. Boston, MA: Springer, 115–22.

Byrnes, H. (ed.). (2007). Perspectives. *Modern Language Journal* 91: 641–85. https://doi.org/10.1016/j.asw.2008.12.001.

Chapelle, C. A., Chung, Y. R. and Xu, J. (eds.). (2008). *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment*. Ames, IA: Iowa State University.

Chapelle, C. A. and Douglas, D. (2006). *Assessing Language Through Computer Technology*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511733116.

Chen, H. and Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the generalized DINA model framework. *Language Assessment Quarterly* 13: 218–30. https://doi.org/10.1080/15434303.2016.1210610.

Cheng, J., D'Antilio, Y. Z., Chen, X. and Bernstein, J. (2014). *Automated Assessment of the Speech of Young English Learners*. Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications. Baltimore, MD: Association for Computational Linguistics, 12–21.

Cheng, L., Watanabe, Y. and Curtis, A. (2004). *Washback in Language Testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J. (ed.). (2001). *Setting Performance Standards*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J., Bunch, M. B. and Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice* 23: 31–50. https://doi.org/10.1111/j.1745-3992.2004.tb00166.x.

Clay, M. M. (2000). *Running Records for Classroom Teachers*. Auckland, NZ: Heinemann Publishers.

Collins, A. (1990). Reformulating testing to measure learning and thinking. In N. Frederiksen, R. Glaser, A. Lesgold and M. Shafto (eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates, 75–88.

Cumming, A. (2009). Language assessment in education: Tests, curricula, and teaching. *Annual Review of Applied Linguistics* 29: 90–100. https://doi.org/10.1017/S0267190509090084.

Davidson, F. and Lynch, B. K. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. London, UK: Yale University Press.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist* 41: 1040–48. https://doi.org/10.1037/0003-066X.41.10.1040.

Edelenbos, P. and Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence.' *Language Testing* 21: 259–83. https://doi.org/10.1191/0265532204lt284oa.

Effatpanah, F., Baghaei, P. and Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia* 9: 1–23. https://doi.org/10.1186/s40468-019-0090-y.

Evanini, K., Heilman, M., Wang, X. and Blanchard, D. (2015). Automated scoring for the "TOEFL Junior"® comprehensive writing and speaking test: Research report. *ETS Research Report Series*: 1–13. https://doi.org/10.1002/ets2.12052.

Fox, J. D. (2009). Moderating top-down policy impact and supporting EAP curricular renewal: Exploring the potential of diagnostic assessment. *Journal of English for Academic Purposes* 8: 26–42. https://doi.org/10.1016/j.jeap.2008.12.004.

Frederiksen, N., Glaser, R., Lesgold, A. and Shafto, M. (eds.). (1990). *Diagnostic Monitoring of Skill and Knowledge Acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment*. London, UK: Routledge.

———. (2009). Test architecture. Test retrofit. *Language Testing* 26: 123–44. https://doi.org/10.1177/0265532208097339.

Gibson, A. (2007). *CASI Grades 4 to 8 Reading Assessment Teacher's Guide*, 2nd edn. Toronto, ON: Thomson Nelson.

Glazer, S. M. and Searfoss, L. W. (1988). Reexamining reading diagnosis. In S. M. Glazer, L. W. Searfoss and L. M. Gentile (eds.), *Reexamining Reading Diagnosis: New Trends and Procedures*. Newark, NJ: International Reading Association, 1–11.

Gu, Z. and Jang, E. E. (2009). *Investigating the Diagnostic Value of Multiple-Choice Options for Cognitive Diagnostic Assessment*. Paper presented at the Canadian Society for the Study of Education (CSSE), Ottawa, Canada.

Harding, L., Alderson, J. C. and Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing* 32: 317–36. https://doi.org/10.1177/0265532214564505.

Henson, R. and Templin, J. (2008). *Implementation of Standards Setting for a Geometry End-of-Course Exam*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Holton, D. and Clarke, D. (2006). Scaffolding and metacognition. *International Journal of Mathematical Education in Science and Technology* 37: 127–43. https://doi.org/10.1080/00207390500285818.

Hudson, T. and Lynch, B. (1984). A criterion-referenced measurement approach to ESL achievement testing. *Language Testing* 1: 171–210. https://doi.org/10.1177/026553228400100204.

Huff, K. and Goodman, D. P. (2007). The demand for cognitive diagnostic assessment. In J. P. Leighton and M. J. Gierl (eds.), *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge, UK: Cambridge University Press, 19–60. https://doi.org/10.1017/cbo9780511611186.

Jang, E. E. (2005). *A Validity Narrative: Effects of Reading Skills Diagnosis on Teaching and Learning in the Context of NG TOEFL*. Unpublished doctoral dissertation, the University of Illinois, Urbana-Champaign.

———. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying fusion model to LanguEdge assessment. *Language Testing* 26: 31–73. https://doi.org/10.1177/0265532208097336.

———. (2010). *Implications of Assessment of School Age L2 Students in Ontario*. Symposium at Canadian Association of Language Assessment, Montreal, Canada.

Jang, E. E. (2017). Cognitive aspects of language assessment. *Language Testing and Assessment*: 163–77. https://doi.org/10.1007/978-3-319-02326-7_11-1.

Jang, E. E., Dunlop, M., Park, G. and van der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Language Testing* 32: 359–83. https://doi.org/10.1177/0265532215570924.

Jang, E. E. and Ryan, K. (2003). Bridging gaps among curriculum, teaching and learning, and assessment [Review of the book *Large-scale assessment: Dimensions, dilemmas, and policy]. Journal of Curriculum Studies* 35: 499–512. https://doi.org/10.1080/00220270305521.

Jang, E. E. and Sinclair, J. (2019). *Advancing Diagnostic Language Assessment Through Natural Language Processing-Based Machine Learning Approaches*. Paper presented at the Colloquium, Assessing Young Learners' English Language Proficiency, AAAL, Atlanta, GA.

Jang, E. E., Vincett, M., van der Boom, E. H., Lau, C. and Yang, Y. (2017). Considering young learners' characteristics in developing a diagnostic assessment intervention. In M. K. Wolf and Y. G. Butler (eds.), *English Language Proficiency Assessments for Young Learners*, vol. 2. New York, NY: Routledge, 193–213. https://doi.org/10.4324/9781315674391-11.

Kim, A. Y. (2015). Exploring ways to provide diagnostic feedback with an ESL placement test: Cognitive diagnostic assessment of L2 reading ability. *Language Testing* 32: 227–58. https://doi.org/10.1177/0265532214558457.

Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77: 1121–34. https://doi.org/10.1037/0022-3514.77.6.1121.

Kunnan, A. and Jang, E. E. (2009). Diagnostic feedback in language testing. In M. Long and C. Doughty (eds.), *The Handbook of Language Teaching*. Oxford, UK: Blackwell Publishing, 610–25. https://doi.org/10.1017/S0261444808005569.

Lantolf, J. P. (2009). Dynamic assessment: The dialectic integration of instruction and assessment. *Language Teaching* 42: 355–68. https://doi.org/10.1017/S0261444808005569.

Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics* 1: 49–72. https://doi.org/10.1558/japl.v1.i1.49.

Lantolf, J. P. and Poehner, M. E. (2008). Dynamic assessment. In N. H. Hornberger (ed.), *Encyclopedia of Language and Education*. New York: Springer, 273–84.

Larsen-Freeman, D. and Cameron, L. (2008). Research methodology on language development from a complex systems perspective. *The Modern Language Journal* 92: 200–13. Retrieved from www.jstor.org/stable/25173023.

Lee, Y. W., & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly* 6: 172–89. https://doi.org/10.1080/15434300902985108.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice* 23: 6–15. https://doi.org/10.1111/j.1745-3992.2004.tb00164.x.

———. (2009). Mistaken impressions of large-scale cognitive diagnostic testing. In R. P. Phelps (ed.), *Correcting Fallacies About Educational and Psychological Testing*. Washington, DC: American Psychological Association, 219–46. https://doi.org/10.1037/11861-000.

Leighton, J. P. and Gierl, M. J. (2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/cbo9780511611186.

Little, D. (2006). The common European framework of reference for languages: Content, purpose, origin, reception and impact. *Language Teaching* 39: 167–90. https://doi.org/10.1017/S0261444806003557.

McCarthy, A. M. and Christ, T. J. (2010). Test review. In J. M. Beaver and M. A. Carter (eds.), *The Developmental Reading Assessment*, 2nd edn (DRA2) (2006). Upper Saddle River, NJ: Pearson. *Assessment for Effective Intervention* 35: 182–85. https://doi.org/10.1177/1534508410363127.

McNamara, T. (2000). *Language Testing*. Oxford, UK: Oxford University Press.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. In P. D. Nichols, S. F. Chipman and R. L. Brennan (eds.), *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers, 43–72.

Mislevy, R. J., Steinberg, L. S. and Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing* 19: 477–96. https://doi.org/10.1191/0265532202lt241oa.

Mousavi, S. A. (2002). *An Encyclopedic Dictionary of Language Testing*, 3rd edn. Taiwan: Tung Hua Book Company.

Nichols, P. D., Meyers, J. L. and Burling, K. S. (2009). A framework for evaluating and planning assessments intended to improve student achievement. *Educational Measurement: Issues and Practice* 28: 14–23. https://doi.org/10.1111/j.1745-3992.2009.00150.x.

Nicols, P. D., Chipman, S. F. and Brennan, R. L. (eds.). (1995). *Cognitively Diagnostic Assessment*. Hillsdale, NJ: Lawrence Erlbaum Association, Publishers.

North, B. and Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15: 217–63. https://doi.org/10.1177/026553229801500204.

Pekrun, R., Goetz, T., Titz, W. and Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist* 37: 91–105. https://doi.org/10.1207/S15326985EP3702_4.

Poehner, M. E. and Lantolf, J. P. (2005). Dynamic assessment in the language classroom. *Language Teaching Research* 9: 233–65. https://doi.org/10.1191/1362168805lr166oa.

Ranjbaran, F. and Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation* 55: 167–79. https://doi.org/10.1016/j.stueduc.2017.10.007.

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *Journal of Psychoeducational Assessment* 34: 782–99. https://doi.org/10.1177/0734282915623053.

Ravand, H. and Robitzsch, A. (2015). Cognitive diagnostic modeling using R. *Practical Assessment, Research & Evaluation* 20: 1–12. https://doi.org/10.7275/5g6f-ak15.

Rayner, S. (1998). Educating pupils with emotional and behaviour difficulties: Pedagogy is the key!. *Emotional and Behavioural Difficulties* 3: 39–47. https://doi.org/10.1080/1363275980030206.

Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing* 18: 429–62. https://doi.org/10.1177/026553220101800407.

Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. *Journal of Educational Psychology* 93: 211–22. https://doi.org/10.1037/0022-0663.93.1.211.

Roussos, L., DiBello, L., Henson, R., Jang, E. E. and Templin, J. (2009). Skills diagnosis for education and psychology with IRT-based parametric latent class models. In S. E. Embretson and J. Roberts (eds.), *New Directions in Psychological Measurement with Model-based Approaches*. Washington, DC: American Psychological Association, 35–69.

Rupp, A., Templin, J. and Henson, R. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York, NY: Guilford.

Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership* 63: 66–70.

Sinclair, J., Jang, E. E. and Rudzicz, F. (2021). Using machine learning to predict children's reading comprehension from linguistic features extracted from speech and writing. *Journal of Educational Psychology.* (Advance online publication). https://doi.org/10.1037/edu0000658.

Smith, L. B. and Thelen, E. (2003). Development as a dynamic system. *Trends in Cognitive Sciences* 7: 343–48. https://doi.org/10.1016/S1364-6613(03)00156-6.

Spolsky, B. (1992). The gentle art of diagnostic testing revisited. In E. Shohamy and A. R. Walton (eds.), *Language Assessment for Feedback: Testing and Other Strategies*. Dubuque, IA: Kentall, Hunt, 29–41.

Stewart, R. (2016). *The Psychometric Properties of the Developmental Reading Assessment*. Unpublished Master's thesis. Retrieved from https://qspace.library.queensu.ca/bitstream/handle/1974/14823/Stewart_Rachel_A_201608_MED.pdf?sequence=1&isAllowed=y.

Stiggins, G. (1997). *Student Centered Classroom Assessment*. Upper Saddle River, NJ: Prentice Hall.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement* 20: 345–54. Retrieved from www.jstor.org/stable/1434951.

Toprak, T. E., Aryadoust, V. and Goh, C. (2019). The 3 log-linear cognitive diagnosis modeling (LCDM) in second language listening assessment. In V. Aryadoust and M. Raquel (eds.), *Quantitative Data Analysis for Language Assessment Volume II: Advanced Methods*. London, UK: Routledge, 56–78. https://doi.org/10.4324/9781315187808.

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K. and Crowther, D. (2016). Flawed self-assessment: Investigating self-and other-perception of second language speech. *Bilingualism: Language and Cognition* 19: 122–40. https://doi.org/10.1017/S1366728914000832.

van Geert, P. and van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior & Development* 25: 340–74. https://doi.org/10.1016/S0163-6383(02)00140-6.

Vygotsky, L. S. (1978). Socio-cultural theory. *Mind in Society* 6: 52–8.

Winke, P. M. (2013). The effects of input enhancement on grammar learning and comprehension: A modified replication of Lee (2007) with eye-movement data. *Studies in Second Language Acquisition* 35: 323–52. https://doi.org/10.1017/S0272263112000903.

Woolley, G. (2010). Developing reading comprehension: Combining visual and verbal cognitive processes. *Australian Journal of Language and Literacy* 33: 108–25.

Yi, Y. S. (2017a). In search of optimal cognitive diagnostic model (s) for ESL grammar test data. *Applied Measurement in Education* 30: 82–101. https://doi.org/10.1080/08957347.2017.1283314.

———. (2017b). Probing the relative importance of different attributes in L2 reading and listening comprehension items: An application of cognitive diagnostic models. *Language Testing* 34: 337–55. https://doi.org/10.1177/0265532216646141.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Section 4

Assessing the language skills


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

12

Assessing speaking

Fumiyo Nakatsuhara, Nahal Khabbazbashi, and Chihiro Inoue

Introduction

Over half a century ago, Lado (1961: 239) stated:

> the ability to speak a foreign language is without doubt the most highly prized language skill. . . . Yet testing the ability to speak a foreign language is perhaps the least developed and the least practiced in the language testing field.

The field of spoken assessment has come a long way since Lado's observation. One need look no further than the wide range of international and local speaking tests available worldwide as well as the wealth of research studies on the subject. The choices of tests offered to users vary along several parameters such as task format, delivery mode, and scoring methods, all of which have important implications in terms of construct coverage and practical demands. Recent advances in digital technologies and disruptive innovations have been an accelerating force in challenging current practices, introducing new research methodologies, and contributing to the diversification of speaking assessments resulting in a field in which the state of the art is constantly in flux.

Before exploring some of these parameters and assessment practices in more depth, it is important to first provide a historical overview of the testing of speaking. While any such account is by its nature partial "and at the mercy" of available evidence (Weir *et al*., 2013: 1), we can only understand where we are now by looking back at where we have come from.

Historical perspectives

While spoken assessment has a long history in content-based educational settings, where writing was considered ancillary to speaking (Latham, 1877; Russel, 2002), the history of assessing second language (L2) speaking seems to have emerged more recently; see Fulcher (2015a) for a comprehensive overview of the history of speaking assessment and areas of prominent research.

DOI: 10.4324/9781003220756-17

The Cambridge Certificate of Proficiency in English was launched in 1913 and was predominantly designed for English teacher trainees. It is one of the earliest L2 proficiency tests that featured speaking components – dictation, read-aloud, and conversation tasks in an examiner-candidate format, as well as a written paper on knowledge of phonetics (Weir *et al*., 2013). In its original form (1913–45), the test mirrored prominent approaches to L2 speaking in late-nineteenth-century Europe, such as the Direct Method and the Oral Method with a primary focus on oral fluency (Palmer, 1921) and the Reform Movement, which emphasised the importance of spoken language – in particular phonetics (Sweet, 1899).

The experience of the two World Wars and pressing military needs led to an exponential growth in L2 oral assessment research and practices in the mid-twentieth century. In particular, the US Foreign Service Institute (FSI) Oral Proficiency Interview (OPI), operationalised in 1952, was a significant initial step toward defining and measuring a multi-trait speaking construct reliably. The original form of the test was accompanied by a six-band holistic scale with performance descriptors defining criterial features at each level, and a checklist of five analytic criteria (accent, comprehension, fluency, grammar, and vocabulary) was added in 1958 to address variation in scores psychometrically (Sollenberger, 1978). The OPI test was seminal in the history of speaking assessment, paving the way for the booming of subsequent oral assessment initiatives in the US in the 1960s to 1980s (Fulcher, 2003). These include the Interagency Language Roundtable (ILR) speaking test and the American Council for the Teaching of Foreign Languages (ACTFL) OPI. In the UK, the University of Cambridge Local Examinations Syndicate (UCLES) also started denoting performance descriptors in the 1970s in an attempt to define the test construct more explicitly and to standardise speaking examiners' rating judgments. However, the development of rating scales and performance descriptors in those days was primarily based on expert judgment and intuition rather than empirical evidence (Fulcher, 2003). The fields of language testing, second language acquisition, and applied linguistics were still young, with little research that could inform the understanding of the nature of L2 speaking ability and performance.

The landscape of the field looked vastly different in the 1990s; theoretical models such as Bachman's (1990) model of communicative language ability started to emerge, along with empirical research exploring those models. Several approaches to developing and validating rating scales – and, by extension, understanding the measured speaking construct – were examined. They include Fulcher's (1996) analysis of learner speech to develop data-driven rating scales, Pollitt and Murray's (1996) investigation into examiners' perceptions of proficiency when rating spoken performances, and North and Schneider's (1998) measurement-driven approach to scale development as embodied in the CEFR.

Since the 1990s, the constructs measured in spoken tests have become increasingly more diverse as represented in large-scale international speaking assessments. To illustrate, in addition to the traditional examiner-candidate interview format as in the IELTS speaking test since 1989, a paired speaking format was introduced in the Cambridge Main Suite examinations during the 1990s and early 2000s. This allowed for a broadening of the speaking construct and the assessment of richer language of an interactional nature (Galaczi, 2014). While the importance of *a social and interactional trait* became more prominent on the one hand, an automated approach to scoring spoken responses, as used in the Pearson Test of English Academic launched in 2009, focused emphasis on *a cognitive, psycholinguistic construct*, on the other (van Moere, 2012). Semi-direct speaking tests – as characterised by computer (or online) delivery of tasks, digital response capturing, and human rating – such as the British Council's Aptis speaking test introduced in 2012 (O'Sullivan and Dunlea, 2020), target a selected range of spoken language elicited in monologic and simulated Q and

A formats. Advances in computer technologies facilitated widening the construct assessed in semi-direct tests; for instance, TOEFL iBT, launched in 2005 (Chapelle *et al*., 2008), introduced integrated tasks of listening and speaking and of reading, listening, and speaking. In direct speaking tests as well, to embrace the role of listening necessary for effective interaction (Nakatsuhara, 2018), candidates' listening comprehension skills are now assessed as part of some interactional speaking tests – see, for example, the speaking and listening rating scales for Trinity's Integrated Skills in English (ISE) examinations (Trinity College London, 2021).

Critical issues

This historical overview illuminates some of the major influences – theoretical, pedagogical, social, political, and practical – on the conceptualisation and operationalisation of the speaking construct and how it has shifted over different periods of time. In the twenty-first century, it is technology that is arguably the strongest influence presenting both challenges and opportunities in the assessment of speaking. In the next few sections, we draw on the latest research and practice to examine some of the most important parameters in speaking assessment in the digital age and highlight critical issues.

Speaking tasks and measured constructs

At the heart of speaking assessment are the constructs that we intend to tap into, and test tasks play a fundamental role in operationalising these constructs and eliciting performances. While several task classification systems exist, a useful approach is to distinguish tasks in terms of the conditions in which candidates' output language is elicited. On this basis, we have broadly classified the most common speaking task types into four categories of elicited imitation, independent, integrated, and interactive tasks and will discuss each in more detail, highlighting their relative strengths and limitations.

Elicited imitation tasks

Elicited imitation (EI) refers to an assessment technique in which candidates are required to listen to utterances and repeat them verbatim or to read aloud written sentences (Yan *et al*., 2016). While their popularity declined in the 1980s with the communicative approaches to language teaching and assessment, they have made a comeback in recent years with the rise of automated speech evaluation (ASE) systems. The use of EI tasks is closely related to the functioning of the speech recogniser element of ASE systems; predictable speech can be recognised and scored with greater accuracy and precision than spontaneous unpredictable speech (Xi *et al*., 2012).

There has been much debate in the language testing literature regarding EI tasks and their construct validity: for example, their successful completion has been suggested to reflect parsing and comprehension of the input (Bley-Vroman and Chaudron, 1994) while others attribute it to phonological short-term memory capacities (Gathercole and Baddeley, 1993). Van Moere (2012: 325) suggests that these tasks can tap into processing competence defined as "the efficiency with which learners process language" – a core competence for any L2 spoken production and interaction. Yan *et al*.'s (2016) meta-analysis of EI studies generally supported their use as a reliable measure of L2 spoken proficiency. On the other hand, these tasks have been criticised for their narrow representation of the speaking construct and their

limited authenticity (c.f. Galaczi, 2010; Yan *et al*., 2016); the tasks, in themselves, do not require candidates to generate ideas or to draw on lexical resources and syntactic knowledge to translate ideas into speech – processes which are essential for the spontaneous production of speech (Field, 2011). A suggestion put forward by Yan *et al*. (2016) is to re-evaluate EIs in light of the research and to combine (rather than replace) them with other communicative language tasks in order to build on strengths but also to address their limitations.

Independent tasks

Independent speaking tasks require candidates to "draw on their own ideas or knowledge [in order] to respond to a question or prompt" (Brown *et al*., 2005: 1) and commonly include question and answer (Q&A), picture description, and individual long-turn tasks. Among these (simulated) Q&A tasks – in which questions are posed by an examiner (or a computer) – elicit spontaneous responses from candidates. The difficulty level of questions can be graded according to the degree of demands in relation to required cognitive processes such as conceptualisation and lexical and grammatical encoding (Field, 2011). Questions can be formulated to ask about simple, factual information immediately relevant to the candidates or can elicit more extended and coherent responses of an abstract nature. While computer-delivered tests currently use only scripted questions, one of the benefits of this format in face-to-face tests is its authenticity, allowing the interlocutor to employ semi-scripted or unscripted questions to probe the proficiency level of candidates as the interview sequences unfold. However, interlocutor variability has been problematised and extensively researched for this format, suggesting that striking a balance between interlocutor standardisation and authentic interaction is key to the successful use of this task (e.g., Brown, 2003; Lazaraton, 2002; Seedhouse and Nakatsuhara, 2018).

The individual long turn, which requires more organised and coherent speech than in interview or picture description tasks, is usually regarded as the most challenging independent task type. These tasks typically provide candidates with verbal and/or visual prompt(s) based on which extended stretches of speech are elicited. One of the issues in these tasks is the extent to which the topic of the prompt and candidates' background knowledge of the topic affect their performance (Khabbazbashi, 2017). The provision of pre-task planning time is therefore considered valuable to minimise variance related to the conceptualisation demands caused by the topic, though research suggests that the benefit of planning time for enhancing candidates' performances under testing conditions is often mixed or unclear (e.g., Wigglesworth and Elder, 2010).

Integrated tasks

Independent tasks have been criticised for the absence of input and for not allowing candidates an "equal footing in terms of background knowledge" (Weigle, 2004: 30). Consequently, integrated speaking tasks which require candidates "to speak about a topic for which information [has] been supplied from another source" (Jamieson *et al*., 2008: 76) have attracted greater attention in recent years (see Plakans, this volume). Technology has facilitated the use of such task types in computer-mediated tests (e.g., TOEFL iBT, Oxford Test of English) owing to the ease of presenting various modalities in a single test setting.

While recent research has suggested that integrated tasks are "not immune to the influence of prior topical knowledge on scores" (Huang *et al*., 2018: 43) as posited in the literature, various other benefits of integrated tasks have been put forward. Integrated tasks

are considered to better reflect real-life communicative acts, especially in the educational and professional target language use (TLU) domains, and to be tapping into a different construct from independent speaking tasks. Specifically, they not only require the reading and/or listening and speaking skills, but also engage other cognitive skills, such as those for selecting, organising, and transforming source information for the production of output (Barkaoui *et al*., 2012; Brown *et al*., 2005). In the endeavour to better understand the construct of integrated tasks, critical topics for investigation include the balance between input and output skills in terms of source difficulty and factor structures. For example, research has offered interesting insights into the extent to which scores derived in interactive test tasks can be attributed to input skill(s) rather than speaking skills (Sawaki *et al*., 2009), whether or not lower and higher proficiency candidates are differentially impacted by the input task (Ockey, 2018), and the extent to which scores from integrated and independent tasks can be defensively combined for reporting purposes (Lee, 2006).

Interactive tasks

The focus of this task type is the interactional aspects of spoken language where candidates are typically paired or grouped to interact with one other. Interactive tasks gained in popularity in response to concerns raised about the asymmetrical nature of interaction and pre-allocated question-answer turns in traditional examiner-candidate interview tasks (e.g., Johnson, 2001; Seedhouse and Nakatsuhara, 2018). Paired and group tasks enable candidates to demonstrate a wider range of language functions by allowing them to negotiate ongoing interactional organisation (Brooks, 2009; Galaczi, 2014). Fulcher (2003: 189) also notes the potential of these tasks to tap into more cognitively and strategically demanding performances, which are likely to mirror non-test conversations. As a result, these tasks are considered suitable to assess candidates' interactional competence (Galaczi and Taylor, 2018; Plough *et al*., 2018) in addition to more linguistically oriented competence.

However, one long-standing issue with these tasks is the interlocutor effect, which relates to the potential impact on the performance of the person(s) candidates happen to be paired or grouped with. A number of studies have investigated the influence of candidates' own and their partners' characteristics such as personality, proficiency level, L1 background, gender, and acquaintanceship (e.g., Berry, 2007; O'Sullivan, 2008). Findings from these studies are often mixed, suggesting complex interactions among those factors, as well as with external factors such as the size of groups (Nakatsuhara, 2011), making it practically impossible to control for all interlocutor variables. Related concerns include the separability of test scores for individuals involved in the co-construction of interaction (May, 2011) and score reliability due to increased variability associated with co-participants (Bonk and Ockey, 2003; Van Moere, 2006). While these issues might appear to diminish the value of these interactional tasks, Brown (2003: 20) notes that such variation cannot be construct irrelevant, "especially where the construct can be interpreted as encompassing interpersonal communication skills." Nevertheless, attempts should be made to ensure fairness to candidates: for example, by offering multiple test tasks or multiple test occasions (Van Moere, 2006).

Test delivery modes

The construct coverage of a speaking test depends on the types of elicitation tasks as described so far, and computer-mediated tests used to be associated with only certain kinds of task – elicited imitation and fundamentally monologic task types. However, recent

advances in technologies have started widening the range of tasks that computer-mediated speaking tests can deliver. In other words, the affordance of technologies has started showing the potential to expand the range of the construct which computer-mediated speaking tests can tap into (see also Xi, this volume; Sawaki, this volume).

For example, the use of video-conferencing technology to deliver a face-to-face speaking test enables reciprocal interaction without relying on the proximity of speakers. It also reflects the ways in which we communicate nowadays in the social, educational, and professional domains. An increasing body of research has investigated the cognitive, contextual, and scoring validity of video-conferencing-delivered tests, as well as examining their feasibility (e.g., Nakatsuhara *et al*., 2017; Ockey *et al*., 2019; Zhou, 2015). These studies also suggest that the speaking construct under the video-conferencing condition include more explicit meaning negotiation and clearer indication of turn taking than face-to-face communication.

The application of virtual environments has also been explored for the potential of enhancing authenticity of communication. Ockey *et al*. (2017) explore the possibilities offered by virtual environments in providing candidates with an immersive and true-to-life experience that can facilitate the elicitation of collaborative discussions. Furthermore, the use of spoken dialogue systems (SDS) in conjunction with virtual environments seems like a promising avenue. A review of literature on educational games indicates how virtual characters can generate responses and ask follow-up questions while giving some support and feedback to learners (e.g., Morton *et al*., 2012). However, challenges still remain, particularly in terms of high word error rates in L2 speech recognition and the difficulties in predicting all possible L2 utterances even in highly defined contexts. As Litman *et al*. (2018) note, current computer-human interactions are mostly limited to an utterance level and do not generate responses over several turns.

Marking models and rating approaches

We have thus far discussed different ways of eliciting candidates' spoken language. Once elicited, another critical issue relates to how these speech samples are rated. Marking models – the approaches used for assigning scores/ratings to performances – have been shown to affect rater marks (e.g., Barkaoui, 2011) and are therefore an important aspect of a test's validity argument. Hunter *et al*. (1996: 62) refer to a "continuum of scoring practices" with a holistic approach as a "wide-angle assessment of the whole, without clearly delineated criteria" on the one end and atomistic scoring with "microscopic assessment of constituent parts" without an overall perspective on candidate performance on the other (see Galaczi and Lim, this volume; Ockey, this volume for a discussion of methods used for ensuring and evaluating judgment quality or different measurement techniques).

A review of the literature suggests holistic, analytic, and part scoring as the most widely used human-mediated marking models in speaking assessment with performance decision trees (Fulcher *et al*., 2011) and binary-choice scales (O'Grady, 2019) gaining in popularity. Advancements in speech processing and machine learning technologies are also transforming the ways in which speaking is tested, as evidenced in the increasing use of automated approaches to the marking of speaking (Chen *et al*., 2018; Xi *et al*., 2012). Definitions and discussions of the strengths and limitations of these approaches are documented extensively elsewhere (c.f. Davies, 2018; Galaczi and Lim, this volume; Khabbazbashi and Galaczi, 2020; Xi, this volume). Empirical research on how these different approaches compare, however, is rather limited, particularly in speaking assessment. In those cases in which such

comparisons have been made (e.g., Chen *et al*., 2018; Harsch and Martin, 2013; Xi, 2007), there has been a reliance on correlations in reporting empirical relationships. This may disguise the impact of scoring approaches on individual candidate marks, and therefore, it is important to consider alternative ways of comparison with an explicit focus on the "practical significance" (Fulcher, 2003: 65). For example, in their study on the comparison of holistic, analytic, and part marking models in speaking, Khabbazbashi and Galaczi (2020) showed that despite high significant correlations, the choice of marking model had an impact on candidates' final CEFR classifications. An implication of this study relates to the comparison between human and machine marks in automated systems. Automated speaking evaluation technologies largely rely on human-awarded marks as the "gold standard" for training and evaluation of systems (Chen *et al*., 2018; see Fulcher, 2010: 296–97 for a discussion of issues related to using a "gold standard"). Given that those marks are influenced by choice of marking model, by extension, they have an impact on the source data for machine learning and system evaluations. As a result, these should be taken into account and made transparent when reporting human-machine agreement levels.

Main research methods

The previous sections touched upon some of the important parameters in the assessment of speaking; we now focus our attention on the different approaches used for examining these factors. The field of speaking assessment has benefitted from a wide range of research methodologies, including analysis of TLU domains and speaking test corpora (Cushing, this volume), analysis of scoring validity using CTT and IRT (Brown, this volume; Ockey, this volume), and analysis of rater perceptions and rater behaviour (Davis, this volume; Galaczi and Lim, this volume), among others. However, since these research methods are shared with the assessment of other skills, here we examine research methodologies unique to spoken assessment.

Over the past three decades, there has been a proliferation of studies that have analysed speaking test discourse rather than or supplementary to the analysis of scores. Results of such research are valuable sources for the design, description, and most crucially the validation of speaking tests as they allow for an examination of the extent to which the intended construct of a speaking test is actually reflected in elicited performances. We now review three methods of analysis which have been found particularly useful in speaking assessment and consider their strengths and limitations.

Conversation analysis

One of the most popular qualitative analysis methods used to investigate discourse in spoken language tests is conversation analysis (CA) (Sacks *et al*., 1974). Since Lazaraton's (2002) pioneering application of CA techniques on a range of Cambridge examinations in the early 1990s, CA has revealed various interactional features of spoken discourse in interview, paired, and group oral formats (e.g., Brown, 2003; Galaczi, 2014; Lam, 2018; Seedhouse and Nakatsuhara, 2018). These studies have mainly focused on how the interactional organisation reflects the institutional aim of eliciting construct-relevant discourse, to what extent the organisation resembles the CA benchmark of "ordinary" conversation, and in what ways high-scoring and low-scoring candidates demonstrate their interactional competence.

Following the conventions of CA, these studies regard repeated listening/viewing of audio or video recordings for production of transcripts as an important part of analysis for

discovery (Hutchby and Wooffitt, 1998). The analysis is based on analysts' development of "emic" perspectives to understand "why that now?" in interaction (Schegloff and Sacks, 1973: 299). It is worth highlighting that coding and quantification in CA are quite controversial since the quantification of CA data is considered premature when our understanding of the phenomena that we may wish to count is still partial (Schegloff, 1993). However, Heritage (1995: 404) acknowledges that quantification could be successful for well-defined elements in "institutional talk," to which speaking test interactions belong. It is therefore not uncommon to see some sort of quantification in CA research in the field of speaking assessment, in order to supplement purely qualitative CA findings (e.g., Galaczi, 2014; Nakatsuhara, 2011).

Language function analysis

While CA is used to describe interactional organisation, language function analysis is a useful tool to provide summative information of the nature of spoken utterances in relation to cognitive and contextual demands posed by test tasks. For example, language functions observed in relevant TLU domains can inform test and task design. At test development and piloting stages, they can be used to verify or modify draft test specifications and materials. Their analysis can also be used as part of ongoing validation research of operational tests to monitor test version comparability and to ensure that there is sufficient overlap between intended and observed language functions.

For example, O'Sullivan *et al*.'s (2002) function checklist consists of an extensive table of functions categorised under *informational*, *interactional*, and *managing interaction* functions. While originally designed for the analysis of paired spoken discourse in the Cambridge General English examinations, the list has been modified and applied to other speaking tests, such as the IELTS Speaking Test (e.g., Brooks, 2003) and the Aptis Speaking Test (e.g., O'Sullivan and Dunlea, 2020). It has also been utilised to investigate examiners' input language during a speaking test (Nakatsuhara, 2018) and real-life group discussions in class (Ducasse and Brown, 2011).

While the analysis of language functions is useful to generate an overall picture of what candidate language is like, a statistical summary of elicited language functions cannot capture the level of sophistication displayed in realising each language function (Green, 2012). This limitation points to the need for a mixed-methods approach that also involves qualitative description of a sample of actual utterances.

Micro-linguistic analysis

Another useful approach to investigating features of candidate performance is micro-linguistic analysis. This is commonly conducted in studies looking into either scoring validity (e.g., seeking to validate "subjective" human ratings) or refining the rating scales by investigating the values from the "objective" measures that represent the ratings (e.g., Hsieh and Wang, 2017; Tavakoli *et al*., 2020) or context validity (e.g., evaluating test administration conditions such as the provision of planning time; see Iwashita *et al*., 2001; O'Grady, 2019; Wigglesworth and Elder, 2010).

Such micro-linguistic analyses often involve calculating variables of complexity – both lexical and syntactic – accuracy, and fluency (CAF), derived from studies in task-based research, and some of the analyses can now be facilitated or partly automated by software such as Praat (Boersma and Weenink, 2016). The discussion of variables of lexical

complexity can be found in Pallotti (2021), including vocd-D value and the percentages of words found in certain vocabulary lists. For syntactic complexity, there are different ways in which candidates might syntactically complexify their performance (Norris and Ortega, 2009), including the number of words per syntactically defined unit and the amounts of coordination and subordination. For accuracy, commonly used variables include the percentage of error-free clauses and weighted errors per clause ratio (Foster and Wigglesworth, 2016). Regarding fluency, de Jong (2016) provides a comprehensive summary of the three types: speed fluency (e.g., speech rate), breakdown fluency (e.g., the number and length of silent pauses over total speaking time), and repair fluency (e.g., the number of repetitions, false starts, and reformulations). For integrated tasks, the successful reproduction of the input text is crucial for task achievement, and therefore, some studies have compared the idea units in the input and candidate performances (e.g., Brown *et al*., 2005; Frost *et al*., 2012).

While there are criticisms against the use of such micro-linguistic analyses on the grounds that the variables may be too crude and that low-inference analytic categories may fail to capture the reality of human communication (Fulcher, 2015b), a number of studies have found fair to good agreement between the objective variables and subjective human ratings (e.g., Inoue, 2016). Such approaches can therefore offer useful recommendations to enhance the contextual and scoring validity of tests.

Future directions

In this chapter, we first looked back at the history of speaking assessment and subsequently considered where we are now in terms of critical parameters in the testing of speaking in the digital age. One striking observation is the role that technology has played in shaping the speaking construct – for example, in narrowing the construct where there is an overreliance on elicited imitation tasks and in broadening the construct in the case of integrated listening/reading into speaking tasks. In ASE systems, technology has removed human interaction from the equation, whereas in video-conferencing delivery and virtual environments, it has facilitated interaction in innovative ways. As Galaczi (2010: 47) argues, there is no "best way" to test speaking, and the important criterion is to evaluate different approaches in light of "fitness for purpose." In looking into the future, it is hard to predict in what ways technology might influence speaking assessment. What is important, however, is to always go back to the construct and ensure that it is technology at the service of the construct and not the other way around.

In gazing into our crystal ball, here are some questions we can ask: What might the speaking construct look like in the future? Are current models of communicative competence adequately capturing the complexities of language use? Elder *et al*. (2017: 15), for example, criticise the overreliance of existing models on linguistic features at the expense of other, more complex "non-linguistic cognitive, affective, and volitional" factors involved in interaction. But how can we, as language testers, draw a line between what is construct relevant and construct irrelevant? In an era when technology affords "seamless" blending of visual and text-based input and between written and spoken modes of delivery, and hybrid spoken/written discourse is common, is it still defensible to test the four skills independently (c.f. scenario-based assessment; Purpura, 2016)? What would the construct of L2 speaking look like in the future when in-ear translations may become a reality? As Litman *et al*. (2018: 305) speculate, will the day "come when what we want to assess is the ability to communicate both with humans and with machines"? With the increasing use of ASE systems, will there be a new set of ethical questions? For instance, can machines be the ultimate

arbiter of whether or not L2 speakers can study or live abroad? What about accountability? Who is responsible, should anything go wrong? Can hybrid human-machine approaches to scoring as proposed by Isaacs (2018) and de Jong (2018) be used to address such ethical concerns as well as making the best use of what ASE systems and human raters can offer?

The answers to these questions are unforeseeable. It is essential, however, that we consider all newly emerging questions in light of "the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other models of assessment" (Messick, 1989: 13). To this end, we should continuously seek to understand the ever-changing TLU domains and what we mean by the term "speaking".

Further reading

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Longman, Pearson Education.
This book, although written some time ago, provides a thorough guide to designing and implementing speaking tests, as well as useful and critical summaries of research in L2 speaking assessment.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching* 48: 198–216.
This article illustrates a timeline of L2 speaking assessment research since the mid-nineteenth century by summarising publications which advanced our understanding of 12 selected themes in the field.

Lim, G. (ed.). (2018). Conceptualizing and operationalizing speaking assessment for a new century [special issue]. *Language Assessment Quarterly* 15.
The articles in this special issue consider important aspects of the speaking construct, such as interactional competence, collocational competence, fluency, and pronunciation and whether and to what extent they have been assessed while reflecting on the potential role of technology in enhancing assessment practices.

Taylor, L. (2011). *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Cambridge: UCLES, Cambridge University Press.
This edited volume presents a review of relevant literature on the assessment of speaking and provides a systematic framework for validating speaking tests.

References

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance: Assessment in Education*: Principles, Policy & Practice* 18: 279–93. https://doi.org/10.1080/0969594X.2010.526585.

Barkaoui, K., Brooks, L., Swain, M. and Lapkin, S. (2012). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics* 34: 304–24. https://doi.org/10.1093/applin/ams046.

Berry, V. (2007). *Personality Differences and Oral Test Performance*. Frankfurt: Peter Lang.

Bley-Vroman, R. and Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In A. Mackey and S. Gass (eds.), *Research Methodology in Second-Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum, 245–61.

Boersma, P. and Weenink, D. (2016). *Praat: Doing Phonetics by Computer* [Computer software]. Retrieved from www.PRAAT.org/.

Bonk, W. J. and Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing* 20: 89–110. https://doi.org/10.1080/15434303.2016.1236797.

Brooks, L. (2003). Converting an observation checklist for use with the IELTS speaking test. *Research Notes* 11: 20–21.

———. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing* 26: 341–66. https://doi.org/10.1177/0265532209104666.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20: 1–25. https://doi.org/10.1191/0265532203lt242oa.

Brown, A., Iwashita, N. and McNamara, T. (2005). *An Examination of Rater Orientations and Test-Taker Performance on English-for-Academic-Purposes Speaking Tasks*. TOEFL Monograph No. TOEFL-MS-29. Retrieved from www.ets.org/Media/Research/pdf/RR-05-05.pdf.

Chapelle, C., Enright, M. and Jamieson, J. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. London: Routledge.

Chen, L., Zechner, K., Yoon, S., Evanini, K., Wang, X., Loukina, A. . . . Ma, M. (2018). Automated scoring of nonnative speech using the SpeechRater SM v. 5.0 engine. *ETS Research Report Series, ETS RR*: 18–10. https://doi.org/10.1002/ets2.12198.

Davies, L. (2018). Analytic, holistic, and primary trait marking scales. In J. I. Liontas, M. DelliCarpini and TESOL International Association (eds.), *The TESOL Encyclopaedia of English Language Teaching*, 1–6. https://doi.org/10.1002/9781118784235.eelt0365.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching* 54: 113–32. https://doi.org/10.1515/iral-2016-9993.

———. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly* 15: 237–54. https://doi.org/10.1080/15434303.2018.1477780.

Ducasse, A. M. and Brown, A. (2011). The role of interactive communication in IELTS Speaking and its relationship to candidates' preparedness for study or training contexts. *IELTS Research Reports* 12. Retrieved from www.ielts.org/-/media/research-reports/ielts-rr-volume-12-report-3.ashx.

Elder, C., McNamara, T., Kim, H., Pill, J. and Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the nonlanguage specialist can tell us. *Language & Communication* 57: 14–21. https://doi.org/10.1016/j.langcom.2016.12.005.

Field, J. (2011). Cognitive validity. In L. Taylor (ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Cambridge: UCLES, Cambridge University Press, 65–111.

Foster, P. and Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics* 36: 98–116. https://doi.org/10.1017/S0267190515000082.

Frost, K., Elder, C. and Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing* 29: 345–69. https://doi.org/10.1177/0265532211424479.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13: 208–38. https://doi.org/10.1177/026553229601300205.

———. (2003). *Testing Second Language Speaking*. London: Longman, Pearson Education.

———. (2010). *Practical Language Testing*. London: Hodder Education.

———. (2015a). Assessing second language speaking. *Language Teaching* 48: 198–216.

———. (2015b). *Re-Examining Language Testing: A Philosophical and Social Inquiry*. London and New York: Routledge.

Fulcher, G., Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing* 28: 5–29. https://doi.org/10.1177/0265532209359514.

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (ed.), *Computer-Based Assessment of Foreign Language Speaking Skills*. Luxembourg, LU: Publications Office of the European Union, 29–51. Retrieved from www.testdaf.de/fileadmin/Redakteur/PDF/Forschung-Publikationen/Volume_European_Commission_2010.pdf.

———. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired tests? *Applied Linguistics* 35: 553–74. https://doi.org/10.1093/applin/amt017.

Galaczi, E. D. and Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly* 15: 219–36. https://doi.org/10.1080/15434303.2018.1453816.

Gathercole, S. E. and Baddeley, A. D. (1993). Phonological working memory: A critical building block for reading development and vocabulary acquisition? *European Journal of Psychology of Education* 8: 259–72.

Green, A. (2012). *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*. Cambridge: UCLES, Cambridge University Press.

Harsch, C. and Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice* 20: 281–307. https://doi.org/10.1080/0969594X.2012.742422.

Heritage, J. (1995). Conversation analysis: methodological aspects. In U. M. Quasthoff (ed.), *Aspects of Oral Communication*. Berlin: Water de Gruyter, 391–418.

Hsieh, C. and Wang, Y. (2017). Speaking proficiency of young language students: A discourse-analytic study. *Language Testing* 36: 27–50. https://doi.org/10.1177/0265532217734240.

Hutchby, I. and Wooffitt, R. (1998). *Conversation Analysis*. Cambridge: Cambridge University Press.

Huang, H. T. D., Hung, S. T. A. and Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing* 35: 27–49. https://doi.org/10.1177/0265532216677106.

Hunter, D. M., Jones, R. M. and Randhawa, B. S. (1996). The use of holistic versus analytic scoring for large-scale assessment of writing. *The Canadian Journal of Program Evaluation* 11: 61–85.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal* 44: 487–505. https://doi.org/10.1080/09571736.2015.1130079.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly* 15: 273–93. https://doi.org/10.1080/15434303.2018.1472264.

Iwashita, N., McNamara, T. and Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to test design. *Language Learning* 51: 401–36. https://doi.org/10.1111/0023-8333.00160.

Jamieson, J., Eignor, D., Grabe, W. and Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. A. Chapelle, M. K. Enright and J. M. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY: Routledge, 55–95.

Johnson, M. (2001). *The Art of Non-Conversation: A Re-Examination of the Validity of the Oral Proficiency Interview*. New Haven, CT and London: Yale University Press.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing* 34: 23–48. https://doi.org/10.1177/0265532215595666.

Khabbazbashi, N. and Galaczi, E. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment *Language Testing* 37: 333–60. https://doi.org/10.1177/0265532219898635.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests. A Teacher's Book*. London: Longman.

Lam, D. M. K. (2018). What counts as "responding"? Contingency on previous speaker contribution as a feature of interactional competence. *Language Testing* 35: 377–401. https://doi.org/10.1177/0265532218758126.

Latham, H. (1877). *On the Action of Examinations Considered as a Means of Selection*. Cambridge: Dighton, Bell and Company.

Lazaraton, A. (2002). *A Qualitative Approach to the Validation of Oral Language Tests*. Cambridge: UCLES, Cambridge University Press.

Lee, Y. W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing* 23: 131–66. https://doi.org/10.1191/0265532206lt325oa.

Litman, D., Strik, H. and Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities, *Language Assessment Quarterly* 15: 294–309. https://doi.org/10.1080/15434303.2018.1472265.

May, L. (2011). *Interaction in a Paired Speaking Test: The Rater's Perspective*. Frankfurt: Peter Lang.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement* (3rd edition). London and New York, NY: Palgrave Macmillan.

Morton, H., Gunson, N. and Jack, M. (2012). Interactive language learning through speech-enabled virtual scenarios. *Advances in Human-Computer Interaction* 2012: 1–14. https://doi.org/10.1155/2012/389523.

Nakatsuhara, F. (2011). Effects of the number of participants on group oral test performance. *Language Testing* 28: 483–508. https://doi.org/10.1177%2F0265532211398110.

———. (2018). Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests. In G. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam and Philadelphia, PA: John Benjamins, 205–25.

Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly* 14: 1–18. https://doi.org/10.1080/15434303.2016.1263637.

Norris, J. M. and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30: 555–78. https://doi.org/10.1093/applin/amp044.

North, B. and Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing* 15: 217–62. https://doi.org/10.1177/026553229801500204.

Ockey, G. J. (2018). The degree to which it maters if an oral test tasks require listening. In G. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam and Philadelphia, PA: John Benjamins, 193–204.

Ockey, G. J., Gu, L. and Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly* 14: 346–59. https://doi.org/10.1080/15434303.2017.1400036.

Ockey, G., Timpe-Laughlin, V., Davis, L. and Gu, L. (2019). Exploring the potential of a video-mediated interactive speaking assessment. *ETS Research Report Series, ETS RR – 19–05*. https://doi.org/10.1002/ets2.12240.

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing* 36: 505–26. https://doi.org/10.1177/0265532219826604.

O'Sullivan, B. (2008). *Modelling Performance in Tests of Spoken Language*. Frankfurt: Peter Lang.

O'Sullivan, B. and Dunlea, J. (2020). *Aptis General technical manual Ver 2.1*. Technical Reports, TR/2020/001. Retrieved from www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.1.pdf.

O'Sullivan, B., Weir, C. J. and Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing* 19: 33–56. https://doi.org/10.1191/0265532202lt219oa.

Pallotti, G. (2021). Measuring complexity, accuracy and fluency (CAF). In P. Winke and T. Brunfaut (eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing*. New York: Routledge, 201–10.

Palmer, H. E. (1921). *The Oral Method of Teaching Languages*. Cambridge: Heffers.

Plough, I., Banerjee, J. and Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing* 35: 427–45. https://doi:10.1177/0265532218772325.

Pollitt, A. and Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic and N. Saville (eds.), *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium*. Cambridge: UCLES, Cambridge University Press, 74–91.

Purpura, J. E. (2016). Second and foreign language assessment. *The Modern Language Journal* 100(Supplement): 190–208. https://doi.org/10.1111/modl.12308.

Russell, D. R. (2002). *Writing in the Academic Disciplines: A Curricular History,* 2nd edn. Carbondale, IL: Southern Illinois University Press.

Sacks, H., Schegloff, E. and Jefferson, G. (1974). The simplest systematics for the organization of turn-taking for conversation. *Language* 50: 696–735. https://doi.org/10.2307/412243.

Sawaki, Y., Stricker, L. J. and Oranje, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing* 26: 5–30. https://doi.org/10.1177/0265532208097335.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction* 26: 99–128. https://doi.org/10.1207/s15327973rlsi2601_5.

Schegloff, E. A. and Sacks, H. (1973). Opening up closings. *Semiotica* 8: 289–327. http://doi.org/10.1515/semi.1973.8.4.289.

Seedhouse, P. and Nakatsuhara, F. (2018). *The Discourse of the IELTS Speaking Test: The Institutional Design of Spoken Interaction for Language Assessment*. Cambridge: Cambridge University Press.

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (ed.), *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service, 89–103.

Sweet, H. (1899). *The Practical Study of Languages*. London: Dent.

Tavakoli, P., Nakatsuhara, F. and Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal* 104: 169–91. https://doi.org/10.1111/modl.12620.

Trinity College London. (2021). *ISE Rating Scales*. www.trinitycollege.com/qualifications/english-language/ISE/ISE-results-and-certificates/ISE-rating-scales

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing* 23: 411–40. https://doi.org/10.1191/0265532206lt336oa.

———. (2012). A psycholinguistic approach to oral language assessment. *Language Testing* 29: 325–44. https://doi.org/10.1177/0265532211424478.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing* 9: 27–55. https://doi.org/10.1016/j.asw.2004.01.002.

Weir, C. J., Vidaković, I. and Galaczi, E. D. (2013). *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*. Cambridge: UCLES, Cambridge University Press.

Wigglesworth, G. and Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly* 7: 1–24. https://doi.org/10.1080/15434300903031779.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test for operational use. *Language Testing* 24: 251–86. https://doi.org/10.1177/0265532207076365.

Xi, X., Higgins, D., Zechner, K. and Williamson, D. (2012). A comparison of two scoring methods for an automated speech scoring system. *Language Testing* 29: 371–94. https://doi.org/10.1177%2F0265532211425673.

Yan, X., Maeda, Y., Lv, J. and Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing* 33: 497–528. https://doi.org/10.1177%2F0265532215594643.

Zhou, Y. (2015). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. *Language Testing in Asia* 5, Article number 2: 1–16. https://doi.org/10.1186/s40468-014-0012-y.

13
Assessing listening

Elvis Wagner

Introduction

The ability to understand spoken language is obviously a vital component of the construct of second language (L2) proficiency, and thus it must be included in any assessment purporting to assess an L2 test taker's communicative competence. The assessment of L2 listening ability, however, is more challenging in many ways than the assessment of the other language skills, and major questions and disagreements exist regarding how L2 listening ability can best be assessed. This chapter provides an overview of a number of current critical issues in assessing L2 listening (i.e., interactive speaking and listening ability, the use of audio-visual texts, the use of real-world spoken language, spoken texts with different accent varieties, listening test washback) and recommends that test developers assess L2 listeners' communicative competence by creating test tasks similar to and representative of the types of L2 listening tasks that are found in real-world listening contexts.

Historical perspectives

Buck (2001) provides a succinct historical overview of L2 listening testing. According to Buck, there are three main approaches to testing second language ability: discrete-point, integrative, and communicative. The discrete-point approach is associated with audio-lingualism and behaviorism, and Lado (1961) is the language tester most associated with this approach. Discrete-point testing focuses on the basic elements of language. The units of linguistic knowledge are identified, and then a representative sample of these elements are assessed in order to get a measurement of a test taker's overall language ability. For listening, then, the types of tasks might include individual multiple-choice or true/false questions that assess finite bits of language at the sentence level but may not reach discourse levels of comprehension. Much of the focus is on decoding, and so tasks that involve phonemic discrimination and minimal pairs are common (Buck, 2001).

Oller (1979) criticized this idea of discrete-point testing and advocated for integrative testing, arguing that while "discrete items attempt to test knowledge of language one bit

DOI: 10.4324/9781003220756-18

at a time, integrative tests attempt to assess a learner's capacity to use many bits at the same time" (p. 37). Instead of focusing on short, unrelated bits of texts, integrative testing of listening ability focuses on longer discourse-level texts, with the goal of assessing the processing of spoken language, rather than focusing on assessing knowledge of language. With integrative testing of listening, typical tasks include listening cloze, dictation, sentence repetition, and even translation (Buck, 2001).

Buck (2001) states that the communicative language approach that emerged in the 1970s greatly influenced L2 testing, especially the notion of communicative competence. As a result, language test developers started to focus on a test taker's ability to use the target language to communicate, rather than trying to measure how much that test taker knew about the target language. With L2 listening, then, the focus shifted to assessing the listener's ability to understand the communicative intent of spoken texts and the ability to use language for a communicative purpose (Buck, 2001). This was accomplished by providing more authentic texts, providing a communicative purpose in the testing situation, and creating test tasks that were more similar to real-world listening tasks (although Buck stresses that defining and operationalizing *communicative* and *authentic* is very difficult to do within the constraints of a listening test).

In the new millennium, the trend in communicative tests of L2 listening seems to have evolved to focus on the use of more integrated tasks (see Plakans, this volume). This is similar to yet different from Oller's idea of "integrative" testing. The idea of integrated tasks focuses more on the ability to use language in multiple modes within a communicative context. With communicative language teaching, the four-skills approach to teaching language came to be seen as artificial and unnecessarily constraining (Savignon, 2018), with the realization that communicative competence in a language often involves multiple skills: that people read something in order to write about it or listen to a text in order to respond orally or in writing. Similarly, language testers realized that a four-skills approach to testing was constraining and not representative of real-life language use, and so tasks that required integrating language skills became more common: "Treating language as holistic or as an amalgamation of reading, writing, listening, or speaking, rather than as isolated abilities may allow for different or more substantive interpretations of performances that are transferrable to real world domains" (Plakans, 2012: 249). The TOEFL iBT is a good example of the increased use of integrated test tasks involving L2 listening ability. The current version of the TOEFL iBT has a number of integrated tasks (e.g., read a passage, listen to a recording of a spoken text, write a response; listen to a recording of a spoken text, then give an oral response). A special type of integrated task involving listening is interactive speaking and listening, which is described later in this chapter.

Critical issues and topics

Within the field of L2 assessment, it seems to be broadly recognized that the idea of communicative competence is paramount and that it is necessary to attempt to assess a learner's ability to communicate in the target language within a particular domain (with that domain being specified by the test developer). That is, rather than trying to assess a learner's knowledge about the language, assessments should focus instead on assessing the learner's ability to use the target language to communicate with others in that specified domain (target language use, or TLU, domain). In L2 listening assessment, then, the goal of an assessment is to measure a test taker's ability to process and comprehend spoken language that is representative of the spoken language found in the TLU domain (Buck, 2001; Wagner, 2014a; see

also Ross, this volume). Doing this should minimize what Messick (1989, 1996) described as the two main threats to construct validity: construct-irrelevant variance and construct underrepresentation.

Interactive speaking and listening

When defining and operationalizing L2 listening ability, it is important to consider the different types of listening that an L2 learner must be able to do. In some settings (i.e., academic contexts), the ability to understand one-way speaking is very important. For example, the ability to understand an academic lecture, in which one person does virtually all the speaking, is obviously an important ability in an academic TLU domain. But in the vast majority of real-world contexts (and TLU domains), the ability to understand spoken language entails interactive speaking and listening ability. That is, interactional/conversational spoken language involves listening to another speaker and responding orally.

This interactive speaking and listening is in some ways easier than one-way listening, because (usually) the speaker accommodates to the listener; if the speaker notices that the listener has difficulty understanding what the speaker has said, the speaker will repeat or rephrase the message in an attempt to make it comprehensible. In many one-way listening contexts, however (e.g., movies or television shows, academic lectures), speakers are usually less attuned or not able to note listeners' difficulty, and thus they are less likely to accommodate their spoken messages. However, interactive speaking and listening is also very cognitively demanding, in that the L2 listeners must focus on the spoken message, interpret it, and reply appropriately, with minimal time allowed to process the spoken input and to formulate the spoken output. In other words, the listening involved in interactive speaking and listening requires different skills and abilities than one-way listening. If interactive speaking and listening ability is prominent in the TLU domain of interest, it would seem vital to include interactive speaking/listening test tasks because using only one-way listening test tasks would result in construct underrepresentation. Yet many or most of the listening test tasks traditionally employed on L2 listening tests emphasize monologic listening ability because (1) the "four-skills" approach is more familiar and easier to assess, and (2) the interactive approach is more difficult to standardize and score.

The four-skills approach to language testing, in which each of the traditional four language skills (i.e., reading, writing, listening, and speaking) is tested separately, continues to be prevalent (Ockey and Wagner, 2018a). There are obvious advantages to such an approach because each of the skills can be isolated and assessed, and a separate score can be given for each skill. And even with integrated test tasks, it is still possible to provide separate skill scores, even if the skills are ostensibly integrated in the test task. Using truly interactive speaking and listening test tasks presents difficulties for test developers because of the impossibility in fully standardizing the procedure. Test developers have to be willing to sacrifice some degree of standardization in the hopes of making more valid inferences about test takers' interactive speaking and listening ability. In addition, because interactive speaking and listening tasks usually require rating scales and expert judgment, the reliability of the scoring might be lower than in monologic listening test tasks with right/wrong answers and objective scoring (Ockey and Wagner, 2018a). Nevertheless, real-life communicative language use entails interaction, and thus L2 assessments that purport to assess communicative language ability should include test tasks that involve interactive listening and speaking ability. Indeed, Douglas (1997) argued that listening and speaking should be integrated in assessment because they "are theoretically and practically very difficult to separate" (p. 25).

Different aspects of assessing interactive speaking and listening ability have been researched extensively in the literature: the use of different scales or subscales to measure oral interactional ability, including turn taking, negotiating meaning, response appropriateness, and topic initiation (e.g., Bonk and Ockey, 2003; Ockey *et al*., 2015; Taylor and Galaczi, 2011); the extent to which listening ability contributes to scores on interactive listening and speaking tasks (e.g., Brooks, 2009; Choi and So, 2018; Galaczi, 2014; Ockey, 2014); and the comparison of one-way listening test scores with interactive listening and speaking test scores (e.g., Hildsdon, 1995; Nakatsuhara, 2012; Ockey, 2018). This research has demonstrated that, despite the challenges involved in implementation, assessing listening through the use of interactive speaking and listening test tasks is viable and effective and, indeed, is necessary if the goal is to truly assess a test taker's communicative competence.

The use of audio-visual texts

Listening ability involves the ability to understand spoken (aural) input. But listeners are also able to utilize nonverbal information transmitted by the speaker to comprehend the aural input. The issue for language test developers, then, is whether to include the nonverbal information in the spoken input used on L2 listening tests by using audio-visual texts or to restrict the verbal input by using audio-only texts. This issue has been researched extensively, especially the question of whether using audio-visual texts leads to increased test performance compared to using audio-only texts. Most of the research results indicate that the use of audio-visual texts results in statistically significantly higher scores compared with audio-only texts (e.g., Batty, 2018; Sueyoshi and Hardison, 2005; Wagner, 2010, 2013), although some studies have found no statistically significant difference (e.g., Batty, 2015; Coniam, 2001; Gruba, 1993) or even an advantage for audio-only input (Suvorov, 2009). This research has also provided clues as to why the audio-visual texts might lead to increased performance (e.g., lip movements, facial expressions, gestures, visual contextual information), why the nonverbal components of the audio-visual texts might be distracting for test takers, and what test takers focus on while watching the video (Batty, OnlineFirst).

While these studies are important in examining how test takers' performance might differ under the different testing conditions, it is also vital to consider the target language use domain of the construct being measured. There are, of course, some real-life contexts in which the spoken input that a listener hears is audio only (e.g., talking on the phone, listening to the radio), but in most real-world L2 listening contexts, the listener is able to see the speaker and thus able to utilize the nonverbal information provided by the speaker and the context of the speaking situation. Research shows that some listeners are better able than others to utilize this nonverbal information to assist them in comprehending the aural input (e.g., Burgoon, 1994; Cross, 2011; Sueyoshi and Hardison, 2005; Wagner, 2008, 2010). Wagner (2008, 2010, 2014a) has argued repeatedly that

> if the ability to utilize the nonverbal aspects of spoken language is part of the TLU domain, then it should be part of the construct definition of L2 listening ability and should be operationalized on tests that purport to measure test takers' communicative competence.
>
> *(Wagner and Ockey, 2018: 144)*

Yet many or most high-stakes L2 listening tests continue to use audio-only input. Wagner and Ockey (2018) described how three of the most prominent English proficiency tests (the

IELTS, PTE Academic, and TOEFL iBT) almost exclusively used audio-only input, and even the few times that visual information was included in the input, it usually consisted of static visual input that served as context visuals (e.g., charts, diagrams, or pictures), rather than dynamic video input. The theoretical justification for using audio-visual input on L2 listening tests is strong, yet its use is surprisingly limited, most likely because using audio-visual input requires more resources to produce and present to test takers (Wagner, 2014a).

The use of real-world spoken language

The characteristics of spoken language vary greatly based on the context of the speaking situation (e.g., informal versus formal setting, familiarity versus unfamiliarity between speakers, age, gender, education level of speakers, etc.). Researchers have described different continua that represent the different varieties that are possible, focusing on how written language (that is then read aloud) differs from spoken language. For example, Tannen (1982) describes a continuum of orality, with literate texts at one end and purely oral texts at the other end. Similarly, McCarthy and Carter (1995) describe how spoken texts can range from "speakerly" to "writerly" (p. 216). Wagner (2016, 2018) writes about a continuum of scriptedness, describing how spoken texts can vary according to the level of scriptedness or planning involved. At one end of the continuum are texts that are composed and written, revised, and edited, then read aloud. At the other end of the continuum are texts that are composed and uttered almost simultaneously. Because of the planning time involved, the characteristics of texts at opposite ends of the continuum are quite different. Texts that are extensively planned and edited and then read aloud tend to have written grammatical norms, with more complex syntax and a more straightforward and linear organizational scheme (Chafe, 1985). In contrast, texts that are unplanned tend to have oral grammatical norms (Carter and McCarthy, 2006) and extensive hesitation phenomena (i.e., silent and filled pauses, repetitions, and false starts) (Fox Tree, 1995). Unscripted texts tend to be less linearly organized (because time constraints minimize the speaker's ability to plan how to organize their utterance) and have more instances of connected speech (Chafe, 1985).

L2 listeners have to be able to understand many different types of spoken language. Test developers need to identify the target language-use domain for their particular test, identify the types of spoken texts that are commonly found in that domain, and then use spoken texts on their test that are representative of the types of spoken texts found in that particular TLU domain (Ockey and Wagner, 2018b). But many high-stakes English proficiency tests use almost exclusively spoken texts that are from the scripted end of the scriptedness continuum. Wagner (2016) reviewed four English proficiency tests (i.e., IELTS, MELAB, PTE Academic, and TOEFL iBT) commonly used for English-medium university admissions and found that almost all the spoken texts used as prompts on the listening sections were scripted. (The only exception was that some of the texts used for the PTE Academic seemed to be unscripted.) Similarly, Wagner and Wagner (2016) reviewed three prominent English proficiency tests from Asia (the CET from China, the EIKEN from Japan, and the GEPT from Taiwan) and found that of the seven different listening tasks on these tests, only one of the tasks (the two-person interview task from the EIKEN) seemed to use unscripted spoken texts. Yanagawa (2016) found similar results for the Center Listening Test (CLT) in Japan. Field (2013) examined the spoken texts used for the listening tasks of the Cambridge ESOL suite of tests and found that the texts lacked the characteristics of unplanned real-world spoken language. Field was especially critical of the fact that the writers of the test appeared to use written texts as the source of the spoken texts.

It is unclear why test developers are reluctant to use unscripted spoken texts on their tests and instead use scripted spoken texts almost exclusively. There does not seem to be a construct-based theoretical rationale for doing so. (It is difficult to imagine a TLU domain that encompasses being solely able to comprehend scripted spoken language.) Rather, this overreliance on scripted spoken texts seems to be a function of relying on the tried and true (L2 listening tests have traditionally used scripted spoken texts) and the fact that L2 test developers can create scripted texts to specifically address predetermined test specifications (e.g., length of text, number of comprehension questions, topic, etc.) (Wagner, 2016, 2018). As Buck (2001) describes it, "[S]peakers need to speak spontaneously to produce realistic spoken language, but test-developers need to control what the speaker says in order to make test items" (p. 87). However, the shortcoming of this sort of approach is that using scripted texts exclusively results in a very narrow operationalization of L2 listening ability. L2 listeners vary in their ability to comprehend unscripted spoken language, and this ability is certainly construct relevant (Wagner, 2014a, 2016, 2018). To not include unscripted spoken texts on L2 listening tests can result in threats to validity due to construct underrepresentation (Messick, 1989, 1996).

Spoken texts with different accent varieties

L2 listening comprehension can be affected by the accent of the speaker. The research has shown that familiarity with a particular accent variety can facilitate comprehension (Gass and Varonis, 1984) and that L2 listeners are able to adapt to adjust their listening behavior to understand an unfamiliar accent (Bradlow and Bent, 2008; Gass and Varonis, 1984; Harding, 2018). The issue of accent variety is an important consideration for L2 listening test developers because they need to consider what accent or variety to use on their tests. The obvious answer would seem to be to use the "standard" variety of the language where the test is taking place, but in an age of globalization and multiculturalism, choosing the most appropriate accent varieties to use on an L2 listening test can be challenging. English provides a good example of this challenge. For a language proficiency test in the UK, for instance, one could choose to use "standard British." But "standard" is a nebulous term and depends on one's personal beliefs and perspective (Milroy and Milroy, 1999). In addition, it seems unlikely that an L2 learner would only encounter standard spoken British outside the classroom or test; much more likely is that this learner would encounter a variety of different regional and socio-economic accents, including L2-accented English.

Some large-scale English proficiency tests such as the TOEFL iBT and IELTS use a variety of geographical accents (e.g., British, American, Australian, New Zealand, etc.), yet test developers seem reluctant to use non-prestige varieties (e.g., African-American Vernacular English, Southern American, Liverpudlian English, etc.). L2 listening ability includes being able to understand all sorts of accent varieties, not just the dominant/standard/prestige variety. Similarly, the idea of using L2-accented English for spoken texts on listening assessments has not been widely adopted, even though there are more L2 English speakers in the world than there are L1 English speakers. Again, L2 listeners need to be able to understand different speech varieties, including different varieties of L2-accented English, because being able to accommodate to unfamiliar speech varieties is necessary for communicative competence (Canagarajah, 2006).

Test washback and the teaching and learning of L2 listening

The issue of test washback and impact has garnered a large amount of research in recent decades (Wall, 2012; Cheng and Sultana, this volume), and justifiably so, as there is an increasing recognition that high-stakes assessments can have a real impact (both positive and negative) not only on learners but also on teachers, materials developers, curriculum developers, and national educational systems (Bailey, 1996). For whatever reason, however, the impact of L2 listening assessment has not been the focus of this extensive washback literature. Indeed, a review of the ILTA 2018 language testing bibliography indicated that the "Washback" category had 115 articles, chapters, or books dedicated to language test washback, but none of these 115 works had "listening" in its title. This is unfortunate because how L2 listening is tested can have a direct effect on how it is taught. Wall (2012) cautions that the research shows "that the introduction or revision of a test will not automatically cause changes to occur in teaching and learning" (p. 84) and that it is difficult to predict concretely how a particular test or components of a test will influence teaching and learning. Nevertheless, if a high-stakes, influential test includes a particular skill or task in its test tasks, then it seems more likely that learning systems (i.e., teachers, curriculum developers, materials developers) will address that skill or task in the classroom, and learners will be more likely to prioritize that skill or task in their learning. All the critical issues and topics described earlier (assessing listening as part of interactive speaking and listening, the use of audio-visual texts, the use of real-world spoken language, spoken texts with different accent varieties) are issues that are relevant to the teaching of L2 listening as well and have been suggested as issues that L2 listening teachers should address in the classroom (Field, 2008; Rost, 2015; Wagner, 2014b). At the very least, test developers have to consider how the inclusion or exclusion of different aspects of L2 listening ability on their tests can impact how L2 listening is taught.

Main research methods

L2 listening ability was sometimes referred to as "the forgotten skill," in reference to the lack of assessment research (Douglas, 1988). However, in the last 20 years, it has been the focus of a great deal of research. Perhaps the most prevalent research method associated with L2 listening research uses classical test theory, especially in attempts to build construct validity arguments involving intervention or differential groups studies (Brown, 2012). An example of this type of study is Wagner (2010), who investigated how the use of videotexts affected test taker performance. One group took a listening test in which the spoken input was delivered with an audio-visual text while a comparable ability group took the same listening test but with the spoken input delivered via audio only. Wagner found that the group that had the audio-visual texts scored statistically significantly higher than the audio-only group. Studies like this are helpful in examining how manipulating a particular variable affects test taker performance and can be useful in building a validity argument for a test.

Another research method commonly found in L2 listening assessment research involves using regression analysis to investigate how varying components of the test format, item types, and variables in the test-taking process (e.g., item types, text characteristics, length of text, hesitation phenomena, etc.) can effect or predict test taker performance. Brunfaut (2016) provides an extensive list of factors that have been researched that can affect listening test

task difficulty, including task-related factors (i.e., the listening input, the listening test task procedures, the interactions between the task input and output) and listener-related factors (cognitive characteristics of the test takers, linguistic characteristics of the test takers, affective characteristics of the test takers). Two example studies in this vein are Freedle and Kostin (1999) and Revesz and Brunfaut (2013). Freedle and Kostin (1999) examined the listening section of the TOEFL and came up with a list of variables that might affect test taker performance, including 11 item variables, 31 text variables, 39 text/item overlap variables, and 4 item-type variables. They performed a hierarchical regression and found that while item type variables accounted for 7.8 percent of the variance, the text and item variables accounted for a much larger percentage (37.5 percent) of the variance. Freedle and Kostin concluded that this presents evidence in support of the construct validity of the test. Similarly, Revesz and Brunfaut (2013) used regression analysis to investigate the effect of a number of listening task factors (e.g., speed of delivery, linguistic complexity, explicitness, etc.) on task difficulty. They found five significant predictors of task difficulty: proportion of function words, incidence of academic words, lexical diversity, lexical density, and causal content.

Item response theory (IRT) has also been used as a research method in L2 listening assessment research, in an attempt to examine how well different types of listening tests can link observable performances on the listening test to their underlying listening ability (Ockey, this volume). An example of a L2 listening test study using IRT is Batty (2018). In this study, Batty used many-facet Rasch analysis to examine how the use of audio-visual and audio-only input affected test taker performance on an L2 listening test. The test had two different types of items: explicit items (in which the answer was stated in the spoken text) and implicit items (in which the answer was not explicitly stated but had to be inferred by the listener). Batty found that test takers scored higher when they had audio-visual input than when they had audio-only input and that this effect was larger for the implicit items than for the explicit items.

These traditional methods of researching L2 listening have been informative and influential, yet L2 listening presents challenges to researchers (as does L2 reading) because it is an internal process. Unlike speaking and writing, in which the test taker's output can be analyzed, listening researchers have to make inferences about the processes going on inside the listener's head. Three ways that researchers have dealt with this challenge in their research is to employ survey methodology, verbal report methodology, and eye-tracking technology.

Survey methodology has been used extensively in L2 listening testing research. Often this research examines listening test takers' opinions, attitudes, or beliefs about different aspects of the test, test format, and spoken texts, as well as issues like motivation and anxiety related to L2 listening testing (e.g., Liao *et al*., 2018; Winke and Lim, 2014; Zhang, 2013). Survey research is useful in many ways for researching L2 listening assessment, especially when it is combined with other research methodologies to triangulate the survey data with test performance data.

Verbal report methodology has been used to examine how L2 listeners process spoken information and the internal thought processes of L2 listeners while taking a listening test (e.g., Ockey, 2007; Wu, 1998; Wagner, 2008). Verbal reports involve the test taker verbalizing or speaking aloud what they are thinking or doing while they perform some task (e.g., while taking a listening test). This act of verbalizing is sometimes referred to as talk alouds, think alouds, or verbal protocols. Obviously, verbal report methodology has a number of limitations (e.g., the subjects' reports of their internal thought processes might not be accurate, the act of verbally reporting might change the listening process, etc.) (Ericsson and Simon, 1993), but even with these many limitations, the methodology has been used

extensively in L2 listening test research because it provides a means of capturing the "complex cognitive skills that underlie item responses and to determine how these are mobilized in particular listening tasks" (Brindley and Slatyer, 2002: 391). As with survey methodology, verbal report methodology is especially useful when combined with other analyses, in order to illuminate the listening test–taking process. An example of this is Wagner (2008). In this study, the quantitative analysis found that the group that was presented with a videotext scored higher on an L2 listening test than a comparable group that was presented with an audio-only text. Wagner then used verbal report methodology to examine possible reasons why the videotext was helpful for the test takers.

Similarly, eye-tracking technology has increasingly been used to examine what listening test takers orient to (as measured by eye gaze) when taking a listening test, including what components of the visual text they focus on (e.g., Batty, 2020; Suvorov, 2015) and what test takers do when taking a video listening test while also taking notes (Cublio and Winke, 2013). Suvorov (2018) combined eye-tracking technology with verbal report methodology to examine how test takers use the visual information provided by an audio-visual text to answer comprehension questions.

Future directions

Bachman and Palmer (1996) defined test authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a target language use task" (p. 23). Using this definition of authenticity, Ockey and Wagner (2018b) argued that by making L2 listening test tasks more similar to and representative of real-life listening tasks, test developers are able to make more valid inferences about the test takers' ability beyond the testing context and to make a more convincing validity argument in support of their test for a particular purpose. Making the test tasks more authentic includes not only the spoken input that is delivered to the test takers, but also how the input is delivered and how test takers are supposed to respond to that input. Technology can be used to make the test tasks more authentic to real-life listening tasks. An obvious way technology can be used is through audio-visual texts, but as noted earlier, audio-only texts continue to be the norm on most listening tests (Wagner, 2014a).

Technology can also be used to make listening test tasks more authentic to real-world listening tasks by providing more context and allowing for easier administration of integrated tasks. Scenario-based assessment is premised on the idea that technology presents a rich context in which test takers receive written, audio, and visual input and have to perform a number of tasks, individually and with other test takers, in order to successfully complete the scenario (Turner and Purpura, 2016). Technology is emerging that allows test developers to create interactive listening and speaking tasks, in which the computer will provide spoken input, the test taker will have to listen and respond appropriately, and then the computer will respond orally to the test taker's spoken response, and so on (Ockey and Chukharev-Hudilainen, 2021). Such a system should allow for a more systematic and fair way to provide the spoken input that test takers must listen and respond to.

Further reading

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.

This book remains the most comprehensive and important work on L2 listening assessment. In the book, Buck provides an overview of L2 listening comprehension, focusing on how listening differs

from the other language skills. He also gives his "default listening construct," which has been hugely influential for listening test developers. He also provides numerous examples of test tasks, suitable texts, illustrative tests of L2 listening, and practical suggestions for creating L2 listening tests.

Geranpayeh, A. and Taylor, L. (eds.). (2013). *Examining Listening: Research and Practice in Assessing Second Language Listening*. Studies in Language Testing, vol. 35. Cambridge: UCLES, Cambridge University Press.

This edited volume is part of the series *Studies in Language Testing* and examines topics related to the listening components of the Cambridge suite of assessments. (The other books in the series focus on assessing writing, speaking, reading, and young language learners.) Like the other books in the series, this volume follows Weir's (2005) socio-cognitive framework for language test validation. Of particular interest is John Field's chapter that examines the issue of cognitive validity in relation to L2 listening.

Ockey, G. and Wagner, E. (2018). *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins.

In this hybrid monograph/edited volume, Ockey and Wagner identified four themes related to L2 listening assessment that have been the focus of extensive research: the use of real-world spoken texts, using different types of speech varieties as input, the use of audio-visual texts, and interactive listening as part of the construct of L2 listening ability. The authors provide extensive reviews of the literature for each of these four themes, and then each theme is addressed by two or three empirical studies. They argue that what these four themes have in common is the idea of using more authentic test tasks in order to increase the validity of listening assessments.

References

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing* 13: 257–79. https://doi.org/10.1177/026553229601300303.

Batty, A. O. (2015). A comparison of video- and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing* 32: 3–20. https://doi.org/10.1177/0265532214531254.

———. (2018). Investigating the impact of nonverbal communication cues on listening item types. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins, 161–75. https://doi.org/10.1075/lllt.50.11bat.

———. (2020). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing*. [OnlineFirst]. https://doi.org/10.1177/0265532220951504.

Bonk, W. J. and Ockey, G. J. (2003). A many-facet Rasch analysis of the L2 group oral discussion task. *Language Testing* 20: 89–110. https://doi.org/10.1191/0265532203lt245oa.

Bradlow, A. R. and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106: 707–29. https://doi.org/10.1016/j.cognition.2007.04.005.

Brindley, G. and Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing* 19: 369–94. https://doi.org/10.1191/0265532202lt236oa.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing* 26*:* 341–66. https://doi.org/10.1177/0265532209104666.

Brown, J. D. (2012). Classical test theory. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. Abingdon, UK: Routledge, 323–35. https://doi.org/10.4324/9780203181287.ch22.

Brunfaut, T. (2016). Assessing listening. In D. Tsagari and J. Baneerjee (eds.), *Handbook of Second Language Assessment*. Boston, MA: De Gruyter, Inc., 97–112. https://doi.org/10.1515/9781614513827-009.

Buck, G. (2001). *Assessing Listening*. Cambridge: Cambridge University Press.

Burgoon, J. (1994). Nonverbal signals. In M. Knapp and G. Miller (eds.), *Handbook of Interpersonal Communication*. London, UK: Routledge, 344–93.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly* 3: 229–42. https://doi.org/10.1207/s15434311laq0303_1.

Carter, R. and McCarthy, M. (2006). *Cambridge Grammar of English: A Comprehensive Guide. Spoken and Written English Grammar and Usage*. Cambridge, UK: Cambridge University Press.

Chafe, W. (1985). Linguistics differences produced by differences between speaking and writing. In D. Olson, D. Torrance and A. Hildyard (eds.), *Literacy Language and Learning*. Cambridge, UK: Cambridge University Press, 105–23.

Choi, I. and So, Y. (2018). A measurement model for listen-speak tasks. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins, 227–45. https://doi.org/10.1075/lllt.50.15cho.

Coniam, D. (2001). The use of audio or video comprehension as an assessment instrument in the certification of English language teachers: A case study. *System* 29: 1–14. https://doi.org/10.1016/s0346-251x(00)00057-9.

Cross, J. (2011). Comprehending news videotexts: The influence of the visual content. *Language Learning and Technology* 15: 44–68.

Cublio, J. and Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. *Language Assessment Quarterly* 10: 371–97. https://doi.org/10.1080/15434303.2013.824972.

Douglas, D. (1988). Testing listening comprehension in the context of the ACTFL proficiency guidelines. *Studies in Second Language Acquisition* 10: 345–61. https://doi.org/10.1017/s0272263100007336.

———. (1997). *Testing Speaking Ability in Academic Contexts: Theoretical Considerations*. TOEFL Monograph Series, No. 8. Princeton: Educational Testing Service.

Ericsson, K. and Simon, H. (1993). *Protocol Analysis: Verbal Reports as Data*, 2nd edn. Cambridge, MA: MIT Press.

Field, J. (2008). *Listening in the Language Classroom*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511575945.

———. (2013). Cognitive validity. In A. Geranpayeh and L. Taylor (eds.), *Examining Listening: Research and Practice in Assessing Second Language Listening*. Studies in Language Testing, vol. 35. Cambridge, UK: UCLES, Cambridge University Press, 77–151.

Fox Tree, J. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language* 34: 709–38. https://doi.org/10.1006/jmla.1995.1032.

Freedle, R. and Kostin, I. (1999). Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* 16: 2–32. http://dx.doi.org/10.1177/026553229901600102.

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired tests? *Applied Linguistics* 35: 553–74. https://doi.org/10.1093/applin/amt017.

Gass, S. M. and Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning* 34: 65–89. https://doi.org/10.1111/j.1467-1770.1984.tb00996.x.

Gruba, P. (1993). A comparison study of audio and video in language testing. *JALT Journal* 15: 85–88.

Harding, L. (2018). Listening to an unfamiliar accent: Exploring difficulty, strategy use, and evidence of adaptation on listening assessment tasks. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins, 227–45. https://doi.org/10.1075/lllt.50.07har.

Hilsdon, J. (1995). The group oral exam: Advantages and limitations. In J. Alderson and B. North (eds.), *Language Testing in the 1990s: The Communicative Legacy*. Hertfordshire, UK: Prentice Hall International, 189–97.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. London, UK: Longman.

Liao, Y. F., Wagner, E. and Wagner, S. (2018). Test-takers' attitudes and beliefs about the spoken texts used in EFL listening tests. *English Teaching & Learning* 42: 227–36. http://doi.org/10.1007/s42321-018-0013-5.

McCarthy, M. and Carter, R. (1995). Spoken grammar: What is it and how can we teach it? *ELT Journal* 49: 207–18. https://doi.org/10.1093/elt/49.3.207.

Messick, S. (1989). Validity. In R. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: American Council on Education and Macmillan, 13–103.

———. (1996). Validity and washback in language testing. *Language Testing* 13: 242–56. https://doi.org/10.1177/026553229601300302.

Milroy, J. and Milroy, L. (1999). *Authority in Language: Investigating Standard English*. London, UK: Routledge.

Nakatsuhara, F. (2012). The relationship between test-takers' listening proficiency and their performance on the IELTS speaking test. In L. Taylor and C. J. Weir (eds.), *IELTS Collected Papers 2: Research in Reading and Listening Assessment*. Cambridge, UK: Cambridge University Press, 519–73.

Ockey, G. J. (2007). Construct implication of including still image or video in computer-based listening tests. *Language Testing* 24: 517–37. https://doi.org/10.1177/0265532207080771.

———. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. *English for Specific Purposes* 35: 17–29. https://doi.org/10.1016/j.esp.2013.11.003.

———. (2018). The degree to which it matters if an oral test task requires listening. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins, 193–202. https://doi.org/10.1075/lllt.50.13ock.

Ockey, G. J. and Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics*.

Ockey, G. J., Koyama, D., Setoguchi, E. and Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing* 32: 39–62. https://doi.org/10.1177/0265532214538014.

Ockey, G. J. and Wagner, E. (2018a). An overview of interactive listening as part of the construct of interactive and integrated oral test tasks. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins, 179–92. https://doi.org/10.1075/lllt.50.c12.

———. (2018b). *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins. https://doi.org/10.1075/lllt.50.

Oller, J. (1979). *Language Tests at School*. London, UK: Longman.

Plakans, L. (2012). Writing integrated items. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. Abingdon, UK: Routledge, 249–61. https://doi.org/10.4324/9780203181287.ch17.

Revesz, A. and Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition* 35: 31–65. https://doi.org/10.1017/s0272263112000678.

Rost, M. (2015). *Teaching and Researching Listening*, 3rd edn. London, UK: Taylor and Francis. https://doi.org/10.4324/9781315732862.

Savignon, S. (2018). Communicative competence. In J. Liontas (ed.), *The TESOL Encyclopedia of English Language Teaching*. Oxford, UK: Wiley-Blackwell. Retrieved from https://onlinelibrary.wiley.com/doi/full/10.1002/9781118784235.eelt0047.

Sueyoshi, A. and Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning* 55: 661–99. https://doi.org/10.1111/j.0023-8333.2005.00320.x.

Suvorov, R. (2009). Context visuals in L2 listening tests: The effects of photographs and video vs. audio-only format. In C. A. Chapelle, H. G. Jun and I. Katz (eds.), *Developing and Evaluating Language Learning Materials*. Ames, IA: Iowa State University, 53–68.

———. (2015). The use of eye tracking in research on video-based second language L2 listening assessment: A comparison of context videos and content videos. *Language Testing* 32: 463–83. https://doi.org/10.1177/0265532214562099.

———. (2018). Test-takers' use of visual information in an L2 video-mediated listening test: Evidence from cued retrospective reporting. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins, 145–60. https://doi.org/10.1075/lllt.50.10suv.

Tannen, D. (1982). The oral/literate continuum in discourse. In D. Tannen (ed.), *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, NJ: Ablex, 1–33.

Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking, 30*. Cambridge, UK: Cambridge University Press, 171–233.

Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Baneerjee (eds.), *Handbook of Second Language Assessment*. Boston, MA: De Gruyter, Inc, 255–72. https://doi.org/10.1515/9781614513827-018.

Wagner, E. (2008). Video listening tests: What are they measuring? *Language Assessment Quarterly* 5: 218–43. https://doi.org/10.1080/15434300802213015.

———. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing* 27: 493–513. https://doi.org/10.1177/0265532209355668.

———. (2013). An investigation of how the channel of input and access to test questions affect L2 listening test performance. *Language Assessment Quarterly* 10: 178–95. https://doi.org/10.1080/15434303.2013.769552.

———. (2014a). Assessing listening. In A. Kunnan (ed.), *Companion to Language Assessment*, vol. 1. Oxford, UK: Wiley-Blackwell, 47–63. https://doi.org/10.1002/9781118411360.wbcla094.

———. (2014b). Using unscripted spoken texts to prepare L2 learners for real world listening. *TESOL Journal* 5: 288–311. https://doi.org/10.1002/tesj.120.

———. (2016). Authentic texts in the assessment of L2 listening ability. In J. Banarjee and D. Tsagari (eds.), *Contemporary Second Language Assessment*. London, UK: Continuum, 438–63. https://doi.org/10.5040/9781474295055.ch-005.

———. (2018). Texts for listening instruction and assessment. In J. Liontas (ed.), *The TESOL Encyclopedia of English Language Teaching*, vol. 3. Oxford, UK: Wiley-Blackwell, 1544–55. https://doi.org/10.1002/9781118784235.eelt0626.

Wagner, E. and Ockey, G. J. (2018). An overview of the use of audio-visual texts on L2 listening tests. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam: John Benjamins, 129–44. https://doi.org/10.1075/lllt.50.c9.

Wagner, E. and Wagner, S. (2016). Scripted and unscripted spoken texts used in listening tasks on high stakes tests in China, Japan, and Taiwan. In V. Aryadoust and J. Fox (eds.), *Current Trends in Language Testing in the Pacific Rim and the Middle East: Policies, Analyses, and Diagnoses*. Newcastle upon Tyne, UK: Cambridge Scholars Publishing, 103–23.

Wall, D. (2012). Washback. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. Abingdon, UK: Routledge, 79–92. https://doi.org/10.4324/9780203181287.ch5.

Winke, P. and Lim, H. (2014). *The Effects of Testwiseness and Test-Taking Anxiety on L2 Listening Test Performance: A Visual (Eye-Tracking) and Attentional Investigation*. IELTS Research Reports Online Series 2014/3. Cambridge, UK: British Council, Cambridge English Language Assessment/IDP, IELTS Australia.

Wu, Y. (1998). What do tests of listening comprehension test? – A retrospection study of EFL testtakers performing a multiple-choice task. *Language Testing* 15: 21–44. https://doi.org/10.1191/026553298673885021.

Yanagawa, K. (2016). Examining the authenticity of the center listening test: Speech rate, reduced forms, hesitation and fillers, and processing levels. *JACET Journal* 60: 97–115.

Zhang, X. (2013). Foreign language listening anxiety and listening performance: Conceptualizations and causal relationships. *System* 41: 164–77. https://doi.org/10.1016/j.system.2013.01.004.

14
Assessing writing

Ute Knoch

Introduction

The ability to write in a second language is widely regarded as essential to L2 learner advancement across social, educational, and workplace settings. The assessment of second language writing ability has therefore been at the heart of second language assessment for many decades. This chapter outlines the key debates and issues, both historical and current, that have influenced practices in second language writing assessment. It proposes a reconceptualization of L2 writing assessment practices by organizing key considerations by evidence, scoring, interpretation, and use, drawing on the framework developed by Hill and McNamara (2012) and focusing more closely on work in the area of written corrective feedback. The chapter concludes by setting out future directions for second language writing assessment research and practice, focusing particularly on the intersection of writing assessment and written corrective feedback, the assessment of writing development, and the assessment of collaborative writing.

Historical perspectives

The historical development of writing assessment has been documented in a number of publications (e.g., Crusan, 2014; Eckes *et al*., 2016; Hamp-Lyons, 2001; Yancey, 1999). Historical trends or changes in writing assessment can essentially be understood to be related to tensions between reliability, validity, and practicality. The earliest records of writing assessments stem from China, where, as far back as the Chou period (111–771 B.C.), essay testing was used in the process of selecting officials. During these early times, anonymity of marking, as well as the use of more than one examiner, were already key considerations. Early documented writing assessment practices for university selection, as well as the selection of public servants in Britain, Europe, and the United States, show that these traditions in the mid-to-late nineteenth century also involved the collection of essays. However, particularly in the United States, there were soon concerns voiced about the reliability of the scoring of these essays, and in the mid-twentieth century, multiple-choice tests replaced essays as a method of assessment. This change was largely driven by the strong tradition

 DOI: 10.4324/9781003220756-19

of psychometrics in the US and resulted in a greater focus on form in writing and a concern over associated construct underrepresentation of writing skills (Hamp-Lyons, 2001). From the end of the 1970s, several universities in the US re-introduced direct assessments of writing to deal with concerns about the validity of such indirect assessments (Messick, 1989). More recently, Hamp-Lyons (2001) and Yancey (1999) described another "wave" or "generation" of writing assessment which came in the form of portfolio assessments. These assessments were seen to be more reflective of the teaching process in writing classrooms and therefore better at capturing the construct of writing with an associated increase in validity, but at the same time, there were concerns about the marking reliability and practicality of these assessments (Lam, 2018).

More recently, writing assessment has been caught by a further wave – technology. This trend is motivated by practicality and has moved not only into high-stakes assessments but also into the classroom, where automated essay evaluation and feedback systems are increasingly being used to supplement or replace teacher feedback practices. This most recent development is accompanied by concerns around validity associated with the scoring engines of automated systems, which are still only able to capture a limited construct (e.g., Deane, 2013). While technology has taken on an important role in writing assessment, there have been other recent developments that have broadened our conceptualization of writing assessment: for example, the design of innovative task types. All traditions described here have had a direct impact on the teaching and learning of writing: a phenomenon known as washback (see Cheng and Sultana, this volume; also Bailey, 1996; Cheng *et al*., 2015; Wall, 2012). A move to indirect assessments, for example, was accompanied by less direct writing instruction and practice in classrooms and an associated lesser focus on the social aspects of writing or writing as human communication. On the other hand, integrated assessments (e.g., listening, reading-to-write tasks) (Gebril and Plakans, 2013) have had a positive influence on preparing students for the writing demands of university.

Critical issues and current research

Language assessment specialists are involved in the process of collecting evidence, scoring this evidence, interpreting the score, and using the score or making suggestions about score use. This process can be visually represented, as set out in Figure 14.1.

In the case of writing assessment, these four dimensions are each associated with a number of considerations which apply to all or most assessment contexts and purposes (Table 14.1).

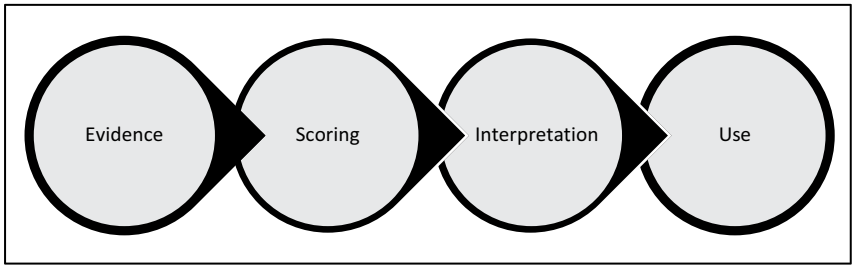


Figure 14.1 Process of writing assessment

Table 14.1 Dimensions and considerations in writing assessment

Dimensions	*Considerations*	*Examples/explanation*
Evidence	What is assessed?	Assessment construct, possibly established by conducting a needs analysis; prioritized aspects of writing
	How is evidence collected?	Writing assessment may be planned/scheduled or less formal/planned in classroom context.
	When is evidence collected?	Once off; integrated at specific time in teaching cycle
	Who is assessed?	Individual/pairs/groups
Scoring	Who scores/corrects the performance?	Language testing company; automated scoring platform; teacher; learners (e.g., self or peer)
	How are the scoring criteria represented?	Scoring criteria (various layouts – analytic, holistic, diagnostic) Scoring/feedback may be visible to learners or embedded in other tasks
	What is the purpose of the scoring?	High-stakes decisions; formative feedback; identifying position on continuum (e.g., standards document); providing feedback in relation to institutional criteria/objectives
	How is scoring reliability established?	Statistical analysis; moderation meeting; scorer training; no reliability check
Interpretation	What does the score/result tell us about the writer?	What does the score/result mean to users (e.g., does it place the writer on a developmental continuum of writing ability/language proficiency?) How can the score be interpreted by users?
Use	How is the score/result used?	What is the purpose of the score/results use? Are the score/results used for any high-stakes decisions? Does the score/result inform teaching and/or learning decisions? (e.g., assign level, change level, modify teaching plans, accountability decisions)
	Who uses the score/result?	Teachers, learners, school, management, educational authorities
	What are the consequences of score use?	Are the consequences of score use intended/unintended?

Source: Adapted from Hill and McNamara (2012: 397).

In the following sections, I will consider each of the four dimensions in more detail, describing the most important considerations in Table 14.1. In relation to "evidence," I focus particularly on the construct assessed. When discussing "scoring," the focus is on scoring rubrics, scoring reliability, and automated scoring. In the dimension of "interpretation," I discuss the reporting of outcomes in relation to learner progression, as indicated through learning objectives, standards frameworks, and portfolios. Finally, in the dimension of "use," I focus on feedback, in particular various types of feedback provided in the different traditions of written corrective feedback and second language writing assessment.

Evidence

Second language writing is complex and multi-faceted and differs depending on purpose, context, and task type. Key agents involved in the assessment context therefore need to

make decisions about what aspects of writing to prioritize. What is assessed, or the *construct* of an assessment, is a concept that is difficult to pin down. What is observed usually represents a sample of language use behavior, and this is considered to represent the ability of the person providing the sample. This is also the case for writing assessment in which there is no agreed theory of writing or consensus on the construct of L2 writing (Hamp-Lyons, 1990). Constructs are complex, and various stakeholders may have different relationships with these constructs and view them differently. Constructs may be clearly stated or hidden. Macqueen (2021) makes the useful distinction between stated, operationalized, theoretical, and perceived constructs. The stated construct is what the assessment designer or teacher (in a classroom context) claims an assessment elicits and measures. The operationalized construct, on the other hand, is what is actually assessed. The theoretical construct is less tangible and represents the "unobservable ability which is assumed to explain differences in assessee responses" (Knoch and Macqueen, 2020: 45). Ideally, the stated and operationalized constructs should overlap. The operationalized construct, however, is not visible to the public, but is rather represented in test specifications and/or test materials or evidenced through validation research. Finally, the perceived construct is what users understand is being assessed. In contexts in which assessments and feedback are less standardized, the construct of an assessment may be more hidden. In many classroom contexts, for example, teachers may not be aware of the construct they are operationalizing when providing feedback on writing – the feedback may be more based on their own experiences as learners, which have shaped their beliefs and practices (e.g., Ferris, 2014; Goldstein, 2006; Gurzynski-Weiss, 2016).

For writing test designers, regardless of the stakes of the assessment, it is important to reflect on where the construct comes from. Depending on the purpose and the context of the assessment, the construct may be based on assessment-external or assessment-internal aspects. As an example of an *assessment-external source of test construct*, the assessment construct may be based on curriculum documents or external standards documents, such as the relevant writing scales in the Common European Framework of Reference (Council of Europe, 2001, 2018). Curriculum documents are often vague and may not provide sufficient information to inform a test construct by themselves, but they may be used in combination with other sources of construct. Two other assessment-external construct sources may be a description of the target language use (TLU) domain (based on a needs analysis) or a theoretical model of language or writing (Bachman and Palmer, 1996; Grabe and Kaplan, 1996). We will look at these two sources of construct in more detail.

Test constructs that are directly based on the TLU domain to which the assessment is designed to predict are often created following a *needs analysis*. Needs analysis refers to any technique of collecting data from a target language use domain which is used to shape the assessment construct. This is an important process because assessment designers are often not insiders of the domain, and without this process of collecting information about the typical communication tasks in the domain, as well as the language typically used to complete these tasks, intuition about the domain may be incorrect. The majority of work on needs analyses (both conceptual and practical) has occurred in the area of language curriculum design (e.g., Brown, 2016; Long, 2005); much less theoretical work has been done in this area in language assessment. Recently, Knoch and Macqueen (2020) proposed a multi-componential model of the aspects that may be collected during such a needs analysis. They followed Hyland's (2006) call to broaden what is included in a needs analysis and adapted these considerations for language assessment research and practice. The model includes five key components: (1) domain analysis, which includes an analysis

of the TLU writing tasks and language, a domain task processing analysis, an examination of the TLU writing criteria (indigenous criteria), and the TLU writing standards; (2) a test-taker analysis, which may focus on test-taker proficiency and typical language difficulties experienced when writing in the domain, and a test-taker perceptions, needs and motivation analysis; (3) a test requirement analysis examining the writing test prerequisites, score reporting needs, examination of the decisions, and inferences that need to be made on the basis of the writing scores; (4) policy analysis, which explores the relevant laws and policy documentation; and (5) a means analysis, which documents the resources available for test development, administration, and validation. Knoch and Macqueen (2020) also document the process of moving from a needs analysis to a test blueprint, a process which is mostly not documented. A well-known example of a test that was developed following a needs analysis is the Occupational English Test (OET). McNamara, the developer of the OET, documents in detail how he went about developing the OET (see, e.g., McNamara, 1996). He interviewed doctors teaching overseas-trained medical professionals, conducted a survey of overseas-trained medical professionals who had already successfully registered in Australia, and also did a careful review of the literature on clinical communication. He then focused on those tasks that were mentioned as being the most frequent, important, and difficult for overseas-trained medical professionals. In the case of writing, this resulted in the inclusion of a referral letter task in which the test takers get a set of case notes and need to write a letter of referral or discharge.

Theoretical models of language are also at times used as the basis of test constructs. More general models of language (e.g., Bachman and Palmer, 1996) are not specifically focused on writing and may therefore result in more general test constructs. To date, few models of writing are available. These models can be divided into two groups – those that are focused on processing during writing and those that set out a model of texts. All these models are from the literature on L1 writing, and few specific models focusing on L2 writing are available. The L1 processing models (e.g., Bereiter and Scardamalia, 1987; Hayes and Flower, 1980; Hayes, 1996; Kellog, 1996) are generally cognitive in nature. They all show that during the composing process, a writer conducts several activities simultaneously. These include aspects such as macro-planning (gathering ideas against task demands), organizing (organizing ideas, identifying relationships between ideas), micro-planning (focusing on the part of text that is currently being written), translating propositional content into linguistic forms, monitoring of processes, and possibly revising. The models differ in how these different components are related, how the recursiveness of writing is treated, and the role of memory. However, none of these models specifically focuses on L2 writers, whose writing processes may be less automatic and may be marked by a higher cognitive load, resulting in a more laborious and slower process of converting ideas to linguistic form. An early model focused on L2 writing created by Grabe and Kaplan (1996) has also been prominent in the literature. Although not a sequential model, it shows how various aspects of processing interact. To account for L2 writing, models need to focus more closely on the processes relating to how ideas are converted into linguistic forms (including lexical retrieval, graphic transcription, and fluent access to forms), how this impacts fluency, and how writing processes in L1 may influence those in L2 composing. What all the processing models are missing, however, due to their cognitive nature, is a focus on the social context in which writing takes place. Despite the depth of research available into these cognitive processing models of writing, their insights are still limited for test developers, who usually focus their attention on the output only and require guidance for the development of tasks and rating scales. For this reason, rating scales (the operational representations of test constructs) are

more likely to be based on more general theories of writing (e.g., Grabe and Kaplan, 1996) or on other sources of test construct.

Assessment-internal sources of writing assessment constructs may be related to the tasks that are used, as well as teacher/rater background and past experiences, which all may shape how writing is assessed. In rater-mediated assessment, the construct of the assessment is usually represented in the scoring criteria, if available, as well as in the way raters apply these criteria.

Sometimes the construct is shaped by other forces: for example, by considerations of practicality or technological considerations. In the case of automated essay scoring (AES) systems, for example, the construct may be limited by the capabilities of the technology. We will consider this in more detail in the section on scoring.

Scoring

The second dimension in Table 14.1 is scoring. Discussions around scoring in language assessment more broadly and writing assessment more specifically have generally focused on one of three topics in rater-mediated assessment contexts: (1) scoring rubrics, (2) establishing scoring reliability, and more recently (3) automated scoring. In the following section, each of these areas will be briefly reviewed as they relate to writing assessment. (Please also refer to Galaczi and Lim, this volume; Xi, this volume for more detailed discussions of scoring performance tests and automated scoring.)

Scoring rubrics

In rater-mediated writing assessment, scoring rubrics are typically used to guide the rating process and to ensure that the operationalized test construct is consistently applied. These scoring rubrics can take various formats (e.g., holistic, analytic, or diagnostic) and may have different conceptual underpinnings. What they have in common is that they are a representation of the test construct and set out the scope of the aspects of writing that are to be assessed and prioritized. The choice of which aspects of writing are represented in the rating scale depends on the scope of the construct of the assessment but may also be dependent on the rating scale development method chosen (Fulcher, 1996; Fulcher *et al*., 2011; Knoch *et al*., 2021). The rating scales of well-known writing tests differ in how they combine aspects into scale criteria and what criteria are included in the first place. For example, the Task 2 Writing rating scale of the International English Language Testing System (IELTS) includes four separate criteria (task response, coherence and cohesion, lexical resource, grammatical range and accuracy) while the holistic TOEFL independent rating scale includes bullet points at each score level focusing on task fulfilment, organization and development, coherence and cohesion, and language.

Of course, not all writing assessment draws on the use of rubrics, which will be discussed later in this chapter. In classroom-based assessment, teachers often offer written corrective feedback (WCF) when they are providing formative assessment to learners (during the course of their studies) while they may draw on rating scales as part of summative assessment processes as a score is required. The use of diagnostic feedback is also becoming more common in writing classrooms, with the advantage that this kind of feedback is typically more systematic than WCF practices. This is because teachers may provide WCF unevenly across different skill areas (e.g., grammar, vocabulary, organization, etc.) without realizing it. Diagnostic rubrics ensure even construct coverage during feedback provision but may

have the disadvantage that learners are less likely to be able to identify specific errors in their writing due to the more generic feedback indicators used.

Scoring reliability

Establishing acceptable levels of scoring reliability has been another major focus of language assessment specialists, and writing assessment has been no exception. Several methods have been used to ensure reliability in rater-mediated writing assessment, including rater training (Davis, 2016; Weigle, 1998), re-standardization of raters at regular intervals, ongoing monitoring of raters in high-stakes testing, and the use of more than one rater, as well as the use of score resolution techniques. Rater training is conducted to ensure that raters understand the test construct and come to a consistent interpretation of the scoring criteria (Weigle, 2002). Empirical studies on the effects of training on reliability have shown that training helps raters understand the task demands and test-taker characteristics, increase the understanding of scoring criteria, reduce extreme tendency in terms of lenience and severity, increase rater self-consistency, and reduce individual rater biases (Davis, 2016; Lim, 2011; Weigle, 1999). Rater training can take the form of two quite different models. In more top-down training, raters are required to conform to a "gold standard" set by the organization administering the test. In other contexts, training might take on the form of a community of practice, where standards are agreed collaboratively by a group of professionals (Shaw, 2004; Watts, 2006; Wenger, 1998; Wenger and Snyder, 2000; Wenger *et al*., 2002). In practice, rather than forming a dichotomy, these two training approaches can also form a continuum, with a mixture of both forms being common in various contexts.

Automated scoring

Automated essay scoring (AES) is becoming increasingly popular after significant progress has been made in the development of such programs, in particular in large-scale assessment contexts "in response to the increasing demand for more efficient constructed-response scoring and the maturation of computer technologies" (Williamson *et al*., 2012: 4). Such systems usually work "by extracting a set of attributes and using some machine learning algorithm to model and predict the final score" (Zupanc and Bosnic, 2015: 387). There are several reasons such AES systems are becoming more prevalent. Firstly, as already mentioned, they provide for efficiency and cost effectiveness in the scoring process. Large numbers of texts can be evaluated, and scores can be returned within hours of the test administration. Proponents of AES also argue for the systems' scoring objectivity and consistency, as these are not susceptible to biases or fatigue, like human raters. However, there are also concerns about automated scoring. These concerns are usually directly related to the construct that is represented in AES systems, which are generally only able to assess a very limited set of writing features, mostly focusing on structure and grammatical accuracy. Deane (2013) argues that AES systems evaluate the quality of the text, whereas human raters consider the writing proficiency of the learner as a writer. He further suggests that in direct, timed writing assessments scored by human raters, certain writing skills are prioritized whereas others are de-emphasized: "explicit text features and content are foregrounded, while social and process elements of writing expertise are backgrounded" (p. 11). It is therefore important to consider the construct when a scoring method is selected, also potentially taking washback on learning and teaching into account. A more detailed exploration of AES is beyond the scope of this chapter. I refer the reader to Xi (this volume) for further information.

Interpretation

It is common practice in language assessment that scores are not provided to learners in isolation. Usually, score meaning is explained in terms of what a learner who received a certain score or who falls within a certain score range can do. In classroom contexts, this score meaning is usually expressed in terms of the learning objectives or specific competencies that learners should attain in a course. In other contexts, test designers may draw on standards frameworks such as the Common European Framework of Reference (CEFR, Council of Europe). Such frameworks are increasingly used by textbook developers to sequence their course materials and by test developers to link their tests. The use of standards frameworks is attractive for teachers and assessment developers because they present language proficiency as a linear continuum on which learners progress over time. In fact, showing writing development is one of the purposes of writing assessment, as this information is needed for summative reporting, accountability purposes, curriculum development, and program evaluation (please also see Chapter 34 this volume).

However, there are limitations in language frameworks and how well they can represent writing development as gathered through assessment data. Firstly, these frameworks are generally not developed based on longitudinal writing data. The CEFR, for example, is based on intuitive teacher judgments, and there is limited empirical evidence, for example, for the task hierarchy represented in the framework. The descriptions are also very general and broad, and teachers may therefore overlook important indicators of learning as certain features do not appear in the framework. It has also been shown that one direct result of this generality is a narrowing of the curriculum (Fulcher, 2010). The broadness of the levels represented in these frameworks also means that it is often difficult to document smaller instances of writing development, a problem that has been compared with the map metaphor, in which maps of larger areas provide less fine-grained detail (Elder and O'Loughlin, 2007). The descriptions are also often inconsistent across levels and focus mostly on general proficiency and are therefore less applicable to academic or professional writing development. There have, of course, been attempts to deal with these issues: (1) work has been done to create more detailed descriptions of the levels, (2) plus levels have been created (see, e.g., Council of Europe, 2018), and (3) specific purpose CEFR frameworks have also been developed for various purposes (e.g., sign languages, young learners, professional purposes).

One branch of research that lends support to the non-linear nature of second language development more broadly and writing development more specifically is that of complex dynamic systems theory (see, e.g., Fogal and Verspoor, 2020; Verspoor *et al*., 2008; Verspoor and Behrens, 2011). Studies drawing on this theory argue that ignoring developmental nuances of individual learners may bypass the highly contextualized and dynamic nature of writing development (Byrnes, 2014; Fogal, 2019). Work in this area is only recently emerging and may be useful in providing language assessors with further information for test score interpretation in the future.

Use

Scores and feedback generated from an assessment situation can be used in a variety of ways, including making decisions of various stakes, informing teaching and learning, and for accountability purposes. In the context of classroom-based writing assessment, the use of assessment information has received less focus in the research literature. This is likely because research on written corrective feedback (mostly situated in the field of second

language acquisition or college composition) and research on writing assessment (from the area of language assessment) are based in conceptually quite different fields and therefore differ in focus and scope. Studies on WCF have often narrowly focused on feedback on grammatical form (e.g., Bitchener and Ferris, 2012; Bitchener and Knoch, 2010; Bitchener *et al*., 2005; Ellis *et al*., 2008; Guenette and Lyster, 2013; Sheen, 2007; Van Beuningen *et al*., 2008), and therefore the construct of the assessment, as well as the feedback received by test takers, is more narrow than the type of constructs often included in writing assessments (which usually also focus on organization, content, coherence, lexis, etc.). At the same time, teachers' WCF practices provide very detailed feedback to learners, often in scaffolded format, which can lead to revisions and noticing (Schmidt, 1995, 2001) and may lead to the acquisition of forms. Developers of automated writing evaluation (AWE) platforms claim that such systems are able to mimic this type of teacher-generated feedback. Research on student responses to WCF have focused on the students' engagement with the feedback as well as their noticing (Schmidt, 1995; Swain and Lapkin, 1995). Another aspect that has been investigated is the learners' uptake of the feedback in revisions, as well as their retention of such linguistic forms (Min, 2006) and learners' perceptions of the feedback (e.g., Montgomery and Baker, 2007). More recently, the same concepts have also been investigated in studies on AWE systems (Chapelle *et al*., 2015; Dikli and Bleyle, 2014; Li *et al*., 2015; Zhang and Hyland, 2018).

What has received much less attention is how learners engage with score reports generated by more standardized assessments. Similarly, how teachers are able to use the information they learn when marking essays and providing feedback to students in classroom contexts is still under-researched.

Future directions

Linking assessing writing with work on written corrective feedback

In the earlier sections, I have argued that research on WCF in classroom contexts and work done on writing assessment (both in classroom contexts and more generally) has failed at times to consider the theoretical basis and body of work in the other discipline. This has led to the fact that the two disciplines have largely developed in isolation from each other, with the focus of writing assessment more on large-scale standardized tests (with the associated focus on constructs and scoring reliability), and researchers on WCF focusing on classroom contexts, researching the efficacy of different types of feedback as well as, more recently, on learner engagement with the feedback. However, the field of language assessment is increasingly becoming interested in classroom assessment (see Fox *et al*., this volume; Turner and Purpura, 2015), and this has been associated with a shift to research on learning-oriented assessment and, in the case of writing assessment more specifically, to diagnostic writing assessment (see Jang and Sinclair, this volume; Kim, 2011; Wagner, 2015), as well as dynamic assessment (see Antón and García, this volume; Lantolf and Poehner, 2004). It could be argued that the two fields have much to learn from each other, and great cooperation would enrich both disciplines. For the purpose of this chapter, I have attempted to blend these two fields and reconceptualize writing assessment in a number of continua, drawing on the dimensions presented in Table 14.1.

Regarding the evidence that is collected, the construct of what is assessed is often quite different when work on writing assessment and WCF are compared. Research on WCF makes the distinction between focused feedback (which only addresses one or a very limited

number of error types in a student's writing sample) and unfocused feedback, which provides feedback to any error type. However, the feedback has often exclusively focused on grammatical forms, at the expense of taking a broader view of writing. In language assessment, the focus has been generally much broader, also scoring aspects of a writer's lexical range and accuracy, as well as content, organization, coherence, and cohesion. The focal construct can therefore be represented along a continuum, as shown in Table 14.2.

When examining scoring, it is worthwhile considering the systematicity and the level of detail created in the scoring process. In terms of systematicity, it can be argued that teacher WCF is likely to be less systematic than when a scoring rubric is implemented, although this is subject to research. When the level of detail and guidance in the scoring and feedback is considered, a continuum such as the one in Table 14.3 can be created.

When it comes to score use, the purpose of the assessment and the associated scoring can also be represented as a continuum (Table 14.4). Please note that the test types in the middle column are only examples; not all tests report scores or provide feedback in the same way.

The series of tables indicating continua relating to writing assessment show that a closer interaction between work in written corrective feedback and writing assessment would broaden both fields and aid learners in the process.

This merging of the two paradigms of language assessment and classroom feedback may be less useful for large-scale tests, as providing more detailed feedback may be seen as impractical. However, with the recent growth in and increased sophistication of automatic

Table 14.2 Continuum of writing constructs in different assessment types

↓	Focused WCF
	Unfocussed WCF
	WCF + other aspects of writing (content, organization, etc.)

Table 14.3 Continuum of level of detail in scoring

↓	Overall score only
	Indirect feedback (e.g., correction symbol)
	Direct feedback
	Direct feedback + explanation

Table 14.4 Continuum of score and feedback provision related to different assessment types

	Assessment type	*Score/feedback provision*
↓	Screening tests, placement tests	Overall score only
	Proficiency tests, achievement tests	Overall score + subtest scores
	Classroom-based assessments, including diagnostic tests	Detailed non-scaffolded feedback prioritized over scores; designed to encourage learning; improve teaching
	Dynamic assessment; high-quality teacher-written corrective feedback	Scaffolded feedback is provided, within learner's zone of proximal development; scores not important

writing evaluation, it has now become feasible to provide feedback more easily on a larger scale. Similarly, a number of large-scale test providers are also looking at moving into the space of learning-oriented assessment. In both these cases, it may be useful for test developers to look more closely at the literature on written corrective feedback and to examine how current best practice in written corrective feedback and language assessment can be merged.

Collaborative writing

One writing activity that is becoming increasingly popular in L2 classrooms is collaborative writing, which can be defined as the co-authoring of a single text by two or more writers (Storch, 2017). This approach is increasingly being used in university settings and has been shown to pose a challenge to L2 writers, who are often required to collaborate with native English–speaking peers. Jointly authored texts are also very common in workplace settings (see, e.g., Knoch *et al*., 2016) and are becoming increasingly more common in many contexts due to technological affordances, and it is for these reasons important that L2 learners are exposed to such writing processes and also assessed on this type of writing. However, the assessment of collaborative writing has not received much attention in the language assessment literature, and collaborative writing has not been incorporated into any well-known language assessments, while paired and group speaking assessments are increasingly common.

It is important to make a distinction between cooperative writing and collaborative writing (Storch, 2017). During the former, writers divide up the task and work on different sub-tasks, which are later collated. Collaborative writing, on the other hand, means that all co-authors are engaged in all aspects of the task. Cooperative writing is probably more authentic in academic and workplace settings, but both types of writing pose challenges for language assessment. The choice of task is important. While in the L2 teaching literature, a distinction has been made between meaning-focused and language-focused tasks (e.g., Alegria de la Colina and Garcia Mayo, 2007), with a preference for language-focused tasks that elicit language-related episodes from the students, language assessors are more likely to choose meaning-focused tasks, which are more authentic to writing in academic or workplace contexts.

Technological possibilities make it possible to implement collaborative writing online: for example, in Google Docs. Such platforms have the advantage that not only the final product is collected, but data about the writing process is also available. Perry (2019), while not focused on the *assessment* of collaborative writing, was able to collect not only the writing products and a log of changes made but also the online interaction of pairs collaborating on a piece of writing. Learners composed a text while communicating using synchronous computer-mediated communication (SCMC). Any of these aspects could be used in the assessment of the collaboration. This has the advantage of providing an insight into writing processes, an aspect currently neglected by the majority of writing assessments, which are product focused. What the washback would be on students' writing practices and preparation, however, needs to be further explored.

There are a number of important areas that warrant research in relation to collaborative writing assessment. Firstly, what is the impact on text quality and composing processes of co-authoring texts in person or in online synchronous or asynchronous environments? What is the impact of learners of different proficiency levels (e.g., Leeser, 2004; Storch and Aldosari, 2013), and group sizes (e.g., Fernandez Dobao, 2012) on the quality of the writing? What aspects of the construct of collaborative writing (product, process, type of collaboration) should be included in the assessment criteria, and how can these be best represented? And finally, are raters able to reliably score collaborative writing? More work in this area is

undoubtedly going to appear with the increasing use of technology in assessment in coming years. One starting point for further research is the literature on paired and group oral assessments, which have investigated issues such as personality (Berry, 2007; Bonk and Van Moere, 2004; Ockey, 2009), test-taker proficiency (Brooks, 2009; Choi and Iwashita, 2016; Davis, 2009; Nakatsuhara, 2004), the quality of the interactions between learners (Galaczi, 2014), group size (Nakatsuhara, 2011), and rating processes (May, 2009).

Assessing writing development over time

Assessing writing development is complex. As Slomp (2012: 81) points out:

> [L]ike intelligence, the construct of writing ability is broad, multifaceted, situated, contextual, and resistant to a monolithic, stable definition. The construct is difficult to pin down, even if only for a moment . . . this challenging task, however, is made all the more difficult when our focus shifts from assessing products . . . to tracing the trajectory of one's development over time and across contexts.

Knoch (2016) argued that measurement of writing development in the language assessment and second language acquisition literature has often focused too narrowly on the development of writing within one text type. Studies have typically either evaluated development cross-sectionally by, for example, comparing writers at different score levels in proficiency tests (e.g., Banerjee *et al*., 2007; Cumming *et al*., 2005; Knoch *et al*., 2014) or class levels (e.g., Henry, 1996; Yau, 1991), or they have documented the development of smaller groups of learners longitudinally on one text type: for example, large-scale standardized tests, such as the IELTS (e.g., Elder and O'Loughlin, 2003; O'Loughlin and Arkoudis, 2009). Knoch proposed instead that writing development should be measured at different levels (as depicted in Figure 14.2).

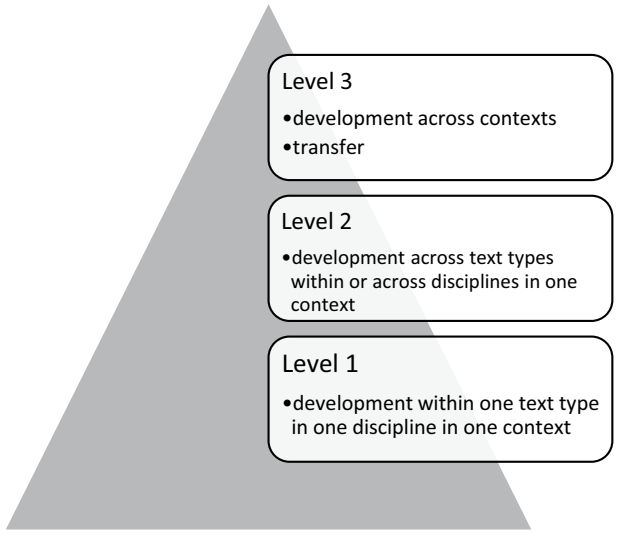


Figure 14.2 Levels of writing development

Level 1 writing development relates to development within one text type in one discipline within one context. The majority of work on writing development has narrowly focused on Level 1 development. Level 2 development focuses on writing development across text types within or across disciplines within one context. For example, students enrolled in higher education would arguably be exposed to various text types within one or more disciplines during their time at university. Level 3 development includes development across contexts and also includes the notion of transfer. For example, when an engineering graduate enters the workplace for the first time, they will be exposed to a variety of new writing practices and tasks (Knoch *et al*., 2016), and therefore, they are expected to develop their writing ability to produce new text types and acquire new writing practices, such as collaborative writing. In such situations, the notion of "transfer" is valued, in which writers are able to transfer skills from one text type and context onto other text types and contexts.

Further work is required in establishing how assessments could be developed to operationalize development at the different levels. One method to gather data on writing development in classroom contexts is through the use of portfolios. A portfolio constitutes a collection of learners' writing samples, typically collected and purposefully collated over a specified period of time, such as a teaching term (Danielson and Abrutyn, 1997; Lam, 2018). This approach is attractive as it provides a more diverse picture of a learner's writing than typical one-shot assessment designs (Weigle, 2002). Learners are able to collate, under the guidance of their teachers, writing samples from one or more genres that were produced within the teaching period. Learners may also include several drafts of the same piece of work and can include a personal reflection on the writing process. The assessment of portfolios is mostly delayed, when learners submit the collection of samples for evaluation (Hamp-Lyons and Condon, 2000). However, portfolio criteria rarely evaluate the "development" of writing in the time period but are generally focused on the product and the processes. While the term *developmental portfolio* exists, this does not seem to have been adopted in L2 writing or assessment practices. A further difficulty faced by assessment developers interested in development at the three levels described here is how to measure progress. It is, for example, not clear whether scoring criteria can be created that are sensitive enough to measure progress in this manner. Further work is clearly needed.

As this chapter has shown, writing assessment has evolved historically. Recent developments in technology have opened up new avenues for writing assessment that make it more feasible to score writing samples automatically. This also provides us with the opportunity to more closely align scholarship in written corrective feedback and language assessment and advance automatic feedback systems. The assessment of collaborative writing, as well as writing development, are areas that will be of interest to researchers in coming years.

Further reading

Weigle, S. C. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

This book is one of the first volumes entirely devoted to the assessment of writing. It presents a detailed, accessible introduction to various important issues, including constructs, scoring, and task development.

Crusan, D. (2010). *Assessment in the Second Language Writing Classroom*. Ann Arbor, MI: University of Michigan Press.

Deborah Crusan's 2010 book is the only volume devoted to the assessment of writing in classroom contexts. It provides a very accessible introduction for teachers and touches on various assessment types, including portfolio assessment, which is often not discussed in other sources.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing* 18: 7–24.
This journal article presents a good discussion of the construct assessment by automated essay scoring systems by discussing its usefulness to different types of writing assessment. The article was awarded the best article award by the International Language Testing Association (ILTA) in 2013.

Weigle, S. C. (2013). Assessment of writing. In C. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. London, UK: Blackwell Publishing.
This encyclopedia chapter provides an excellent short introduction to the assessment of writing, including the nature of second language writing, different types of writing assessments, tasks, and scoring.

References

Alegria de la Colina, A. and Garcia Mayo, M. P. (2007). Attention to form across collaborative tasks by low-proficiency learners in an EFL setting. In G. Mayo (ed.), *Investigating Tasks in Foreign Language Learning*. Clevedon, UK: Multilingual Matters, 91–116.

Bachman, L. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Bailey, K. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing* 13: 257–79. https://doi.org/10.1177/026553229601300303.

Banerjee, J., Francheschina, F. and Smith, A. M. (2007). *Documenting Features of Written Language Production Typical of Different IELTS Band Score Levels*. IELTS Research Reports, vol. 7. Retrieved from www.ielts.org/research/research-reports/volume-07-report-5.

Bereiter, C. and Scardamalia, M. (1987). *The Psychology of Written Composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Berry, V. (2007). *Personality Differences and Oral Test Performance*. Frankfurt: Peter Lang.

Bitchener, J. and Ferris, D. (2012). *Written Corrective Feedback in Second Language Acquisition*. New York, NY: Routledge.

Bitchener, J. and Knoch, U. (2010). Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing* 19: 207–17. https://doi.org/10.1016/j.jslw.2010.10.002.

Bitchener, J., Young, S. and Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing* 14: 227–58. https://doi.org/10.1016/j.jslw.2005.08.001.

Bonk, W. and Van Moere, A. (2004). *L2 Group Oral Testing: The Influence of Shyness/Outgoingness, Match of Interlocutor's Proficiency Level, and Gender on Indivdiual Scores*. Paper presented at the Language Testing Research Colloquium, Temecula.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing* 26: 341–66. https://doi.org/10.1177/0265532209104666.

Brown, J. D. (2016). *Introducing Needs Analysis and English for Specific Purposes*. Oxford, UK: Routledge.

Byrnes, H. (2014). Theorizing language development at the intersection of 'task' and L2 writing: reconsidering complexity. In H. Byrnes and R. M. Manchon (eds.), *Task-Based Language Learning: Insights from and for L2 Writing*. London, UK: John Benjamins, 79–103.

Chapelle, C., Cotos, E. and Lee, J. Y. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing* 32: 385–405. https://doi.org/10.1177/0265532214565386.

Cheng, L., Sun, Y. and Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching* 48: 436–70. https://doi.org/10.1017/S0261444815000233.

Choi, H. and Iwashita, N. (2016). Interactional behaviours of low-proficiency learners in small group work. In M. Sato and S. Ballinger (eds.), *Peer Interaction and Second Language Learning: Pedagogical Potential and Research Agenda*. Amsterdam and Philadelphia, PA: John Benjamins, 113–34.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg: Council of Europe.

———. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe.

Crusan, D. (2014). Assessing writing. In A. Kunnan (ed.), *The Companion to Language Assessment*. New York, NY: John Wiley & Sons, 1–15.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K. and James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing* 10: 1–75. https://doi.org/10.1016/j.asw.2005.02.001.

Danielson, C. and Abrutyn, L. (1997). *An Introduction to Using Portfolios in the Classroom*. Alexandria: Association for Supervision and Curriculum Development.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing* 26: 367–96. https://doi.org/10.1177/0265532209104667.

———. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33: 117–35. https://doi.org/10.1177/0265532215582282.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing* 18: 7–24. https://doi.org/10.1016/j.asw.2012.10.002.

Dikli, S. and Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing* 22: 1–17. https://doi.org/10.1016/j.asw.2014.03.006.

Eckes, T., Mueller-Karabil, A. and Zimmermann, S. (2016). Assessing writing. In D. Tsagari and J. Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin: Walter de Gruyter, 147–64.

Elder, C. and O'Loughlin, K. (2003). *Investigating the Relationship Between Intensive EAP Training and Band Score Gains on IELTS*. Retrieved from www.ielts.org/research/research-reports/volume-04-report-6.

———. (2007). *ELICOS Language Levels Feasibility Study – Final Report*. Melbourne: University of Melbourne, Language Testing Research Centre.

Ellis, R., Sheen, Y., Murakami, M. and Takashima, H. (2008). The effects of focused and unfocused written corrective feedback in an English as a foreign language context. *System* 36: 353–71. https://doi.org/10.1016/j.system.2008.02.001.

Fernandez Dobao, A. (2012). Collaborative writing tasks in the L2 classroom: Comparing group, pair, and individual work. *Journal of Second Language Writing* 21: 40–58. https://doi.org/10.1016/j.jslw.2011.12.002.

Ferris, D. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing* 19: 6–23. https://doi.org/10.1016/j.asw.2013.09.004.

Fogal, G. C. (2019). Investigating variability in L2 development: Extending a complexity theory perspective on L2 writing studies and authorial voice. *Applied Linguistics* 41: 575–600. https://doi.org/10.1093/applin/amz005.

Fogal, G. C. and Verspoor, M. H. (eds.). (2020). *Complex Dynamic Systems Theory and L2 Writing Development*. London, UK: John Benjamins Publishing Company.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13: 208–38. https://doi.org/10.1177/026553229601300205.

———. (2010). The reification of the common European framework of reference (CEFR) and effect-driven testing. In A. Psyaltou-Joycey and M. Matthaioudakis (eds.), *Advances in Research on Language Acquisition and Teaching*. Thessaloniki: GALA.

Fulcher, G., Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing* 28: 5–29. https://doi.org/10.1177/0265532209359514.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics* 35: 553–74. https://doi.org/10.1093/applin/amt017.

Gebril, A. and Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The relationship between discourse features and proficiency. *Language Assessment Quarterly* 10: 9–27. https://doi.org/10.1080/15434303.2011.642040.

Goldstein, L. M. (2006). Feedback and revision in second language writing: Contextual, teacher, and student variables. In K. Hyland and F. Hyland (eds.), *Feedback in Second Language Writing: Contexts and Issues*. Cambridge, UK: Cambridge University Press, 185–205.

Grabe, W. and Kaplan, R. B. (1996). *Theory and Practice of Writing*. New York, NY: Longman.

Guenette, D. and Lyster, R. (2013). Written corrective feedback and its challenges for pre-service ESL teachers. *Canadian Modern Language Review* 69: 129–53. Retrieved from www.muse.jhu.edu/article/506710.

Gurzynski-Weiss, L. K. (2016). Factors influencing Spanish instructors' in-class feedback decisions. *The Modern Language Journal* 100: 255–75. https://doi.org/10.1111/modl.12314.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (ed.), *Second Language Writing: Research Insights for the Classroom*. New York, NY: Cambridge University Press.

———. (2001). Fourth generation writing assessment. In T. Silva and P. K. Matsuda (eds.), *On Second Language Writing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hamp-Lyons, L. and Condon, W. (2000). *Assessing the Portfolio: Principles for Practice Theory and Research*. Cresskill, NJ: Hampton Press.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy and S. Ransdell (eds.), *The Science of Writing: Theories, Methods, Individual Differences*. Mahwah, NJ, Lawrence Erlbaum Associates.

Hayes, J. R. and Flower, L. S. (1980). Identifying the organization of the writing process. In L. W. Gregg and E. R. Sternberg (eds.), *Cognitive Processes in Writing*. Hillsdale, NJ: Erlbaum, 3–30.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *Modern Language Journal* 80: 309–26. https://doi.org/10.1111/j.1540-4781.1996.tb01613.x.

Hill, K. and McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing* 29: 395–420. https://journals.sagepub.com/doi/pdf/10.1177/0265532211428317.

Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book*. London, UK: Routledge.

Kim, Y. H. (2011). Diagnosing EAP writing ability using the reduced reparameterized unified model. *Language Testing* 28: 509–41. https://doi.org/10.1177/0265532211400860.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy and S. Ransdell (eds.), *The Science of Writing: Theories, Methods, Individual Differences, and Applications*. Mahwah, NJ: Lawrence Erlbaum.

Knoch, U. (2016, November). *Measuring Writing Development: Implications for Research and Pedagogy*. Invited keynote presentation at the conference of the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ), University of Auckland, Auckland.

Knoch, U., Deygers, B. and Khamboonruang, A. (2021). Re-visiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing* [OnlineFirst]. https://doi.org/10.1177/0265532221994052.

Knoch, U. and Macqueen, S. (2020). *Assessing English for Professional Purposes: Language and the Workplace*. London, UK: Routledge.

Knoch, U., Macqueen, S. and O'Hagan, S. (2014). *An Investigation of the Effect of Task Type on the Discourse Produced by Students at Various Score Levels in the TOEFL iBT Writing Test*. TOEFL iBT Report – 23, ETS Research Report RR-14–43. Princeton, NJ: Princeton University Press.

Knoch, U., May, L., Macqueen, S., Pill, J. and Storch, N. (2016). *Transitioning from University to the Workplace: Stakeholder Perceptions of Academic and Professional Writing Demands*. IELTS Research Report 2016/1. Retrieved from www.ielts.org/research/research-reports/online-series-2016-1.

Lam, R. (2018). *Portfolio Assessment for the Teaching and Learning of Writing*. Frankfurt, Germany: Springer.

Lantolf, J. P. and Poehner, M. E. (2004). Dynamic assessment of L2 development: Bringing the past into the future. *Journal of Applied Linguistics* 1: 49–74. Retrieved from https://journals.equinoxpub.com/index.php/JAL/article/view/647.

Leeser, M. J. (2004). Learner proficiency and focus on form during collaborative dialogue. *Language Teaching Research* 8: 55–81. https://doi.org/10.1191/1362168804lr134oa.

Li, J., Link, S. and Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing* 27: 1–18. https://doi.org/10.1016/j.jslw.2014.10.004.

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessments: A longitudinal study of new and experienced raters. *Language Testing* 28: 543–60. https://doi.org/10.1177/0265532211406422.

Long, M. (2005). Methodological issues in learner needs analysis. In M. Long (ed.), *Second Language Needs Analysis*. Cambridge, UK: Cambridge University Press.

Macqueen, S. (2021). Construct in L2 assessments of speaking. In T. Haug, W. Mann and U. Knoch (eds.), *Handbook of Language Assessment Across Modalities*. Oxford, UK: Oxford University Press.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing* 26: 397–421. https://doi.org/10.1177/0265532209104668.

McNamara, T. (1996). *Measuring Second Language Performance*. London and New York, NY: Longman.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: Palgrave Macmillan, 13–103.

Min, H. T. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing* 15: 118–41. https://doi.org/10.1016/j.jslw.2006.01.003.

Montgomery, J. L. and Baker, W. (2007). Teacher-written feedback: Student perceptions, teacher self-assessment, and actual teacher performance. *Journal of Second Language Writing* 16: 82–99. https://doi.org/10.1016/j.jslw.2007.04.002.

Nakatsuhara, F. (2004). *An Investigation into Conversational Styles in Paired Speaking Tests*, MA dissertation. University of Essex, Colchester.

———. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing* 28: 483–508. https://doi.org/10.1177/0265532211398110.

Ockey, G. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing* 26: 161–86. https://doi.org/10.1177/0265532208101005.

O'Loughlin, K. and Arkoudis, S. (2009). *Investigating IELTS Exit Score Gains in Higher Education*. IELTS Research Reports, vol. 10. Retrieved from www.ielts.org/research/research-reports/volume-10-report-3.

Perry, G. (2019). *Comparing Writing and Revisions of Second Language Learners in Computer-Mediated Individual and Collaborative Writing*. MA minor thesis. University of Melbourne, Melbourne.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on attention and awareness in learning. In R. Schmidt (ed.), *Attention and Awareness in Foreign Language Learning*. Honolulu, HI: University of Hawai'i National Foreign Language Resource Center, 1–63.

———. (2001). Attention. In P. Robinson (ed.), *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Shaw, S. (2004). Creating a virtual community of assessment practice: Towards 'on-line' rater reliability. *Cambridge Research Notes* 15: 18–20.

Sheen, Y. (2007). The effect of focused written corrective feedback and language aptitude on ESL learners' acquisition of articles. *TESOL Quarterly* 41: 255–83. https://doi.org/10.1002/j.1545-7249.2007.tb00059.x.

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing* 17: 81–91. https://doi.org/10.1016/j.asw.2012.02.001.

Storch, N. (2017). Implementing and assessing collaborative writing activities in EAP classes. In J. Bitchener, N. Storch and R. Wette (eds.), *Teaching Writing for Academic Purposes to Multilingual Students*. New York, NY: Routledge, 130–42.

Storch, N. and Aldosari, A. (2013). Pairing learners in pair work activity. *Language Teaching Research* 17: 31–48. https://doi.org/10.1177/1362168812457530.

Swain, M. and Lapkin, S. (1995). Problems of output and the cognitive process they generate: A step towards second language learning. *Applied Linguistics* 16: 371–91. https://doi.org/10.1093/applin/16.3.371.

Turner, C. E. and Purpura, J. (2015). Learning-oriented assessment in the classroom. In D. Tsagari and J. Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin, Germany and Boston, MA: DeGruyter Mouton.

Van Beuningen, C. G., De Jong, N. H. and Kuiken, F. (2008). The effect of direct and indirect corrective feedback on L2 learners' written accuracy. *ITL International Journal of Applied Linguistics* 156: 279–96. https://doi.org/10.2143/ITL.156.0.2034439.

Verspoor, M. H. and Behrens, H. (2011). Dynamic systems theory and a usage-based approach to second language development. In M. H. Verspoor, K. de Bot and W. Lowie (eds.), *A Dynamic Approach to Second Language Development: Methods & Techniques*. London, UK: John Benjamins, 25–38.

Verspoor, M. H., Lowrie, W. and van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal* 92: 214–31. https://doi.org/10.1111/j.1540-4781.2008.00715.x.

Wagner, M. (2015). *The Centrality of Cognitively Diagnostic Assessment for Advancing Secondary School ESL Students' Writing: A Mixed Methods Study*. ProQuest Central; ProQuest Dissertations & Theses Global; Social Science Premium Collection. 1719259603. Retrieved from http://search.proquest.com.ezp.lib.unimelb.edu.au/docview/1719259603?accountid=12372.

Wall, D. (2012). Washback. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. New York, NY: Routledge, 79–72.

Watts, A. (2006). *Fostering Communities of Practice: A Rationale for Developing the Use of New Technologies in Support of Raters*. Cambridge, UK: Cambridge University Press.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing* 15: 263–87. https://doi.org/10.1177/026553229801500205.

———. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing* 6: 145–78. https://doi.org/10.1016/S1075-2935(00)00010-6.

———. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning and Identity*. Cambridge, UK: Cambridge University Press.

Wenger, E. and Snyder, W. (2000). Communities of practice: The organizational frontier. *Harvard Business Review* 78: 139–45.

Wenger, E., Snyder, W. and McDermott, R. (2002). *Cultivating Communities of Practice: A Guide to Managing Knowledge*. Boston, MA: Harvard Business School Press.

Williamson, D., Xi, X. and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice* 31: 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication* 50: 483–503. https://doi.org/10.2307/358862.

Yau, M. (1991). The role of language factors in second language writing. In L. Malave and G. Duquette (eds.), *Language, Culture and Cognition: A Collection of Studies in First and Second Language Acquisition*. Clevedon, UK: Multilingual Matters, 266–83.

Zhang, Z. and Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing* 36: 90–102. https://doi.org/10.1016/j.asw.2018.02.004.

Zupanc, K. and Bosnic, Z. (2015). Advances in the field of automated essay evaluation. *Informatica* 39: 383–95.

15
Assessing reading

Tineke Brunfaut

Introduction

When tasked with assessing second or foreign language (L2) reading, two fundamental questions arise: (1) What is reading, and, specifically, what does reading in an L2 entail? and (2) How can we assess reading comprehension? I begin this chapter with a concise overview of reading theory, starting with insights into reading in a first language (L1) and describing how this has been expanded on for the context of L2 reading. With respect to the second question, because reading is a receptive skill and comprehension is often not directly observable, various techniques have been developed to reveal text processing and understanding. I therefore give an overview of task formats that are often used in L2 reading assessments. Text comprehension, however, can be affected by a variety of factors arising from the specific features of the text and the questions asked, as well as the characteristics of the reader. In the section "Critical issues and topics," I provide a summary of empirical findings on such factors, since insights into these can help inform test design decisions. The next three sections centre on aspects of validity and validation in testing L2 reading. First, I highlight important recent developments in L2 reading assessment, enabled by technological advances but fundamentally driven by validity rationales. This is followed with an overview of research methods prevalent in reading test validation research. Then, I provide practical recommendations for the construction and evaluation of L2 reading tests that will benefit test quality. In the final section of this chaper, I offer suggestions to broaden the scope of our practices in the testing of L2 reading – in terms of both what and how we assess reading in an L2. In addition, I signal research opportunities that may help optimize L2 reading (test) theory.

Historical perspectives

What is reading?

Early research in reading primarily concentrated on reading in the first language, mostly situated within the field of psychology. Some of the first L1 reading theories differed in their

 DOI: 10.4324/9781003220756-20

emphasis on *bottom-up processing* – whereby readers were thought to sequentially proceed from recognising the print, decoding it into sounds, and recognising words to decoding meaning (Gough, 1972) – versus *top-down processing* – whereby readers were thought to largely draw on their knowledge schemata to predict the text's meaning from limited textual information (Goodman, 1967; Smith, 1971). As more empirical findings became available, scholars recognised that reading is in fact a very interactive activity with different processes operating simultaneously and in parallel (e.g., Kintsch's (1988) *construction-integration model*). Stanovich's (1980) *interactive compensatory model*, for instance, posited that processing problems can be offset by bringing in other processes such as compensating for decoding flaws by making use of context information or knowledge. A different strand of reading research has focused less on the process and more on the product of reading, taking a *componential approach*. Here, the work by Tunmer and colleagues has been influential. Tunmer and Chapman's (2012) *Modified Simple View of Reading* argues that general language comprehension skills and word decoding accuracy/efficiency together predict reading comprehension. One of the most recent models, which brings together various strands of existing L1 reading research, is Francis *et al*.'s (2018) *Complete View of Reading* (CVRi). This dynamically interactive model integrates aspects of componential models (Tunmer and Chapman's work) with description of how textual features impact comprehension (Kintsch and colleagues' work) and adds a developmental angle to capture reading across ages. In this manner, they model the interaction between the reader, the text, and reading processes.

Initial theories on reading in a second or foreign language were strongly influenced by the earlier L1 literature. Only from around 1980 onward did L2 research emerge, valuing and retaining many insights from L1 research but also making a case for L2-specific research due to important differences between reading in a first and a second/foreign language. One particular realisation was that the target-language knowledge base of L2 readers who are in the process of acquiring the L2 is much more limited than that of L1 readers. Other questions arose around issues such as L2 readers' skills in their own L1 and background/cultural knowledge differences. This led to the formulation of the *linguistic interdependence hypothesis* (Cummins, 1979), which argues that L1 and L2 reading skills are highly interlinked, with L1 reading competencies affecting L2 reading ability. An alternative view was the *linguistic threshold hypothesis*, which reasoned that L2 readers first need to reach a certain level of language knowledge in the L2 before they can draw on their L1 reading skills to achieve L2 reading comprehension. This debate is represented well in Alderson's (1984) seminal paper "Reading in a foreign language: a reading problem or a language problem?" The accumulating body of empirical research, however, more consistently supports the linguistic interdependence hypothesis, with more conflicting findings on the linguistic threshold hypothesis (Pae, 2019).

One well-known conceptualisation of L2 reading in language testing is the TOEFL 2000 Reading Framework (Enright *et al*., 2000), which takes a *reader-purpose perspective*. The key consideration is the type of purposes people have for reading texts. The framework builds on work conducted in L1 reading: Carver's (1997) distinction between rauding (reading for general, main points comprehension) and reading to learn (establishing a text organisation representation distinguishing between main points and supporting details); Guthrie's (1988) search reading (skimming and scanning for specific types of information); and Perfetti's (1997) multi-text reading (intertextual meaning building). Since the TOEFL aims to test academic reading abilities in English as a foreign language, the target construct was consequently defined as reading to find information, reading for basic comprehension, reading to learn, and reading to integrate information across multiple texts – with each of these

requiring "a combination of word recognition/processing efficiency and comprehension abilities" (Enright *et al*., 2000: 5). At the same time, the framework was also informed by research insights on differences between L1 and L2 reading, and it recognised the following factors as specifically influencing L2 reading: transfer of L1 reading skills and strategies, facilitation resulting from structural similarities between L1 and L2, cross-linguistic interactions during L2 reading, and processing constraints due to more limited L2 linguistic knowledge (Enright *et al*., 2000: 7–13). Then, by applying this reader-purpose perspective of L2 reading to the broader TOEFL 2000 Framework (Jamieson *et al*., 1999), the reading framework for the TOEFL 2000 test was developed (see Enright *et al*., 2000: 14–39). In turn, the reading framework formed the basis for designing the TOEFL 2000 reading test (later the TOEFL iBT reading test), with three of the four reading purposes operationalised in items targeting basic comprehension, inferencing, and reading to learn. The new reading test also included new task types and longer, more complex texts than earlier test versions (Chapelle *et al*., 2008).

Another, currently influential L2 reading model in language testing is Khalifa and Weir's (2009: 43) *cognitive processing model* for reading comprehension, which they specifically developed to bridge disjoints between reading theories and test practices. The model comprises three main components. First, the *central processing core* describes readers' cognitive processes, which work together to achieve text comprehension. It encompasses the lower-level processes of word recognition, lexical access, syntactic parsing, and establishing propositional meaning, which can all become highly automatized. Word recognition entails orthographic processing (recognising the printed symbols), phonological processing (sounding out the words in mind), and morphological processing (utilizing information on expected grammatical forms). Information on word form and meaning are retrieved from the reader's mental lexicon during lexical access, and words are integrated at the clausal level during syntactic parsing (while decoding the grammatical information in the text). Propositional meaning is established by converting clauses and sentences into units of meaning. In addition, the central processing core involves higher-level processing too. Namely, readers may infer meaning by drawing on world knowledge, topic knowledge, or what they have understood from the text so far. They also build a mental model of the text by integrating individual propositions into their overall meaning framework of the text. This may then lead to the creation of a text-level representation and differentiation between the main ideas and the details of the text. Finally, meaning from an individual text can be supplemented with information from other sources and result in an intertextual representation.

The second key component in this model is the reader's *knowledge base*, which feeds into the ongoing cognitive processing. Lower-level processes ultimately rely on the reader's orthographical, phonological, morphological, lexical, and syntactic knowledge of the language. Higher-level processing draws importantly on the reader's knowledge of the world and the topic and on knowledge of text structure, genre, etc.

The third component concerns *metacognitive activity*. This involves setting goals and determining what reading approach is needed to complete a task. Khalifa and Weir describe this along the dimensions of a) local reading at the sentence and clause level versus global reading to comprehend the text beyond sentence and clause level, and b) careful reading to achieve complete and detailed meaning versus expeditious reading to efficiently target only very specific information. In addition, metacognitive processing engages readers in monitoring their own comprehension levels (in correspondence with the reading goals set) and in remediating comprehension problems.

In their own research, Khalifa and Weir (2009) applied the model to retrospectively describe and evaluate the reading sections of the Cambridge ESOL test suite. The model has also formed the basis for the development of new L2 reading tests such as the British Council's Aptis reading test (O'Sullivan and Dunlea, 2015). In addition, its usefulness for research on the cognitive validity of L2 reading tests has been demonstrated in, for example, Bax and Chan (2016) and Brunfaut and McCray (2015a).

How can we assess reading?

While the physical act of reading might be directly observable, a reader's level of understanding is less straightforward to observe. Therefore, texts are accompanied by comprehension-oriented questions or activities to elicit evidence of understanding. These questions can be categorised into (1) selected-response items, whereby test takers need to demonstrate textual comprehension through choosing from a set of options, and (2) constructed-response items, in which readers need to write down an answer or orally respond to show their level of understanding; this can take the shape of a very short, one-word answer up to a lengthy productive performance. Both approaches have advantages and disadvantages. For example, while selected-response items might be easy and quick to mark, writing such items is time consuming and challenging to do well. In addition, they risk inflating scores through guessing. Constructed-response items are less demanding to write but more challenging to rate, and they might test aspects of writing/speaking due to the response format.

Early L2 reading tests heavily made use of the *multiple-choice format* (e.g., Mundy, 1968), and this remains a highly popular task type due to its versatility in what can be tested. In the 1970s, the *cloze procedure* (Oller, 1973) – whereby each n^{th} word is deleted from a text – was promoted as a format for testing L2 reading comprehension. Editing tasks, in which the reader needs to detect errors in a text (Alderson, 2000: 224), and c-tests, in which the second half of every n^{th} word is deleted (Raatz and Klein-Braley, 1981), also featured in reading assessment. Questions were raised, however, about the extent to which such procedures test reading as opposed to general language proficiency (for a review, see McCray and Brunfaut, 2018), and therefore, these formats no longer prevail in L2 reading assessment. Over time, several variations on the cloze format have been introduced to have control over the words deleted and thus over what is being tested: for instance, open/banked/multiple-choice *gap-fill tasks*. Other frequently used formats in L2 reading assessment are *matching* (e.g., matching headings to paragraphs or descriptions to short texts), *true-false-not given* (whether a given statement aligns with, contradicts, or is not covered in the text), and *information transfer* (e.g., completing a table based on information from the text). Often-employed constructed-response formats in L2 reading assessment are *short-answer questions*, which are also very versatile, and *sentence-completion tasks* (e.g., completing a gap in a statement based on the text's information). A key development is that the majority of L2 reading tests now include a variety of item formats, which enables the testing of a wider range of aspects of reading and reduces potential format effects.

Critical issues and topics

L2 reading comprehension accuracy (as demonstrated in test performance) can be influenced by a range of factors associated with the characteristics of the texts, the items, the test taker, and the interactions between these. Awareness of these factors is important for L2

reading test design and use, to ensure that the assessment instruments are appropriate for the test purpose and test taker population and that the resulting scores represent the intended reading behaviour.

With respect to the impact of *text characteristics*, many linguistic, topic, genre, and length features of texts have been hypothesised to influence L2 reading performance, and a number of these have been empirically confirmed. For example, in a study on the TOEFL reading section, Freedle and Kostin (1993) identified significant, albeit weak correlations between reading item difficulty and the lexical characteristics of the texts, sentence/paragraph/passage length, frequency of referentials within the texts, occurrence of fronted structures in the texts, and the concreteness of the information in the texts. Furthermore, the type of rhetorical organisation of the texts was found to be a significant predictor of the difficulty of the items for the L2 test takers.

As regards question-related L2 reading differences, these encompass *item characteristics* as well as *reading purposes as directed by the task*. For example, a meta-analysis by In'nami and Koizumi (2009) suggested that there is generally a limited effect of the item format (in this case, multiple-choice versus open-ended questions) on L2 reading test performance. They did find, however, that multiple-choice tasks were easier than open-ended tasks in a number of circumstances (i.e., with stem-equivalent items or for test takers with high L2 proficiency). In another example, Löwenadler (2019) observed a significant difference between the extent to which multiple-choice gap-fill tasks versus traditional multiple-choice content questions measured L2-specific, lower-level reading abilities, with the gap-fill tasks assessing these more.

In terms of the *interaction between items and the text* in L2 reading tests, Freedle and Kostin (1993) found that the prominence or location in the text of the information that is needed to answer the item correctly impacted the difficulty of TOEFL reading items, in particular when the items required reading for main ideas or inferencing. The extent of lexical overlap between key information in the text and the correct answer, as well as the length of sentences central to the answer, were also shown to be significant predictors in their study. McCray (2014) similarly confirmed the prominence in the text of the necessary information as a predictive factor, in that case for PTE Academic reading items.

With respect to the *test taker*, Jeon and Yamashita's (2014) meta-analysis of research on ten different cognitive and linguistic reader characteristics showed that readers' L2 grammatical, vocabulary, and morphological knowledge and their L2 listening comprehension and decoding abilities strongly associated with their L2 reading comprehension accuracy. Medium correlations were found between L2 reading comprehension and readers' L1 reading comprehension, L2 orthographic knowledge, L2 phonological awareness, working memory, and, to a lesser extent, their metacognition. In some cases, readers' ages and the linguistic distance between their L1 and L2, as well as characteristics of the reading tasks, moderated the relationship. Kremmel *et al*. (2017) in addition found that L2 test takers' phraseological knowledge accounted for a large amount of the variance in L2 reading performance.

Current contributions and research

In 2009, in his leading volume on reading in a second language, Grabe expressed concerns about the extent to which standardized L2 reading tests at that point assessed core component abilities of L2 reading. Among these components, he listed search processes, text-structure awareness and discourse organisation, summarisation abilities, synthesis skills,

and evaluation and critical reading (Grabe, 2009: 357). In recent years, we have seen at least two important developments in the testing of L2 reading – one largely enabled by practical innovations, the other driven by construct and authenticity rationales. Namely, first, the vast speed of technological advances and their global spread have meant that several language test providers have developed and brought into circulation alternative modes of delivery for existing tests or created entirely new computer-based/online tests (see also Sawaki, this volume). For the testing of reading, these technologies have opened up ways to broaden what we test. For example, in digitally delivered tests, individual tasks can be technically time-restricted, even in large-scale administration settings. This enables the testing of expeditious, search-reading approaches such as skimming a text to gain an overall idea of its content or scanning it for a specific piece of information. This is opposed to careful reading, which involves more detailed, complete, and linear processing of the entire text, which requires more time. Of course, research is needed to determine exactly what time limit will be reasonable and suitable to test search reading with a particular task (and not careful reading). Another advantage of a digital mode lies in task types such as sentence/paragraph re-ordering, which aim to assess understanding of text organisation and cohesion. Using drag-and-drop technology, the test taker can reconstruct the text order to its original and easily reread the recreated text for cohesion. Examples of this can be found in the computer-based/online PTE Academic (https://pearsonpte.com/the-test/format/english-reading/re-order-paragraphs/) and Aptis (www.britishcouncil.org/exam/aptis/reading) reading tests.

Second, aiming to better reflect the integrated nature of language as used in "the real world," many tests now include integrated reading tasks, which require the test taker to read one or more texts and provide a written or spoken response (see also Plakans, this volume). This can, for example, be a summary of the reading input(s) (e.g., PTE Academic – https://pearsonpte.com/the-test/format/english-speaking-writing/summarize-written-text/) or a synthesis of multiple texts to respond to a writing prompt which requires further argumentation on the topic (e.g., the ISE exams' Reading & Writing Task 3 – www.trinitycollege.com/qualifications/english-language/ISE). This widening of the tested construct has also been paralleled with research on integrated testing – especially reading-into-writing – including studies on differences between independent and integrated task performances and studies on integrated performances in terms of scores, task completion processes, and discourse produced (e.g., Chan, 2018; Plakans and Gebril, 2013).

Another area in which there has been an increasing volume of research contributions to L2 reading assessment concerns validation. On the one hand, this has been stimulated by the development of more practical test validation frameworks such as the argument-based approach (see Kane, this volume) and the socio-cognitive approach (Weir, 2005). On the other hand, innovations in research methodology have also enabled and enriched validation research on L2 reading assessment, as discussed in the next section.

Main research methods

A key impetus for research on the testing of L2 reading relates to validity concerns. It is imperative that test developers obtain and provide evidence for the claims they make about their reading tests in light of the test's intended use and suitability for the test taker population (see Chapelle, this volume). Regardless of the validation approach taken, a multitude of data will need to be collected to comprehensively support the validity of inferences based on test takers' scores on the reading test. Next, I outline a number of prevailing research methods in reading test validation research.

Fundamental to the evaluation of any reading test is the running of *descriptive statistics* to establish how the test as a whole, as well as individual items, function. This includes calculating (1) test-level measures of central tendency or score clustering (mean, median, mode) and dispersion or score spread (range, standard deviation) and (2) item-level measures of difficulty (facility value) and discriminatory power (discrimination index/point-biserial correlation). The resulting figures then need to be examined with reference to guidelines provided in the literature (see, e.g., Green, 2013; Brown, this volume) and, importantly, interpreted in relation to the test's purpose and test-taker population (see, e.g., Alderson *et al*., 1995). For instance, while we might expect that a test aiming to assess L2 learners' reading achievement on a course will show many easy items and a negative skew (indicating high reading attainment), such a distribution would not be helpful if it was a placement test. It is also vital that the statistical analysis is conducted alongside, and integrated with, a *qualitative analysis* of the test and its items. This often sheds light on why an item was identified as problematic based on its statistical properties and can help determine ways to improve the item (or drop it altogether).

More sophisticated statistical analyses have, for example, been used in studies exploring whether L2 reading items function differently for test-taker groups with different characteristics while controlling for their L2 reading ability levels. The main rationale for such research is to help identify potentially biased items, which provide groups of L2 learners of similar ability with unequal opportunities to get the item correct and thus pose a threat to validity. Such *differential item functioning* (DIF) studies make use of Rasch-based statistical procedures or classical test theory analyses such as logistic regression or the Mantel-Haenszel method (see Kamata and Vaughn, 2004). It should be emphasized, however, that the presence of DIF – as identified with statistical analysis – does not, by definition, constitute bias; additional *expert content analysis* of any DIF items is necessary to determine whether or not the DIF sources are related to the intended reading construct (which would simply mean item impact) or are unrelated (which would constitute item bias). The main test-taker characteristics for which DIF has been looked into in reading tests are gender (see, e.g., Pae, 2012), L1 background (see, e.g., Abbott, 2007; Koo *et al*., 2013), and academic background (see, e.g., Pae, 2004; Runnels, 2013).

Another important strand of research on the testing of L2 reading has focused on how test takers arrive at the answers they give to reading items. This line of research developed as a result of increasing awareness that (1) expert qualitative analyses of items ultimately tell us "what test constructors believe an item to be testing" (Alderson, 2000: 97), which may or may not relate to how test takers solved the item, and (2) post-hoc quantitative studies, such as those using factor analysis to identify distinct components underlying reading test performance, exclusively focus on the reading product (versus the reading process) and show a number of weaknesses in their assumptions (for a discussion, see Chapter 3 in Khalifa and Weir, 2009). More direct insights into test takers' processes and strategy use while completing a reading test are considered vital, however, to shed light on a test's cognitive validity (as defined by O'Sullivan and Weir, 2011; see also Cohen, this volume). Consequently, a body of research has developed which adopts *verbal protocols* as part of its methodology (Green, 1998). In such studies, the test taker is asked to say out loud what they are reading and thinking when completing the reading test. In some studies, this is done *while* doing the reading test; this is termed a concurrent verbal report. In other studies, it is done immediately *following* the event; this is called a retrospective verbal report. The resulting audio-/video-recorded think alouds are then transcribed and coded by the researcher for reading processes and strategies. Examples of studies which employed verbal

protocols to investigate L2 reading tests are Anderson *et al*. (1991), Phakiti (2003), and Rupp *et al*. (2006). However, while the use of this method has led to valuable insights into test takers' processing, two main limitations have also been pointed out which may affect the quality and validity of the data gathered with verbal protocols. These concern (1) the risk of reactivity – changes in test takers' thought processes and length of processing due to the act of verbalizing – and (2) the risk of veridicality – inaccurate reflections of the processes due to omissions or additions when saying things out loud (Barkaoui, 2011; Bowles, 2010a, 2010b). Alternative methods such as *introspective questionnaires* administered after a reading test similarly suffer from potential incompleteness of reporting due to the time lapse or because some of the processes or strategies actually used by test takers are not pre-listed. Questionnaire data also essentially reflect test takers' perceptions of their reading test processing, which may or may not represent their actual processing.

In search of ways to minimize some of these methodological limitations, researchers have started exploring the potential of *eye-tracking* methodology to gain insights into test taker processing while completing L2 reading tests (Winke, 2014). The use of this method, whereby readers' eye movements are recorded during task completion, is based on the assumption that there is a close link between the point in a text our eyes are looking at and what the object of our attention is at that time (Rayner, 1998). By means of analyses of the frequency, length, and duration of the points at which our eyes are resting (called fixations), of the back- and forward movements our eyes are making while completing the reading test (called saccades), and of larger eye movement transitions between different parts of the task (e.g., instructions, text, items), inferences are made about test takers' cognitive processing during test completion. Studies such as Bax (2013), Brunfaut and McCray (2015a), and McCray and Brunfaut (2018) have shown that this can lead to vital validation information – in these cases, for the reading sections of the IELTS, Aptis, and PTE Academic tests, respectively. It is often useful to combine eye-tracking with *stimulated recalls* (Gass and Mackey, 2017), in which the test taker is shown an on-screen recording with eye movements taken during task completion in order to help them recall and verbally report their processing immediately following task completion (see, e.g., Brunfaut and McCray, 2015a). This allows for data triangulation, thus providing the researcher with more confidence in their conclusions from the eye-tracking analyses, and it also compensates for limitations of eye-tracking methodology (for example, with respect to gathering insights on inferencing while reading).

Recommendations for practice

Probably the most fundamental recommendation for assessing L2 reading is to ensure *construct representation*: namely, that the reading assessment one uses or develops effectively measures key and relevant aspects of reading and that it does so in a sufficiently broad manner. In this way, the intended reading construct will account for learners' scores on the assessment, and the test results will be interpretable. In practice, this means that anyone involved in the selection or development of reading tests needs to start from a sound understanding of what reading in a second/foreign language encompasses and what factors may influence it (see, e.g., Alderson, 2000: 1–84; Grabe and Stoller, 2020: 1–61). What constitutes relevant aspects of reading will depend on the purpose of the assessment and the population being assessed and thus needs to be determined against the specific assessment context. For example, the assessment of decoding and lower-level reading processes is likely to be highly relevant in the context of young, novice L2 readers, whereas the testing

of intertextual representation may not be as relevant for this age group at this level of L2 proficiency.

The next important step is to identify a *suitable approach to test the intended aspects of reading*. This includes determining which item types to adopt in the test. It will be meaningful to consider the general advantages and disadvantages of different item types, as well as challenges associated with writing such items. In addition, it will be equally useful and important to consider to what extent a particular item type lends itself to testing specific aspects of reading. While most item types are reasonably versatile, matching tasks, for example, tend to be especially useful when aiming to test main ideas (for example, through matching headings). True-false-not given tasks, on the other hand, can be very difficult to craft if the aim is to test inferencing since there will often be debate over whether something is implied and can thus be inferred or is "not given" in a text. A traditional multiple-choice format can often work better in that regard.

Another crucial recommendation is to try and keep any potential *construct-irrelevant variance* to a minimum: that is, to prevent factors other than the intended reading processes or skills to affect scores on the assessment. For example, using a range of item types in a reading test may reduce method effects. Also, taking great care during item writing can help avoid construct-irrelevant variance. Burton *et al*. (1991), for example, exemplify several ways in which guessing risks can be reduced through careful multiple-choice item design. This includes making all options within an item structurally and linguistically similar. Another important reason for careful item writing is that well-intended multiple-choice items can easily turn into the testing of comprehension of the multiple-choice options rather than the reading input text if the options are linguistically or cognitively complex. Similarly, if not well designed, short-answer questions may shift the construct to a test of language knowledge or of writing – for example, if a correct response requires (complex) grammatical transformation of information from the reading input or if understanding can only be demonstrated through a lengthier written response. A further consideration is the extent to which background knowledge enables giving the correct answer. Identifying such flaws in items should be carried out at an early stage in test design through item prototyping (see Nissan and Park, this volume).

When developing reading tests, due attention needs to be given to *written-text sourcing*. Texts need to be selected so that they meet the test specifications in terms of test purpose, suitability for the population, target language level, text genre, style, length, and topic. In addition, it will be useful to consider whether a text lends itself to item development. For example, from a construct perspective, if the aim is to evaluate L2 readers' comprehension of main ideas, a multi-paragraph text organized around distinct and important points will enable the development of a relevant task. Or, from an item-format perspective, if a set of multiple-choice items will need to be created, a text which includes many information points will enable the creation of plausible distractor options.

A final important question for item writers concerns what *information in the text to target*. This relates to the so-called meaning-relativity problem, whereby the meaning extracted from the text may differ by reader. As discussed by Sarig (1989: 80), this can lead to the extreme position that "there is no such thing as 'correct' interpretation." This would make the assessment of reading virtually unattainable as it would be impossible to define the correct answer to a reading item. A more workable view is that "[t]he text itself contains a limiting constraint on the infinite range of potential interpretations" (Sarig, 1989: 81) and that we can "[accept] certain readings of a text as 'truer' than others" (Sarig, 1989: 82). The latter view recognises that it is reasonable to expect that many different readers will, for

example, identify the same parts of a text as important. The challenge for item writers is then to establish a "commonly acceptable" textual meaning and to target this in item design, in order to ensure fair testing of meaning construction (as defined by Sarig, 1989). A practical technique that can assist item writers with this is called "textmapping" (see Green, 2017: 57–81). This method is based on the principle of "consensus" on meaning (i.e., a minimum of everyone except one in a group of at least four readers). The common text points then constitute potential target information for items.

Future directions

While there is now a solid body of research available on the construct, as well as the testing, of L2 reading, several theoretical and practical issues remain. For example, in a relatively recent overview, Grabe and Jiang (2014) identified the following as key challenges for reading assessment: (1) how to capture the complexity and multi-componential nature of the reading construct within an assessment and (2) how to elicit and capture, within an assessment context, "natural" reading processes and strategies used in non-testing contexts. One emerging strand of work may be particularly promising in this regard: namely, research and development around digital innovations.

As discussed in the section "Current contributions and research," computer-based/online testing is already broadening the range of reading processes tested, and this will presumably expand further. For example, digital test modes seem to have potential for measuring reading fluency and reading rate as part of assessments. With respect to the second challenge, an interesting avenue concerns the potential of eye-tracking technology beyond its current use as a research method. As the technology is rapidly becoming more sophisticated as well as affordable, it becomes more practicable. It might offer opportunities to complement the currently controlled and test-directed forms of L2 reading in assessment contexts. For example, it might become possible to gain some insights into L2 readers' abilities by collecting eye movement observation data from more "natural" or "authentic" reading: that is, without an accompanying task and/or in non-assessment contexts. Some initial support for this idea can be found in recent research which indicated that a selection of eye movement measures can explain part of the variance in English L2 reading ability (e.g., Brunfaut and McCray, 2015b). Of course, we should not overstate any possible potential of eye tracking for evaluating L2 reading ability or make broad and simple claims in this regard since this methodology is restricted in terms of what component abilities can be captured (e.g., eye movements do not provide information on processes such as inferencing). Nevertheless, it seems worthwhile to explore in future research whether or not eye tracking might have some potential as a complementary source of information, in what ways, and to what extent. All these digital innovations might furthermore increase the diagnostic potential of L2 reading assessments through the inclusion of automated reading measures and the increasing range of reading processes being captured.

Another area for further research relates to the integrated testing of reading. While the existing body of research has provided ample evidence of the context and cognitive validity of such testing approaches (e.g., Chan, 2018; Chan *et al*., 2014; Plakans and Gebril, 2012), there is a scarcity of research related to their scoring validity. On the one hand, generic rating scales which primarily focus on the productive skill (writing/speaking) are often used to mark integrated-task performances (Chan, 2013). On the other hand, it can be challenging for raters to pay attention to aspects of performance content, organisation, and language, as well as to source use, synthesis, and textual borrowing practices (Gebril and Plakans,

2014; Wang *et al*., 2017; Weigle and Montee, 2012). These issues may affect both scoring and construct validity. Thus, more research on rating processes and rating scales used in the integrated testing of reading is recommended.

Finally, further research is needed into L2 reading models of text comprehension. Khalifa and Weir's (2009) model, for example, brought together a variety of prior research through its combined focus on the reader (knowledge base), reading processes (central processing core), and task (metacognitive activity). At the same time, the fact that the existing research base primarily results from older adolescent and adult reading means that current L2 reading theories do not model reading across ages. Young L2 readers, however, are a growing test taker population, especially for English as a foreign language, as a result of the spread of digital media and increases in pedagogic approaches such as EMI (English medium instruction) and CLIL (content and language integrated learning). Apart from learning the L2, young learners are typically developing their cognitive abilities as well as their L1 reading skills in crucial manners at the same time. In this sense, Francis *et al*.'s (2018) CVRi, introduced in the subsection "What is reading?," seems a meaningful model to explore in the context of (assessing) L2 reading since it incorporates developmental aspects. This also suggests that more research on the impact of individual differences (e.g., working memory, self-regulation, age) would be particularly meaningful (see, e.g., Brunfaut *et al*., 2021).

Further reading

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.
This remains one of the most comprehensive handbooks available on the testing of L2 reading, offering both theoretical and practical insights. Starting with an accessible review of the construct of reading and key factors that impact it, Alderson then discusses implications for the testing of reading. The handbook describes different approaches to operationalising reading constructs and designing reading tests, thereby giving due attention to the reader, the target language use domain, text and task characteristics, and test purposes. Practitioners might particularly value the overview of potential task types for assessing L2 reading in Chapter 7, including the many example tasks and the outline of their pros and cons.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
This resource provides an example of a framework and guidelines for validating reading assessments and also illustrates, by means of a widely used reading test, how this can be done in practice. Namely, Chapelle *et al*. demonstrate the application of the argument-based validation approach to the TOEFL reading test (as well as other skills sections of this test).

Grabe, W. and Stoller, F. L. (2020). *Teaching and Researching Reading*, 3rd edn. Oxon and New York, NY: Routledge.
Now in its third edition, Grabe and Stoller's volume is a key resource for anyone working in the area of second/foreign language reading, whether as a researcher or a practitioner. For language testers, Parts I and II are must-reads; these provide in-depth insights into L1 and L2 reading from both theoretical and empirical perspectives and thus essentially describe the construct we are aiming to assess. The review is well-referenced and therefore offers many leads for further reading, as does the final "resources" chapter. The concept boxes presented throughout the chapters reinforce important points and act as summaries to which one can easily return.

Khalifa, H. and Weir, C. J. (2009). *Examining Reading*. Cambridge: Cambridge University Press.
This book provides another worked example of how to validate reading assessments using an established framework. More specifically, Khalifa and Weir describe the use of the socio-cognitive

approach to evaluate the validity of the Cambridge ESOL Main Suite's reading tests (now called Cambridge English Qualifications).

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing* 24: 7–36. https://doi.org/10.1177/0265532207071510.

Alderson, J. C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson and A. H. Urquhart (eds.), *Reading in a Foreign Language*. London: Longman.

———. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Anderson, N. J., Bachman, L., Perkins, K. and Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing* 8: 41–66. https://doi.org/10.1177/026553229100800104.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing* 28: 51–75. https://doi.org/10.1177/0265532210376379.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing* 30: 441–65. https://doi.org/10.1177/0265532212473244.

Bax, S. and Chan, S. (2016). *Researching the Cognitive Validity of GEPT High-Intermediate and Advanced Reading: An Eye Tracking and Stimulated Recall Study*. Taiwan: The Language Training and Testing Center.

Bowles, M. A. (2010a). *The Think-Aloud Controversy in Second Language Research*. New York and Oxon: Routledge.

———. (2010b). Concurrent verbal reports in second language acquisition research. *Annual Review of Applied Linguistics* 30: 111–27. https://doi.org/10.1017/S0267190510000036.

Brunfaut, T., Kormos, J., Michel, M. and Ratajczak, M. (2021). Testing young foreign language learners' reading comprehension: Exploring the effects of working memory, grade level and reading task. *Language Testing* 38: 356–77. https://doi.org/10.1177/0265532221991480.

Brunfaut, T. and McCray, G. (2015a). *Looking into Test-Takers' Cognitive Processing Whilst Completing Reading Tasks: A Mixed-Methods Eye-Tracking and Stimulated Recall Study*. ARAGs Research Reports Online, AR/2015/01. London: British Council. Retrieved from www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final.pdf.

———. (2015b). *Modelling L2 Reading Proficiency Using Eye-Tracking Measures*. Paper presented at the CRELLA/LU Eye tracking Research Day, Luton, UK.

Burton, S. J., Sudweeks, R. R., Merrill, P. F. and Wood, B. (1991). *How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty*. Brigham Young University Testing Services and The Department of Instructional Science. Retrieved from https://testing.byu.edu/handbooks/betteritems.pdf.

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of rauding theory. *Scientific Studies of Reading* 1: 3–43. https://doi.org/10.1207/s1532799xssr0101_2.

Chan, S. H. C. (2013). *Establishing the Validity of Reading-into-Writing Tasks for the UK Academic Context*. Unpublished doctoral dissertation. University of Bedfordshire, Luton, UK.

———. (2018). *Defining Integrated Reading-into-Writing Constructs: Evidence at the B2 C1 Interface*. English Profile Series Studies 08. Cambridge: Cambridge University Press.

Chan, S. H. C., Wu, R. Y. F. and Weir, C. J. (2014). *Examining the Context and Cognitive Validity of the GEPT Advanced Writing Task 1: A Comparison with Real-Life Academic Writing Tasks*. LTTC-GEPT Research Reports, RG-03. Taiwan: Language Training and Testing Center.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research* 49: 222–51. https://doi.org/10.3102/00346543049002222.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P. B., Mulcahy-Ernt, P. and Schedl, M. A. (2000). *TOEFL 2000 Reading Framework: A Working Paper*. TOEFL Monograph Series, TOEFL-MS-17. Princeton, NJ: Educational Testing Service.

Francis, D. J., Kulesz, P. A. and Benoit, J. S. (2018). Extending the simple view of reading to account for variation within readers and across texts: The complete view of reading (CVR*i*). *Remedial and Special Education* 39: 274–88. https://doi.org/10.1177/0741932518772904.

Freedle, R. and Kostin, I. (1993). The prediction of TOEFL reading item difficulty: Implications for construct validity. *Language Testing* 10: 133–70. https://doi.org/10.1177/026553229301000203.

Gass, S. M. and Mackey, A. (2017). *Stimulated Recall Methodology in Applied Linguistics and L2 Research*. New York and Oxon: Routledge.

Gebril, A. and Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing* 21: 56–73. https://doi.org/10.1016/j.asw.2014.03.002.

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Journal of the Reading Specialists* 24: 44–69. https://doi.org/10.1080/19388076709556976.

Gough, P. B. (1972). One second of reading. *Visible Language* 6: 291–320.

Grabe, W. (2009). *Reading in a Second Language: Moving from Theory to Practice*. New York, NY: Cambridge University Press.

Grabe, W. and Jiang, X. (2014). Assessing reading. In A. J. Kunnan (ed.), *The Companion to Language Assessment*. Hoboken, NJ: Wiley-Blackwell, 185–200.

Grabe, W. and Stoller, F. L. (2020). *Teaching and Researching Reading*, 3rd edn. Oxon and New York, NY: Routledge.

Green, A. (1998). *Verbal Protocol Analysis in Language Testing Research: A Handbook*. Cambridge: Cambridge University Press.

Green, R. (2013). *Statistical Analyses for Language Testers*. London: Palgrave Macmillan.

———. (2017). *Designing Listening Tests: A Practical Approach*. London: Palgrave Macmillan.

Guthrie, J. T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly* 23: 178–99. https://doi.org/10.2307/747801.

In'nami, Y. and Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing* 26: 219–44. https://doi.org/ 10.1177/0265532208101006.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P. and Taylor, C. (1999). *TOEFL 2000 Framework: A Working Paper*. TOEFL Monograph Series, TOEFL-MS-16. Princeton, NJ: Educational Testing Service.

Jeon, E. H. and Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning* 64: 160–212. https://doi.org/10.1111/lang.12034.

Kamata, A. and Vaughn, B. K. (2004). An introduction to differential item functioning analysis. *Learning Disabilities: A Contemporary Journal* 2: 49–69.

Khalifa, H. and Weir, C. J. (2009). *Examining Reading*. Cambridge: Cambridge University Press.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review* 95: 163–82.

Kremmel, B., Brunfaut, T. and Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics* 38: 848–70. https://doi.org/10.1093/applin/amv070.

Koo, J., Becker, B. J. and Kim, Y-S. (2013). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing* 31: 89–109. https://doi.org/10.1177/0265532213496097.

Löwenadler, J. (2019). Patterns of variation in the interplay of language ability and general reading comprehension ability in L2 reading. *Language Testing* 36: 369–90. https://doi.org/10.1177/0265532219826379.

McCray, G. (2014). *Statistical Modelling of Cognitive Processing in Reading Comprehension in the Context of Language Testing*. Unpublished doctoral dissertation. Lancaster University, Lancaster, UK.

McCray, G. and Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing* 35: 51–73. https://doi.org/10.1177/0265532216677105.

Mundy, J. (1968). *Read and Think*. Harlow: Longman.

Oller, J. W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning* 23: 105–18. https://doi.org/10.1111/j.1467-1770.1973.tb00100.x.

O'Sullivan, B. and Dunlea, J. (2015). *Aptis General Technical Manual, Version 1.0*. Technical Report, TR/2015/005. London: British Council.

O'Sullivan, B. and Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (ed.), *Language Testing: Theories and Practices*. Basingstoke: Palgrave Macmillan, 13–32.

Pae, T. I. (2004). DIF for examinees with different academic backgrounds. *Language Testing* 21: 53–73. https://doi.org/10.1191/0265532204lt274oa.

———. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing* 29: 533–54. https://doi.org/10.1177/0265532211434027.

———. (2019). A simultaneous analysis of relations between L1 and L2 skills in reading and writing. *Reading Research Quarterly* 54: 109–24. https://doi.org/10.1002/rrq.216.

Perfetti, C. A. (1997). Sentences, individual differences, and multiple texts: Three issues in text comprehension. *Discourse Processes* 23: 337–55. https://doi.org/10.1080/01638539709544996.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing* 20: 26–56. https://doi.org/10.1191/0265532203lt243oa.

Plakans, L. and Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing* 17: 18–34. https://doi.org/10.1016/j.asw.2011.09.002.

———. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing* 22: 217–30. https://doi.org/10.1016/j.jslw.2013.02.003.

Raatz, U. and Klein-Braley, C. (1981). The C-test – a modification of the cloze procedure. In T. Culhane, C. Klein-Braley and D. K. Stevenson (eds.), *Practice and Problems in Language Testing*. University of Essex Department of Language and Linguistics Occasional Papers, 26. Colchester: University of Essex.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124: 372–422. https://doi.org/10.1037/0033-2909.124.3.372.

Runnels, J. (2013). Measuring differential item and test functioning across academic disciplines. *Language Testing in Asia* 3. https://doi.org/10.1186/2229-0443-3-9.

Rupp, A. A., Ferne, T. and Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing* 23: 441–74. https://doi.org/10.1191/0265532206lt337oa.

Sarig, G. (1989). Testing meaning construction: Can we do it fairly? *Language Testing* 6: 77–94. https://doi.org/10.1177/026553228900600107.

Smith, F. (1971). *Understanding Reading*. New York, NY: Holt, Rinehart & Winston.

Stanovich, K. E. (1980). Toward an interactive compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly* 16: 32–71. https://doi.org/10.2307/747348.

Tunmer, W. E. and Chapman, J. W. (2012). The simple view of reading redux vocabulary knowledge and the independent components hypothesis. *Journal of Learning Disabilities* 45: 453–66. https://doi.org/10.1177/0022219411432685.

Wang, J., Engelhard, G., Raczynski, K., Song, T. and Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing* 33: 36–47. https://doi.org/10.1016/j.asw.2017.03.003.

Weigle, S. C. and Montee, M. (2012). Raters' perception of textual borrowing in integrated writing tasks. In E. E. Van Steendam, M. Tillema, G. Rijlaarsdam and H. Van Den Bergh (eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practises*. London: Brill, 117–45.

Weir, C. J. (2005). *Language Testing and Validation*. Basingstoke: Palgrave Macmillan.

Winke, P. M. (2014). Eye-tracking technology for reading. In A. J. Kunnan (ed.), *The Companion to Language Assessment*. Hoboken, NJ: Wiley-Blackwell, 1029–46.


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Section 5

Test design and administration


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

16
Test specifications

Yan Jin

Introduction

The origin of "specification" in educational and psychological assessment was traced to the definition provided by Ruch (1929) in Davidson and Lynch (2002): "an efficient generative blueprint by which many similar instances of the same assessment task can be generated" (p. 4). This view of test specification emphasizes its central role in test development and has remained largely unchanged.

In the educational measurement literature, a useful distinction is made between norm-referenced measurement (NRM) and criterion-referenced measurement (CRM) (Glaser, 1963). From a measurement point of view, the two differ in their fundamental assumptions about meaningful score interpretation. Having believed firmly for many years in the superiority of a CRM approach to test development, Davidson and Lynch (2002) came to realize that "the NRM/CRM distinction is not as necessary as we once thought" (p. 7). Instead, they stressed that "regardless of the use of the test score, certain fundamental practices always seem to apply. . . . [A] test of any purpose should be developed in an iterative and consensus-based form, and such development can be achieved through specifications" (p. 7–8).

Drawing on the conceptual framework of evidence centred design (ECD) for the design and delivery of educational assessments (Mislevy, 2003a; Mislevy *et al*., 2003), Fulcher and Davidson (2009) developed a multi-layered architectural documentation which unravels the complexities of test creation, making it possible to "articulate design decisions with a large degree of clarity" (p. 126). At the top of this layered documentation is the model, which presents "a theoretical overview of what we understand by what it means to know and use a language"; the middle layer is the framework, which lays out "the constructs to be tested" and states "test purpose for a particular context of score use"; at the bottom of the structure are test specifications, which elaborate on "the detail that is specific to a particular test for use in the context specified in the framework" (pp. 127–28). Test specifications therefore afford a very useful means of translating design decisions into action – it is the specifications that operationalize test constructs and achieve test purposes.

There is abundant literature on the role that test specifications play in supporting the development of language tests. In the first edition of this book, Davidson (2012) provided

DOI: 10.4324/9781003220756-22

a detailed overview. This chapter aims to take a closer look at how test specifications are developed and used and reflect on the role of specifications for improving test validity and facilitating test validation. Given my interest in the social orientation of test development and use, I considered it necessary to differentiate language tests by their intended purposes and consequences: those for making high-stakes decisions leading to serious consequences and those for classroom uses when decisions of relatively lower stakes are made. Though equally useful in both contexts, specifications may serve different functions and take different forms for language tests and assessments with varying degrees of stakes.

Specifications for high-stakes language testing

All tests have some degree of stakes, the level of which is determined by the importance of the decisions to be made. Bachman and Damböck (2018) differentiated three types of decisions: High-stakes decisions are major, life-affecting decisions (e.g., university admission); medium-stakes decisions are neither very high nor very low stakes (e.g., placing students into different ability levels in a language program); low-stakes decisions are minor decisions (e.g., providing feedback to students on their performance on a classroom quiz) (p. 30–31). This section provides a review of the generative and productive role of specifications in high-stakes language testing. Guided by Fulcher and Davidson's (2009) three-layered structure, the review focuses on how test purposes and constructs, documented in the layer of the framework, can be operationalized in test specifications.

Statement of test purposes

According to Spolsky (1995), in language testing, especially in high-stakes language testing, "the first step must be an honest decision on the purpose of the test" (p. 358). Tests are considered as having high stakes when they are used to make important decisions. The level of stakes is a relative concept, depending on the importance of the decision to be made in a particular social and educational context. Typically, language tests are used for making such high-stakes decisions as college entrance, employment, career promotion, and residency permit (e.g., Cheng and Curtis, 2010; Cizek, 2005; Yu and Jin, 2016). Language tests are even used for making decisions on citizenship, immigration, and asylum (Shohamy and McNamara, 2009). These uses, however, may or may not be intended by the test developer. Test purposes that are stated in specifications refer to those uses of a language test that are endorsed and encouraged by the test developer. That is, statements of test purposes set out the claims of the test developer on the intended uses of a language test, though the actual consequences of the uses need further evidential support.

Specifications of a high-stakes language test, therefore, should begin with a clear understanding of and an explicit statement about test purposes. In general terms, it is easier to define purposes of language tests for specific purposes than those for general purposes. The following examples of two international language tests are cited for a comparison of the specifications of test purposes in the two situations.

Example 1: OET is an international English language test that assesses the language communication skills of healthcare professionals who seek to register and practise in an English-speaking environment (www.occupationalenglishtest.org/test-information/).

Example 2: The IELTS measures the language proficiency of people who want to study or work where English is used as a language of communication. IELTS is available

> in Academic – for people applying for higher education or professional registration – and General Training – for those migrating to Australia, Canada, or the UK or applying for secondary education, training programs, or work experience in an English-speaking environment (www.ielts.org/en-us/what-is-ielts/ielts-introduction).

It can be seen that OET defines its purpose and target audience specifically, whereas the purposes of IELTS are broadly defined to include a diverse range of uses: for study, for work, and even for immigration in the case of General Training. Spaan (2006) provided practical advice on the development of test and item specifications. In her view, the prime consideration to be given to the development of any test is that of purpose. Her explanation of purpose specification included specifying the target audience, which must be defined in terms of "age, academic or professional level, language proficiency level, and possibly geographical location or cultural background" (p. 71) (see also O'Sullivan and Green, 2011 for test-taker characteristics). Spaan (2006) further stressed that "[N]o single method about how to develop a test exists" because "[T]he extent of the process and level of detail will vary depending on the purpose, the stakes, the resources, and the stakeholders" (p. 78–79).

Fulcher and Davidson (2009) cautioned that "[T]ests that do not state purpose are as useless in decision making as are buildings that are designed without users in mind" (p. 125). This is particularly true in the context of high-stakes language testing, where the test developer should be morally responsible for defining, as explicitly and specifically as possible, the intended purposes of their testing services. As pointed out in Fulcher (2010a), "[I]t is test purpose that drives all the other activities associated with the development of a test" (p. 96).

Definition of test constructs

While test purposes address the "why" – why a language test is needed in a particular context – construct definition looks into the "what" issue of test development: what knowledge, skills, and/or abilities should be assessed so that the test score will be meaningful and useful? Therefore, when purposes have been specified in a high-stakes testing context, the next step is to define the construct to be tested: in Fulcher and Davidson's (2009) terms, provide a "framework" within which items and tasks can be designed, delivered, and scored.

The field of language testing in the 1960s and 1970s was influenced by structuralism and adopted mainly a skill/element view of language proficiency, resulting in a heavy reliance on discrete-point items (Carroll, 1961; Oller, 1979). The construct of language tests at this stage was specified in terms of language elements (e.g., phonology/orthography, morphology, syntax, lexicon) and language skills (e.g., auditory comprehension, oral production, reading, writing). However, the interaction between elements and skills was not well understood and clearly specified. When high correlations between these components of the construct were observed and a principal component was generated through factor analysis, Oller (1979) proposed a unitary competence hypothesis (UCH). Construct specification in a unitary view was difficult because language proficiency was deemed an indivisible construct. General competence or proficiency was thought to be best demonstrated through performances on integrative tasks such as dictation, cloze, oral interview, essay writing, and translation.

The UCH was rejected in the 1980s by evidence of divisibility of language competence and a deeper understanding of the role of context in language use. Based on theories of communicative competence (e.g., Hymes, 1972; Canale and Swain, 1980), Bachman (1990) and Bachman and Palmer (1996) developed and refined a model of communicative

language ability (CLA) to serve as the construct definition of language tests. In the model, the construct of communicative competence is viewed as an interaction between test-taker characteristics (e.g., personal characteristics, topical knowledge, and affective schemata) and language ability. Language ability comprises two components: knowledge about the language and strategic competence to mobilize the knowledge for various communicative purposes. The CLA-based construct definition takes into consideration users' language ability as well as contextual facets (e.g., test-taker characteristics) that influence language use.

In recent decades, language standards or frameworks have also been referred to when the construct of a language test is defined. Since its release in 2001, the Common European Framework of Reference for Languages (Council of Europe, 2001) has probably been the most well-known and extensively used standard for understanding and defining language proficiency. Fulcher (2015), however, cautioned that in standards-based systems, "the descriptors are secondary to policy" (p. 132). Commenting on the use of Interagency Language Roundtable (ILR) as standard for language assessment in the military in the US, Fulcher (2015) pointed out that "the verbal descriptors of standards-based documents do not matter as much as the socialised understanding of their meaning" (p. 132). As a response to the problem involved in construct definition, Pragmatic realism was proposed, which prioritized observation of communication and analysis of language in use (see *Pragmatic realism and effect-driven testing* later in this chapter).

To demonstrate the way in which constructs are specified in a language test, we use the example of the SAT Essay Test. The test is targeted at high-school students hoping to start a college education or career. An integrated writing task is designed to give students an opportunity to "demonstrate college and career readiness proficiency in producing a cogent and clear written analysis using evidence drawn from an appropriately challenging source text written for a broad audience" (The College Board, 2015: 73). In such an integrated task, a key contextual facet is the type of input text used to elicit writing performances. The input texts used in the SAT Essay Test are specified in terms of their source, style, content, topic, and text complexity (p. 70):

> All passages are taken from high-quality, previously published sources. While the specific style and content of the passages inevitably vary to some extent given the College Board's commitment to using authentic texts with this task, the passages take the general form of what might be called arguments written for a broad audience. . . . Text complexity of the passages is carefully monitored to ensure that the reading challenge is appropriate and comparable across administrations but not an insuperable barrier to students responding to the source text under timed conditions. Prior knowledge of the passages' topics is not expected or required.

In addition to verbal description, sample texts are used to further illustrate the distinctive features of the source texts. The construct of the integrated task is operationalized and specified by the so-called "content dimensions" for reading, analysis, and writing. The three components of the construct are presented in the form of a detailed checklist of the skills and strategies required of test takers for task completion (see Table 16.1, The College Board, 2015: 73).

When a language test uses performance-based tasks such as essay writing or oral presentation, the scoring criteria of such tasks should be specified and presented because they represent the de facto construct of the tasks. Again, we take the SAT Essay Test as an example. In addition to a comprehensive description of the content dimensions, the information package

Table 16.1 The SAT essay domain

Content dimension	*Description*
Reading	Comprehension of the source text Understanding of central ideas, important details, and their interrelationship Accuracy in representation of the source text (i.e., no errors of fact or interpretation introduced) Use of textual evidence (quotations, paraphrases, or both) to demonstrate understanding of the source text
Analysis	Analysis of the source text and understanding of the analytical task Evaluation of the author's use of evidence, reasoning, and/or stylistic and persuasive elements, and/or features chosen by the student Support for claims or points made in the response Focus on features of the text most relevant to addressing the task
Writing	Use of a central claim Use of effective organization and progression of ideas Use of varied sentence structures Employment of precise word choice Maintenance of a consistent, appropriate style and tone Command of the conventions of Standard Written English

prepared by the College Board also provides the rubric that raters will use to evaluate test takers' performance. According to the rubric, responses are "evaluated for demonstrated comprehension of the source text, the quality of analysis of that source text, and the quality of the writing in the response" The College Board, 2015: 71). Table 16.2 is a summary of the rubric of SAT Essay Test (p. 80; a full rubric can be found at https://collegereadiness.collegeboard.org/sat-essay-scoring-before-march-2016).

Specification of design features

With purposes explicitly stated and constructs clearly defined, the craft of test design in a high-stakes context is to specify the "how": how is the framework of a language test operationalized in its specifications? Derived from the conceptual assessment framework within ECD (Mislevy, 2003a, 2003b), Fulcher and Davidson's (2009) four sub-layers of test specifications are particularly relevant and useful: item specifications, the presentation model, the delivery model, and the test assembly model. Item specifications describe "the items or tasks, and any material such as input texts, upon which they depend"; a presentation model specifies "how the items and tasks are presented to the test takers"; a delivery model explains "how the actual test is delivered, including administration, security and timing"; and an assembly model depicts "how the tasks and items should be combined to produce a test form" (pp. 128–29).

As an example of how the design features of a high-stakes test can be specified, a summary table is presented to show the main features of the College English Test-Spoken English Test (CET-SET) Band 4 (National College English Testing Committee, 2016) (see Table 16.3). The speaking test takes a computer-based format in which instructions are given by an examiner through a pre-recorded video. Two test takers are paired as a working team immediately before taking the test. During the test, the two test takers see each other's photo on the screen, listen to each other through earphones, and talk to each other

Table 16.2 A summary of the rubric of the SAT Essay Test

Reading	*Analysis*	*Writing*
The Reading dimension encompasses students' comprehension of a source text, including the text's central ideas and important details and how they interrelate. Students demonstrate their skill in comprehension in part by making effective use of evidence (quotations, paraphrases, or both) from the source text.	The Analysis dimension encompasses students' evaluation of an author's use of evidence, reasoning, and/ or stylistic and persuasive elements, and/ or features of the text of the students' own choosing. Students demonstrate their skill in analysis in part by using relevant, sufficient, and strategically chosen support for the claims or points they make and by focusing consistently on those features of the text that are most relevant to addressing the task.	The Writing dimension encompasses the cohesiveness of students' written response to the task as well as students' use of language. Students demonstrate their skill in writing in part by providing a precise central claim; creating an effective organization and progression of ideas; successfully employing a variety of sentence structures; using precise word choice; maintaining an appropriate style and tone; and showing command of the conventions of Standard Written English.

Table 16.3 Specification of some features of the CET-SET Band 4

Task order	*Task format*	*Planning/response time*	*Mode of input*	*Mode of speaking*
Warm up	Self-introduction	0/20 seconds	Video	Monolog
Task 1	Read-aloud a text	45/60 seconds	Verbal	Monolog
Task 2	Answer questions	0/40 seconds	Video	Monolog
Task 3	Individual presentation	45/60 seconds	Verbal, graph	Monolog
Task 4	Pair discussion	1/3 minutes	Verbal, graph	Dialog

via microphones. The test starts with a warm-up exercise, through which test takers in a pair get to know each other. The first task is reading aloud: a text of about 120 words is shown on the screen, and test takers read the text after 45 seconds of preparation. Test takers then answer two questions related to the text. This is followed by an individual presentation, in which prompts are shown on the screen, and the two test takers take turns presenting on a given topic. In the final task, after one minute of preparation, the test takers are engaged in a pair discussion for three minutes.

Elaborate and detailed specifications of design features enable test developers to produce test tasks that better resemble tasks in a target language use (TLU) domain. A critical quality of the design of a language test is task authenticity (see Lewkowicz, 2000, for an overview of the authenticity debate). Bachman and Palmer (1996) defined authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task." As one of the six qualities of a language test, authenticity is considered crucial to test usefulness because "it relates the test task to the domain of generalization to which we want our score interpretations to generalize" (p. 23–24). The tasks of the computer-based CET-SET, for example, were designed to simulate monologic and dialogic oral communication in

the real world and maximally operationalize the construct of oral production and interaction (see Jin and Zhang, 2016).

Specifications for classroom-based language assessment

In this section, the role of specifications in medium- and low-stakes contexts is first discussed focusing on the purposes of classroom-based language assessment (CBLA). Key considerations involved in CBLA are then reviewed with reference to Turner and Purpura's (2016) framework of learning-oriented assessment. Finally, sample specifications are presented to demonstrate the design features of specifications for classroom-based language assessment. The switch of terminology, replacing "test" with "assessment," indicates a need to distinguish CBLA from tests for high-stakes decisions (see also Fox *et al*., this volume). "Test specifications," however, is consistently used in both classroom-based and high-stakes contexts because most of the literature takes this label.

Statement of assessment purposes

Large-scale international or national language tests are usually developed and used for making high-stakes decisions. "Non-standardized local assessment carried out by teachers in the classroom" (Leung, 2005: 871), or classroom-based language assessment, is considered as having lower stakes because it is mainly designed and used by teachers as a tool to "gauge and enhance their students' learning" (Tan and Turner, 2015: 30). CBLA serves educational purposes and ties itself closely to curricula and teaching.

In high-stakes testing contexts, teachers are typically excluded from the process of test development and validation, though research has demonstrated that "teacher involvement in high-stakes language testing is a beneficial enterprise for different stakeholders" (Xerri and Briffa, 2018: 341). Specifications of high-stakes tests are developed by testing experts through an "iterative" and "consensus-based" approach (Davidson, 2012: 202). In contrast, teachers are decision makers of CBLA, and they know their students best. When classroom-based assessment activities are developed and delivered, teachers are given sufficient autonomy and flexibility. Specifications of CBLA, therefore, often take a different form: they are broadly defined in curricula and carefully specified and embedded in course descriptions and teaching plans.

Traditionally, two types of decisions are made based on CBLA: formative and summative decisions. Formative assessments lead to activities that are intended to improve instruction and learning, whereas summative tests aim at advancement and certification (see Bachman and Damböck, 2018: 27–29, for a summary of classroom-based assessment decisions).

A more detailed classification of the assessment purposes of CBLA is available at the website provided by the TALE project (http://taleproject.eu). The project team has developed an online, self-study (Moodle-based) training course for both in-service and pre-service language teachers. In its recently published handbook, Vogt and Froehlich (2015) introduced "alternatives in assessment," which were defined as "holistic and learner-oriented assessment" (p. 148). Typical tasks include teacher observation, self-/peer assessment, and portfolio assessment. An important feature of alternative assessment is that it is learner centered and tends toward assessment *for* learning. Purposes of alternative assessment are summarized as follows (p. 151):

1 Enable teachers to develop lesson plans and approaches;
2 Determine learners' ability to use the foreign language in authentic situations;

3 Allow learners to present what they are able to do;
4 Allow learners to have a say in their learning;
5 Serve as evidence of learners' skills;
6 Give teachers a more precise depiction of their learners' language.

Alderson *et al*. (2017) also drew our attention to several so-called new paradigms for "bringing assessment practice more into line with theories of learning" (p. 380): diagnostic assessment, learning-oriented assessment, and dynamic assessment. Though language assessments in these new paradigms are "still on the fringe with regard to wide-scale use and practice" (p. 381), progress has been made in recent years. Take diagnostic language assessment (DLA) as an example. A number of well-known DLA systems have been developed, including DIALANG, DELTA (Diagnostic English Language Tracking Assessment), DELNA (Diagnostic English Language Needs Assessment), and so on (see Alderson *et al*., 2015 for a comprehensive overview of these assessments; see also Jang and Sinclair, this volume for further discussion of DLA).

In a special issue of *Language Testing*, Lee (2015) defines DLA as "identifying test-takers' (or learners') weaknesses, as well as their strengths, in a targeted domain of linguistic and communicative competence and providing specific diagnostic feedback and (guidance for) remedial learning" (p. 303). In DLA, the link between assessment and learning has been clearly established. The following is an example of how assessment purposes are specified in the Guidelines for Users of the web-based DELTA, developed by a team of language assessment experts from three universities in Hong Kong (DELTA, 2018: 3):

> The DELTA provides all stakeholders with valuable information on which decisions can be made about language enhancement provision, both by students individually and by institutions. . . . As students progress through their programme of study, they will take the DELTA at intervals and use the results to make improvements in their overall proficiency so that they reach a point at graduation having made real improvements. In other words, they will be able to "track" their progress to ensure that they remain "on track" to attain the outcome in terms of language proficiency required by the discipline that they intend to enter post-graduation.

In addition to a general description, advice is given on how to make the best use of the DELTA: when to take the DELTA, setting a target DELTA measure, and creating a study program (DELTA, 2018: 11–12). More detailed guidance is provided for each of the three steps. A study program, for example, may include

1 Using the online materials and resources that are linked to the DELTA report;
2 Discussing DELTA results with an English teacher;
3 Visiting the institution's self-access centre for language learning and speaking to an advisor;
4 Enrolling in an enhancement programme that focuses on areas of weakness.

Key considerations in CBLA

Second language (L2) learning, according to Turner and Purpura (2016), is "a highly individual cognitive process" and "a highly intricate socio-cognitive and sociocultural process" (p. 261). Consequently, many factors affect the effectiveness of teaching and learning.

Following an in-depth review of theoretical and empirical research in L2 classroom assessment, Turner and Purpura (2016) advocated a learning-oriented approach to L2 assessment. Learning-oriented assessment (LOA) "highlights learning goals, performance evaluation and feedback, and the role they play in developing individual learning progressions" (p. 260). By doing so, the field of language testing could possibly "move beyond dichotomous depictions of CBA as formative/summative or AOL/AFL, toward an approach that characterizes assessment, learning, and instruction, while different, as intrinsically intertwined" (p. 260). Alderson *et al*. (2017) also noted that within the new paradigms of assessment, "learning and assessment have become not only intertwined, but mutually dependent" (p. 380).

The LOA framework of Turner and Purpura (2016) depicts seven interrelated dimensions for conceptualizing and implementing L2 classroom assessment: the social and educational context, ways to elicit language for assessment, theories of language proficiency, theories of learning, instructional factors, interactional patterns, and learners' affective factors (see pp. 261–68 for detailed explanations). These factors are critical to the design and use of CBLA for improving teaching and learning. More importantly, these considerations are closely interconnected. Language policies (a macro-level contextual factor), for example, may have an impact on the environment of L2 teaching and learning; personalities of teachers and learners (micro-level contextual factors) may shape the way L2 is taught and learnt in classrooms; these macro- and micro-level contextual factors may combine to affect the elicitation methods (the way learners' language performances are elicited) used in L2 classroom assessment. Learner engagement (also a micro-level contextual factor) and their attitudes toward L2 learning (an affective factor) may interact with each other and affect their responses to feedback and their willingness to take intervention measures (interactional patterns). Teachers' understanding of curricula and their knowledge of L2 learning theories may affect their decisions on learning objectives, assessment criteria, elicitation methods, and interpretation of assessment results. LOA, therefore, operates in a complex system in which many factors combine to affect the effectiveness of assessment.

Specifications serve as a means of linking these factors and enhancing the effectiveness of learning-oriented assessment. Bachman and Damböck (2018) commented that the blueprint (i.e., specification) "provides a link between the learning objectives of the language classroom and the content to be assessed, and between the characteristics of the instructional tasks and those of the assessment tasks"; by so doing, the blueprint "helps you [the teacher] make sure that the assessment tasks you include in the assessment adequately sample both the content to be learned and the tasks in the TLU domain, to assure that your interpretations will be meaningful and generalizable" (p. 188).

Specification of design features

As can be seen in the section on assessment purposes, CBLA takes a variety of forms to meet the needs of teaching and learning. A useful distinction was made in Bachman and Damböck (2018) between two modes of CBLA: the implicit mode and the explicit mode (pp. 33–34). In the implicit mode, also referred to as "dynamic assessment" or "informal assessment," assessment activities are instantaneous and become a part of classroom teaching, and students may not be aware of the assessment; the explicit mode of CBLA, or "formal assessment," is a separate activity from teaching and is used mostly for summative decisions.

Turner and Purpura (2016) classified the methods of eliciting language for classroom assessment in a more refined way (pp. 262–64). Basically, there are also two types of

assessment activities: those inside the classroom and those outside the classroom. Inside classrooms, teachers can employ a range of planned language elicitations, which are embedded in instruction, as well as numerous types of unplanned or spontaneous assessments; outside classrooms, assessment can take the form of placement exams, standardized achievement tests, proficiency exams, or aptitude tests.

Specifications may be embedded in teaching plans for instantaneous classroom-based assessment activities, which are in fact mingled with teaching activities. But specifications can be of special use for planned assessment tasks or assessments in an explicit mode for making summative decisions. Bachman and Damböck (2018) suggested that a blueprint of CBLA should include the following components (pp. 180–81):

1 Assessment specifications:
    a) Numbers and descriptive labels of the parts
    b) Numbers and descriptive labels of the tasks per part
    c) Relative importance of the parts and tasks
    d) Time allowed for each part and overall
2 Procedures and instructions for the whole assessment
    a) Procedures
    b) General instructions
3 Assessment task template (for each task type)
    a) Activities and procedures to be followed
    b) Assessment task characteristics
    c) Recording method
    d) Task-specific instructions

In the handbook produced by the TALE project, Green (2015) followed Bachman and Damböck's (2018) recommendation and provided examples of specifications of each component of a CBLA blueprint. Task characteristics, for example, are specified in a three-element structure: instructions, input, and expected responses (Green, 2015: 7). Moreover, effective feedback could be generated following Wiggins's (1998) template in order to help learners understand their own performance (Green, 2015: 9, emphasis in original):

1 what they are doing **well**
2 what they are **struggling with**
3 what they need to **do differently** in order to be **more successful**

The process of developing CBLA may not be as rigorous as that involved in a high-stakes testing context, and teachers can be creative in their construction of assessment tasks. Nonetheless, specifications (or templates) may prove useful for teachers to create multiple tasks and deliver them with a satisfactory degree of consistency. It is also worth pointing out that classroom-based assessment should be conducted on a continuous and cyclical basis so as to fully support teaching and learning and to bring about expected changes.

Specifications for test validity and validation

Following a discussion of the role of specifications in both high-stakes language testing and classroom-based language assessment contexts, we move on to examine the inextricable connection between test/assessment purposes and the design of language tests/assessments and highlight the usefulness of specifications for enhancing test validity and facilitating test validation. At the end of the section, Fulcher's (2015) notion of Pragmatic realism is discussed for insights on developing and using specifications to promote effect-driven testing.

Specifications for test validity

It is widely accepted that the primary purpose of test specifications is to establish equivalence between test forms and ensure the quality of language tests and assessments. In recent decades, the field of language testing has begun to take a social turn, and the scope of test specifications has been extended beyond a narrow focus on reliability or consistency to a broader view of test validity.

Alderson *et al*. (1995) made a deliberate attempt to distinguish between "test syllabus" and "test specification." In their view, a test specification "is a detailed document, and is often for internal purposes only," whereas the syllabus is "a public document, often much simplified, which indicates to test users what the test will contain" (p. 9). In other words, the specification contains all the technical details for guiding the design, development, scoring, and score reporting of a language test. The syllabus, in contrast, is derived from a test specification and is produced for public use. The subtle differentiation, on the one hand, reflected language testers' awareness of the level of specificity for different audiences of test specifications; on the other hand, the differentiation seemed to suggest that the public did not have adequate assessment literacy to understand technical details or simply did not need to know the technical quality of a language test.

Over the past two decades, with an increasing demand for accountability and transparency in language testing as well as the growing assessment literacy of stakeholders, the distinction between specification and syllabus has been slowly fading. Good practice of language testing requires that all stakeholders be informed of test design features as well as the rationale behind the design. Both test developers and test users are expected to be provided with not only a documentation of test design but also evidence of test validity. Evidence needs to be produced, articulated, and presented to support score interpretation and score-based decisions. Specifications provide the best and the most direct means to communicate to stakeholders all the information about a language test or assessment, including, for example, the intended purposes; targeted audience; knowledge, skills, and/or abilities to be assessed; task format; input material; expected performance; scoring criteria; score interpretation; and feedback to test takers/learners.

Davies *et al*. (1999) pointed out that "specifications are important in the establishment of the test's construct validity" (p. 207). In their definition of test specifications, the primary audiences of specifications are internal users such as test developers, item writers, examiners, and raters, but specifications are also referred to by test evaluators and test users. For test evaluation and test use, the crucial questions to be considered are: Does the test assess key competences required in a target domain, and are test scores accurate, reliable, and meaningful for decision-making? These are core questions about a test's validity.

Specifications for a technical approach to validation

The usefulness of specifications in establishing test validity is best captured in the process of test validation. Among various approaches to validation, Weir's (2005) socio-cognitive framework (see also O'Sullivan and Weir, 2011) makes the maximum use of test specifications. The approach advocates test validation at both the *a priori* and *a posteriori* stages of test development and highlights a symbiotic relationship among the three core components of test validity – "at the heart of any language testing activity . . . we can conceive of a triangular relationship between three essential components: the test taker's cognitive abilities; the task and context; and the scoring process" (Taylor and Galaczi, 2011: 172). The socio-cognitive approach prioritizes validity evidence supporting 1) the extent to which test tasks successfully elicit from candidates a set of processes which resemble those employed in real-world communicative events (cognitive validity), 2) the way in which contextual factors influence candidate performances (context validity), and 3) the consistency and dependability of test scores (scoring validity).

The application of the socio-cognitive framework is typically specification driven. Take context validity as an example. The specification of contextual facets first necessitates an analysis of language use in the target domain in order to inform the decisions on the choices of tasks that are "representative of the larger universe of tasks of which the test is assumed to be a sample" (Weir, 2005: 19). Following the decision on choices of tasks, task characteristics are specified so that they best simulate the communicative activities required of candidates in the target domain. By manipulating the characteristics of test tasks, the difficulty level of the tasks can also be calibrated to a suitable proficiency level.

Bachman and Palmer's (1996) test usefulness framework has also adopted a similar "technical approach to validity and validation" (Fulcher, 2015: 116). Building on Bachman (1990), Bachman and Palmer (1996) proposed a framework of language task characteristics, which is essentially a checklist for matching the characteristics of test tasks with those of tasks in TLU domains. The framework is not an exhaustive and fixed set of task characteristics, but it has proved to be a valuable tool and a starting place for designing and analyzing test items or tasks. Elements of item/task specifications, according to the checklist, include the purpose of the test task, the definition of the construct of the task, task setting, time allotment, instructions for task responses, characteristics of input, the expected response, the relationship between input and response, and scoring method (see Bachman and Palmer, 1996: 49–50).

Useful as it is, the so-called "technical approach" has been considered as having "three major disadvantages in comparison with the argument-based approach": the tendency of confirmatory research, insufficient attention to alternative hypotheses that form the weakest link in the argument chain, and the time-consuming and resource-intensive nature of application (Fulcher, 2015: 116–20). Nonetheless, technicalism "remains a powerful force in language testing, perhaps because lists are comfortable to work with, provide some kind of programmatic framework for action, and tend to reassure us that nothing is going to get left out" (p. 117).

Pragmatic realism and effect-driven testing

To address various issues identified in the conceptualization of validity and validation, Fulcher (2015) proposed the notion of Pragmatic realism (with a capital P) by drawing on "the traditional criterion-referenced testing movement" (p. 104). The criterion-referenced

approach "prioritises language and communication above the manipulation of standards and numbers" (p. 140). That is, criterion-referenced practices value language testers' observation of communication and support inferences based on testers' experience with language use. True criterion referencing, according to Fulcher (2015), is not "correlation between test and criterion" but "referencing our test to the criterion of language use" (p. 126).

Pragmatic realism embraces "an optimistic and progressive philosophy of language testing" and subscribes to the notion of "effect-driven testing" (Fulcher, 2015: 126). The term "effect-driven testing" was coined in Fulcher and Davidson (2007) to refer to "test creation in which the ultimate test design decisions are driven by the impacts that the test will have on stakeholders" (p. 51). Fulcher (2010b) explained that the first layer of an architectural documentation for test creation (the model) provides guidance to the selection of constructs relevant to test purposes. In the second layer (the framework), decisions are made about the specific knowledge, skills, and abilities to be included in the test based on the intended uses and desirable effects of the test. At the bottom layer, specifications lay out the design features that will help achieve the intended effects.

An explanation of effect-driven testing is given here through a comparison of the specification-driven and the effect-driven approaches to language testing. (See Figure 16.1 for a schematization of the two approaches.)

In Davidson and Lynch (2002), detailed guidance was provided to language teachers on writing and using language tests. Their approach emphasizes the importance of developing solid test specifications based on language testing research: "we advocate tests that are specification-driven: a specification is an efficient generative recipe for a test that fosters dialogue and discovery at a higher, more abstract level than achieved by analysis of a simple item or task" (p. 7). In specification-driven testing, domain specifications and research of language constructs inform the decisions on test purposes and construct definition. Based on the abstraction of these analyses, the framework of a language test is documented, and specifications are developed for test design, item writing, item assembling, and test delivery. A distinctive feature of the approach is that washback and impact are investigated in a post-hoc manner: that is, after the test or assessment has been implemented (see Cheng and Sultana, this volume).

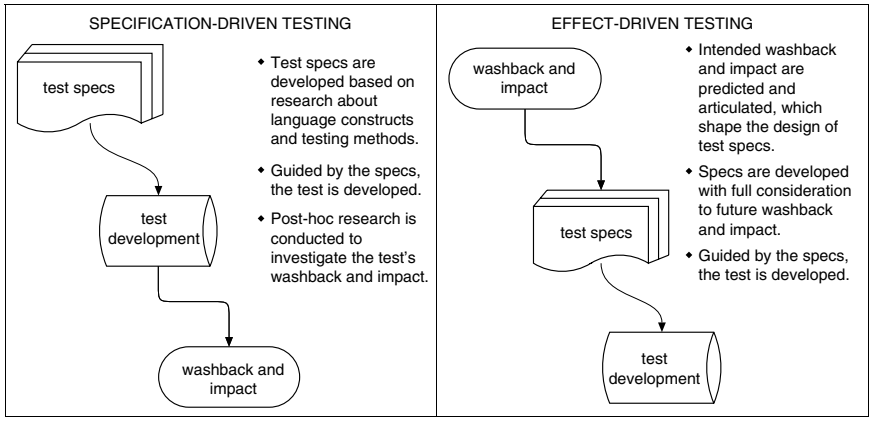


Figure 16.1 A comparison of specification-driven and effect-driving testing

In contrast, when an effect-driven approach is adopted, intended washback and impact are predicted and articulated and shape the design of test specifications. Specifications are developed with full consideration of intended effects in mind, and tests are developed based on the specifications. Washback and impact, in a sense, have been built into the test validation scheme prior to the implementation of the test. Kim and Davidson (2014) encouraged the use of effect-driven specifications, which "detail what types of tasks or items can elicit the behaviors serving as evidence of the desired effect" (p. 794). To facilitate the collection of evidence of test effects, they recommended that "effect considerations" (EC) be added as "a new element to the guiding language in this spec model" (p. 794). The inclusion of such a claim "obligates the test developers to include (in the spec) evidence of how it was developed and of how that development proactively considered the effect the test is intended to achieve" (p. 794). An example was given to show what an EC was like:

> *EC:* The developers of this test spec carefully considered the various stake-holders who will be impacted by this test task. Given that consideration, they then decided upon the observable behaviors of authentic written communication as seen and experienced by those stakeholders. The result is shown here in this spec.

As admitted in Kim and Davidson (2014), effect-driven testing is actually not a new idea. Messick (1996), for example, argued that "rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback" because "in practice, test makers are mainly concerned about adverse consequences that are traceable to sources of test invalidity such as construct under-representation and construct-irrelevant difficulty" (p. 252). Alderson *et al*. (2017) cited the examples of the reform of the language examinations in the Austrian Matura and the planned reform of the secondary school leaving English exam in Luxembourg to illustrate what they called the "washback by design" approach to designing instruments with test uses and consequences at the forefront.

Conclusion and future directions

This chapter began with a review of "the classical utility of test specs to generate items or tasks" (Kim and Davidson, 2014: 789): that is, specification-driven testing in both high-stakes language testing and classroom-based assessment contexts. The usefulness of test specifications was explicated by reference to their role in articulating test/assessment purposes, operationalizing test constructs, linking assessment to curricula, and detailing the design features of tests or assessments. The chapter then considered the way in which test specifications may facilitate validation and improve test washback and impact. The argument of the chapter has been that the usefulness of test specifications should be extended beyond a mechanism for developing equivalent test forms.

Test specifications are integral to language testing and assessment practices. Specifications play a bridging role between construct definition and score interpretation and incorporate the design features that will bring about the effects intended by test developers. Fulcher and Davidson (2009) argued that "the architecture of the test and the way in which its items and tasks are engineered from prototyping to field testing cannot be separated from test purpose, use, and intended effect" (p. 131). Fulcher (2015) reiterated his argument for effect-driven testing: "ethical test design requires the developers to look to the future effects they wish their test to have" (p. 126). From a social perspective, a major function of test specifications is to serve as a valuable instrument for communication with

stakeholders, including test takers and test users. For test takers, test specifications would enable them to develop a good understanding of the test, prepare properly for the test, and perform to their full potential on the test; for test users, specifications would enable them to make informed decisions. In classroom-based contexts, specifications link assessment criteria to curriculum standards, thus facilitating learner-centred and learning-oriented assessment practices.

To conclude his chapter on test specifications, Davidson (2012) pointed out a future of "multiple referencing" – each score report would contain three references: the score stated in normative terms, thick and detailed criterion-based information about the abilities of the test taker, and an active personal investment by the test taker (p. 206). Seven years later, this is still on the agenda for the future. Fortunately, the field has now come to realize the importance of the content aspects of language communication and language use in context: all assessment activities should be based on "a truth universally acknowledged that a language tester, in possession of a purpose for testing, must be in want of a language use domain to which predictions are required" (Fulcher, 2015: 107) (see also the notion of "local construct" in Chalhoub-Deville, 2003). With a closer link between test tasks and language use in the real world, positive effects are more likely to be generated. The exploration and description of language use in context will also lend stronger support to Davidson's (2012) three references, especially criterion-based score interpretation and test takers' self-reflection on their performances.

One notable feature of test specifications in recent years is their increased accessibility and specificity due largely to the availability of the internet. The website of a large-scale and high-stakes language test presents various types of information to stakeholders. Using hyperlinks, sample materials and/or performances are easily demonstrated. Free or commercial apps are also available for test takers to access sample questions, tips and resources for test preparation, online courses, video demonstrations, and practice tests. It is envisaged that the application of technology, including the use of artificial intelligence, will have great potential for building in effect-driven features in task design, test delivery, scoring, and feedback generation, as well as score reporting (see Sawaki, this volume). Taking a more optimistic view of technological development, in the not-so-distant future, scenario-based tasks might be automatically generated, performance on the tasks might be automatically scored, and test takers might be able to experience and take the test on a virtual reality system. Technology may change the landscape of language testing, and by that time, the role of test specifications will have to be redefined.

Further reading

Davidson, F. and Lynch, B. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT and London: Yale University Press.

This book is about language test development using test specifications. Viewed as the basic tool of testcraft, test specifications are introduced and discussed from various perspectives, including their role in specification-driven, inclusive testing.

Fulcher, G. (2015). *Re-Examining Language Testing – A Philosophical and Social Inquiry*. London and New York: Routledge.

This monograph provides a unique analysis of the ideas and social forces that shape the practice of language testing and assessment. The notion of Pragmatic realism proposed in the monograph is enlightening and thought provoking, lending strong support to effect-driven testing.

Fulcher, G. and Davidson, F. (2009). Test architecture, test retrofit. *Language Testing* 26: 123–44.

This article emphasizes the central role of test purposes in language test design and revision and presents a systematic approach to evaluating test revisions or retrofit. The use of the architecture metaphor makes the argument accessible to a broad audience including language testing practitioners.

References

Alderson, C. J., Brunfaut, T. and Harding, L. (2017). Bridging assessment and learning: A view from second and foreign language assessment, *Assessment in Education: Principles, Policy & Practice* 24: 379–87. https://doi.org/10.1080/0969594X.2017.1331201.

Alderson, C. J., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alderson, C. J., Haapakangas, E. L., Huhta, A., Nieminen, L. and Ullakonoja, R. (2015). *The Diagnosis of Reading in a Second or Foreign Language*. New York: Routledge.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. and Damböck, B. (2018). *Language Assessment for Classroom Teachers*. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Canale, M. and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1: 1–47.

Carroll, J. B. (1961). *Fundamental Considerations in Testing for English Language Proficiency of Foreign Students*. Washington, DC: Testing Center for Applied Linguistics.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing* 20: 369–83. https://doi.org/10.1191/0265532203lt264oa.

Cheng, L. and Curtis, A. (eds.). (2010). *English Language Assessment and the Chinese Learner*. London: Routledge.

Cizek, G. J. (2005). High-stakes testing: Contexts, characteristics, critiques, and consequences. In R. P. Phelps (ed.), *Defending Standardized Testing*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 23–54.

The College Board. (2015). *Test Specifications for the Redesigned SAT®*. Retrieved August 15, 2019, from www.collegeboard.org.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Strasbourg: Council of Europe.

Davidson, F. (2012). Test specifications and criterion referenced assessment. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. London and New York: Routledge, 197–207.

Davidson, F. and Lynch, B. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT and London: Yale University Press. Retrieved from www.jstor.org/stable/j.ctt1npx6r.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of Language Testing*. Cambridge: Cambridge University Press.

DELTA. (2018). *Guidelines for Users*. Retrieved October 15, from http://gslpa.polyu.edu.hk/eng/delta_web/.

Fulcher, G. (2010a). *Practical Language Testing*. London: Hodder Education.

———. (2010b). The reification of the common European framework of reference (CEFR) and effect-driven testing. In A. Psyaltou-Joycey and M. Matthaioudakis (eds.), *Advances in Research on Language Acquisition and Teaching*. Thessaloniki, Greece: GALA, 15–26.

———. (2015). *Re-Examining Language Testing – A Philosophical and Social Inquiry*. London and New York: Routledge.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment – An Advanced Resource Book*. London and New York: Routledge.

———. (2009). Test architecture, test retrofit. *Language Testing* 26: 123–44.
Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist* 18: 519–21. https://doi.org/10.1037/h0049294.
Green, A. (2015). The ABCs of assessment. In D. Tsagari, K. Vogt, V. Froehlich, I. Csépes, A. Fekete, A. Green . . . S. Kordia (eds.), *Handbook of Assessment for Language Teachers*, 1–15. Retrieved November 11, 2019, from http://taleproject.eu.
Hymes, D. (1972). Models of the interaction of language and social life. In J. Gumperez and D. Hymes (eds.), *Directions in Sociolinguistics*. New York: Holt, Rinehart and Winston, 35–71.
Jin, Y. and Zhang, L. (2016). The impact of test mode on the use of communication strategies in paired discussion. In G. Yu and Y. Jin (eds.), *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums*. Basingstoke: Palgrave Macmillan, 61–84.
Kim, J. and Davidson, F. (2014). Effect-driven test specifications. In A. J. Kunnan (ed.), *The Companion to Language Assessment, II:7:47*. Chichester, UK: Wiley Blackwell, 788–95.
Lee, Y. W. (2015). Future of diagnostic language assessment. *Language Testing* 32: 295–98. https://doi.org/10.1177/0265532214565385.
Leung, C. (2005). Classroom teacher assessment of second language development: Construct as practice. In E. Hinkel (ed.), *Handbook of Research in Second Language Teaching and Learning*. Mahwah, NJ: Lawrence, 869–88.
Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing* 17: 43–64. https://doi.org/10.1177/026553220001700102.
Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13: 241–56. https://doi.org/10.1177/026553229601300302.
Mislevy, R. J. (2003a). *On the Structure of Educational Assessments*. CSE Technical Report 597. Los Angeles: Centre for the Study of Evaluation, CRESST.
———. (2003b). *Argument Substance and Argument Structure in Educational Assessment*. CSE Technical Report 605. Los Angeles: Centre for the Study of Evaluation, CRESST.
Mislevy, R. J., Almond, R. G. and Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centred Design*. Research Report RR-03-16. Princeton, NJ: Educational Testing Service.
National College English Testing Committee. (2016). *College English Test Syllabus*, revised version. Shanghai: Shanghai Jiao Tong University Press.
Oller, J. (1979). *Language Tests at School*. London: Longman.
O'Sullivan, B. and Green, A. (2011). Test taker characteristics. In L. Taylor (ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Studies in Language Testing 30. Cambridge: UCLES, Cambridge University Press, 36–64.
O'Sullivan, B. and Weir, C. J. (2011). Language testing and validation. In B. O'Sullivan (ed.), *Language Testing: Theory and Practices*. Basingstoke: Palgrave Macmillan, 13–32.
Ruch, G. M. (1929). *The Objective or New-Type Examination: An Introduction to Educational Measurement*. Chicago: Scott, Foresman.
Shohamy, E. and McNamara, T. (2009). Language tests for citizenship, immigration, and asylum. *Language Assessment Quarterly* 6: 1–5. https://doi.org/10.1080/15434300802606440.
Spaan, M. (2006). Test and item specifications development. *Language Assessment Quarterly* 3: 71–79.
Spolsky, B. (1995). *Measured Words*. Oxford: Oxford University Press.
Tan, M. and Turner, C. E. (2015). The impact of communication and collaboration between test developers and teachers on a high-stakes ESL exam: Aligning external assessment and classroom practices. *Language Assessment Quarterly* 12: 29–49. https://doi.org/10.1080/15434303.2014.1003301.
Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking*. Studies in Language Testing, vol. 30. Cambridge: UCLES, Cambridge University Press, 171–233.
Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin: DeGruyter, 255–72.

Vogt, K. and Froehlich, V. (2015). Alternatives in assessment. In D. Tsagari, K. Vogt, V. Froehlich, I. Csépes, A. Fekete, A. Green . . . S. Kordia (eds.), *Handbook of Assessment for Language Teachers*, 148–78. Retrieved November 11, 2019, from http://taleproject.eu.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.

Wiggins, G. P. (1998). *Educative Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco: Jossey-Bass Publishers.

Xerri, D. and Briffa, P. V. (eds.). (2018). *Teacher Involvement in High-Stakes Language Testing*. Switzerland: Springer International Publishing AG.

Yu, G. and Jin, Y. (eds.). (2016). *Assessing Chinese Learners of English: Language Constructs, Consequences and Conundrums*. Basingstoke: Palgrave Macmillan.

17

Evidence-centered design in language testing

Chengbin Yin and Robert J. Mislevy

Introduction

Evidence-centered design (ECD) is a comprehensive framework for describing and articulating the conceptual, inferential, and computational elements of educational assessments, organized around the idea of assessment as evidentiary argument. Inferences concern what learners know or can do in various situations, based on observing what they say, do, or make in a handful of particular situations (Mislevy *et al*., 2003). Assessment is viewed as five coordinated layers of components, processes, and representations. Most directly related to test development are the layers called domain modeling, in which assessment design and use arguments are developed, and the conceptual assessment framework, in which student, evidence, and task models specify the pieces that instantiate the argument (such as scoring methods, measurement models, and schemas for tasks). This chapter lays out the basic ideas of ECD and discusses ways it can be gainfully employed in language testing.

Historical perspectives

Assessment as argument

Validity research concerns ways in which we can build and use assessments that support interpretations and uses of results (Bachman, 2005; Bachman and Palmer, 2010; Kane, 2006, 2013, this volume; Messick, 1994). The elements of an assessment argument are best understood through Toulmin's (1958) argument schema shown as Figure 17.1 (Mislevy, 2018). A *claim* is a proposition we wish to support with data. A *warrant* is a generalization that justifies the inference from the particular data to the particular *claim*. In language testing, we make claims about learners' language proficiency or abilities within and across language domains from observing their task performances. Theory and research concerning processes of language acquisition, stages of language development, and instances of language use can all provide *backing* for the warrant. This backing also informs task design and performance evaluation. When using the test, we reason back up through the warrant from the observations to particular claims. Usually we must qualify our inferences because

DOI: 10.4324/9781003220756-23

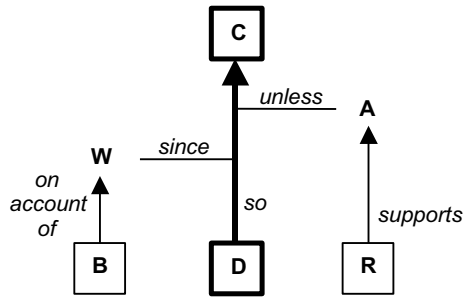


Figure 17.1 Toulmin's general structure for arguments

of *alternative explanations*, which *rebuttal data* might support or undercut: for example, that anxiety rather than language difficulties caused a candidate's poor performance on a speaking task.

Assessment design means specifying the kinds of claims we wish to make, then crafting the operational elements of an assessment that allow us to reason in this manner. Messick (1994) outlines the essential structure:

> A construct-centered approach would begin by asking what complex of knowledge, skills, or other attributes should be assessed, presumably because they are tied to explicit or implicit objectives of instruction or are otherwise valued by society. Next, what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors? Thus, the nature of the construct guides the selection or construction of relevant tasks as well as the rational development of construct-based scoring criteria and rubrics.
>
> *(p. 17)*

Test specifications

Test specifications, also known as test blueprints, have long been used in test development (see Jin, this volume). The original focus was content coverage and item types. Bloom's (1956) taxonomy of learning objectives incorporated considerations of cognition as well. Continued extensions provide additional support for the many specific decisions a test developer must make. In language testing, Bachman and Palmer's (2010) framework of task characteristics describes the features of target language use and assessment tasks. Davidson and Lynch's (2001) book on test specifications helps developers create a detailed blueprint for writing tasks with specified properties to suit the purposes, constraints, and resources of a particular testing context.

Assessment engineering

Another line of research is developing tools to improve the efficiency and quality of task design. Hively *et al.*'s (1968) item shells, Haladyna and Shindoll's (1989) item forms, and Bormuth's (1970) transformational-grammar item generation algorithms were early theory-based approaches to task development. Embretson (1985) illuminated a path toward integrating test theory, task design, and cognitive psychology. Irvine and Kyllonen (2002)

review articles and illustrative examples of theory-based item generation through 2000. In more recent work, the physical features of tasks that are to be varied reflect the features of tasks that are important cognitively (e.g., Gierl *et al*., 2008). Embretson's (1998) cognitive assessment design system and Luecht's (2013) integrated test design, development, and delivery approach are two examples of technology supports for principled task design.

Definitions

The terminology in ECD is best understood through its representation of layers in the design and implementation of educational assessment (Mislevy *et al*., 2002, 2003). Table 17.1 summarizes the roles and key entities of layers, which are discussed later in the context of language assessment.

Domain analysis

Domain analysis marshals information that has implications for assessment in the targeted domain. In Toulmin's terminology, *domain analysis* provides the backing for assessment arguments. In language testing, this means forms, meanings and use, assemblies of knowledge, skills and competencies that language learners typically engage with in target language use situations, and features of tasks that invoke them.

Table 17.1 Summary of ECD layers in the context of language testing

Layer	*Role*	*Key entities*	*Examples of knowledge representations*
Domain analysis	Gather substantive information about what is to be assessed.	Representation forms; analysis of language learning and language use in specific context or content area.	Language framework; language curriculum; proficiency guidelines; syllabi.
Domain modeling	Express assessment argument in narrative form.	Knowledge, skills, and capabilities; task features; potential observations.	Toulmin diagram; design patterns; claims, evidence, and task worksheets.
Conceptual assessment framework	Express assessment argument in specifications for tasks, measurement models, scoring procedures.	Student, evidence, and task model; task model variables; rubrics; measurement models; test specifications.	Task shells; task templates; graphical representation of measurement models.
Assessment implementation	Implement assessment, including writing tasks and fitting measurement models.	Task materials; pre-test data for item and task scoring.	Items and tasks; design of test materials and score reports; item parameter data files.
Assessment delivery	Coordinate interaction of students and tasks; task- and test-level scoring and reporting.	Task as presented; responses as collected; scores as evaluated.	Rendering of test materials; individual and group-level reports; data files.

Source: Adapted from Mislevy and Riconscente (2006)

The inferences we wish to draw about examinees and the psychological perspective we take on language acquisition and use serve as the starting point for *domain analysis*. If the inferences concern learners' ability to perform certain kinds of real-world tasks, the domain analysis will necessarily involve analysis of prototypical real-world tasks, or needs analysis (Long, 2015), and examine target language use (TLU) situations (Bachman and Palmer, 2010). Task-based language assessments, framed in the behaviorial perspective, focus on features of tasks and characterize examinees in terms of their performance in those situations. In contrast, language aptitude tests are cast in trait terms and support inferences about individuals' possession of capabilities conceptualized as latent traits. The *domain analysis* for these tests is thus an analysis of abilities that capture individual between-person differences in language aptitude. In *domain analysis*, we draw on the perspectives and expertise of language educators and researchers to identify characteristics of knowledge, skills, and competencies that are central to assessment design. Analysis of task features and competency-task relationships also occur in this layer. Curriculum materials and standards, language frameworks, and national and regional language proficiency standards are potential sources of information as well.

Domain modeling

The *domain modeling* layer articulates the assessment argument. It does not address technical details such as measurement models or data formats, but rather ways to fill in an argument schema using information identified in *domain analysis*. Proficiency, evidence, and task paradigms; Toulmin diagrams; and design patterns are schematic tools for this purpose. The paradigms are non-technical sketches of the student, evidence and task models discussed in the section on the next layer.

Figure 17.2 is an extended Toulmin diagram in the context of a task-based language assessment. The claim is an aspect of a language user's capabilities. The evidence is particular things the user says or does in the assessment situation. In this diagram, Student X has completed a number of assessment tasks that approximate the real-world language

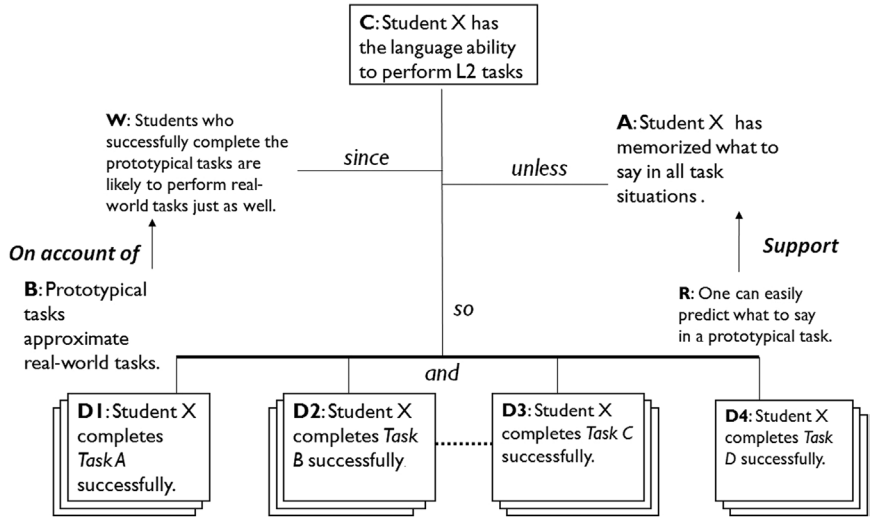


Figure 17.2 Extended Toulmin diagram for assessment

use situations of interest. A behaviorial perspective characterizes students in terms of what responses are produced under what conditions and postulates how likely they are to complete real-world tasks. Backing comes from practical experience and language acquisition research (e.g., Bachman and Cohen, 1998). Patterns of responses across a range of tasks are an indication of what tasks Student X performs well on and those which the student is not able to complete. Alternative perspectives around which assessment arguments can be cast are trait and interactionist (Bachman, 2007; Chalhoub-Deville, 2003; Mislevy and Yin, 2009). The latter is prominent in assessments of language for special purposes, where language capabilities are entwined with the knowledge and activities of a domain (Douglas, 2000; O'Sullivan, 2012). In each case, the nature of the construct shapes the argument, in terms of claims, task features, and evaluations of performances.

Design patterns organize information from *domain analysis* to define a design space for a task designer who needs to write tasks addressing the capabilities at issue (Mislevy, Hamel *et al.,* 2003; Riconscente *et al*., 2015). Table 17.2 shows the attributes of a design pattern. Focal KSAs are the capabilities about which the tasks will be constructed to assess (i.e., the claims). Characteristic features of tasks must be included in some form in order to evoke evidence of the targeted capabilities, and variable features are ways designers can shift focus or difficulty. Potential work products are the kinds of things student can say, do, or make to produce evidence, and potential observations are the qualities of the work that constitute evidence (e.g., accuracy, fluency, appropriateness, correctness of inferences). Table 17.3 is an example of a design pattern for an aspect of reading proficiency, reasoning through cause and effect structures in text.

Table 17.2 Design pattern attributes and relationships to assessment argument

Attribute	*Description*	*Assessment argument component*
Rationale	Provide a rationale for linkage between potential observations and focal knowledge, skills, and abilities (KSAs)	Warrant
Focal knowledge, skills, and abilities (focal KSAs)	The primary knowledge/skill/abilities targeted by this design pattern.	Student model
Additional knowledge, skills, and abilities	Other knowledge/skills/abilities that may be required by tasks generated under this design pattern.	Student model, if construct relevant; alternative explanations, if construct irrelevant
Characteristic features of tasks	Aspects of assessment situations that must be present in the task situation in order to evoke the desired evidence.	Task model
Variable features of tasks	Aspects of assessment situations that can be varied in order to shift difficulty or focus.	Task model
Potential observations	Characteristics of performance that constitute evidence about KSAs (e.g., accuracy, fluency, correct inference).	Evidence model
Potential work products	Performances by which students can produce evidence about focal KSAs (e.g., written text, response choice, oral speech)	Task model

Table 17.3 A design pattern for assessing cause-and-effect reasoning reading comprehension

Summary	This design pattern motivates tasks that require students to reason through cause-and-effect schema expressed in the conventions of written English to make inferences or predictions or provide explanations.
Focal KSAs	Knowledge of how real or imaginary phenomena are connected in terms of causation, what makes an event a cause and what makes an event an effect, and how causal connections are expressed in propositions; Skill to distinguish various forms of causal connections, distinguish causal connections from logical or temporal connections, distinguish signal words for cause-effect relation from those of the other relations, organize propositions as indicated by rhetorical structure, marker words; Ability to reason in inductive or deductive mode for predictions or explanations, trace relationship among causally connected events whether stated explicitly or implicitly, hypothesize in cause-and-effect structures, and recognize signal words for cause-effect relationship.
Add'l KSAs	Ability to decode and to respond orally or in writing. Knowledge of various types of text passages and notations, signs, and symbols used in specific disciplines. Familiarity with the substantive situation being addressed.
Characteristic features of tasks	Prose presentation of a situation involving one or more cause-and-effect relationships; directive requiring examinee to reason about or through those relationships.
Variable features	Nature of C&E relationship(s) with regard to Lakoff's (1987) canonical C&E schema. Structure of C&E relationship(s): simple, chained, multiple effects, multiple possible causes, multiple conjunctive causes. C&E relationship(s) implicit, explicit, or mixed? Direction of reasoning required – cause to effect or effect to cause? Length and complexity of prose passage. Degree of substantive knowledge required.
Potential work products	Selection of adverbials or answer choices, completion of table/graphic organizer, sequencing cards or pictures to structure a narrative, oral or written explanations or predictions, identifying missing elements, retell a story.
Potential observations	Is the C&E relationship appropriately explained or distinguished from other forms of relationship or correctly constructed when the events are not presented in canonical order? Were C&E text conventions used appropriately? Are there conceptual errors or misunderstandings that indicate the C&E relationship has not been properly constructed? Are conventions and rhetorical structures being used appropriately in explanation?

Source: Adapted from Mislevy and Rahman (2009)

Conceptual assessment framework

The *conceptual assessment framework* (CAF) provides a structure for coordinating the substantive, statistical, and operational aspects of an assessment. The operational elements of assessment machinery are specified in the form of three models, which instantiate the assessment argument. The student model operationalizes claims in terms of student variables,

which can be as simple as an overall score or as complex as the person variables in an item response theory or cognitive diagnostic model. Values assigned to student model variables, based on observing performances, correspond to claims about students' capabilities.

The evidence model contains two components: the evaluation component and the measurement model. The *evaluation component* concerns how to identify and characterize salient features of examinee performances – that is, how to score task performances. Through evaluation procedures using answer keys, rating rubrics with examples, automatic scoring procedures, and so on, we determine the values of observable variables from students' work products. What aspects of student work are important and how to evaluate them are governed by the psychological perspective in which the assessment is framed. A test cast in behavioral terms, with its focus on specified aspects of performances in specified contexts, may focus on qualities such as accuracy or complexity of the responses. Tests cast in interactionist/socio-cognitive terms highlight the appropriateness and effectiveness of the choices an examinee makes in light of the language use situation. In interactive tasks, evaluation may draw on natural language processing methods, learning analytics, or data-mining techniques and may comprise multiple stages (as in the automated scoring of spoken responses; Higgins *et al*., 2011). The *measurement model* provides a mechanism for synthesizing such data across tasks or interactions and for quantifying the strength of evidence for claims, such as reliability coefficients and standard errors of measurement. Psychometric models such as classical test theory, item response theory (Wilson, 2005; Ockey, this volume), and cognitive diagnosis models (Lee and Sawaki, 2009; Jang and Sinclair, this volume) are used in this step.

The task model consists of blueprints for assessment tasks and work products. In applications using design patterns, task models reflect choices from the suggested range of potential work products and characteristic and variable features and potential work products and observations. This layer specifies work products and narrows down to central and optional task features that evoke evidence about the targeted claims. For example, performance-based assessment of student teachers will necessarily involve observations of a student teaching a language class and evaluations of qualities of lesson planning and lesson delivery. In contrast, if the claim concerns mastery of content or pedagogical knowledge, a computer-delivered knowledge assessment, in the form of multiple-choice questions or short-essay questions may serve the purpose. In assessments consisting of multiple discrete tasks, constellations of tasks are brought together as an assessment in the *assembly model* to determine the breadth and depth of the domain being assessed. Test specifications and blueprints are generated in this model on the basis of CAF models. In interactive tasks such as simulations and games, specifications for how interactions between test takers and the task environment are effected, data are captured, and evaluations are produced are specified in more complex forms of these basic ECD models (Conrad *et al.,* 2014; Ke *et al.,* 2019).

The implementation and delivery layers

The *assessment implementation* layer includes activities such as developing interfaces, authoring tasks, fine-tuning rubrics and examples, programming automated scoring routines, and fitting measurement models. It is in this layer that assessment-engineering developments in technology for creating such elements as adaptive testing, interaction in simulation environments, and capture and evaluation of complex performances are effected, in accordance with specifications laid out in the CAF. Behrens *et al*. (2012), for instance,

provide worked-through examples of how Cisco Networking Academies designed its next-generation automated performance-based assessment, the NetPass system, using the ECD framework.

The *assessment delivery layer* includes processes for selecting and administering test items, interacting with examinees, reporting scores, and providing feedback to appropriate parties. The information, structure, and rationale of these processes embody the assessment argument laid out in *domain analysis*, are structured by the blueprint laid out in the CAF, and utilize the operational elements created in assessment implementation. Details of delivery system architecture are provided in Almond *et al*. (2002) and, as extended to interactive digital assessments, in Conrad *et al*. (2014).

Critical issues and topics

Conceptions of capabilities and competencies shape the assessment argument. A perspective on language learning and language use suggests the kinds of inferences we should draw about examinees, from what kinds of evidence, and in what kinds of assessment situations (Mislevy, 2018; Mislevy and Yin, 2009).

Definition of constructs

Claims in language testing are usually phrased in terms of constructs and reflect our intended ways to construe aspects of examine capabilities and competencies. In language testing, approaches vary as to the nature of the construct to be assessed (Bachman, 2007). From the trait perspective, a construct is defined in terms of aspects of examinees' capabilities that are stable across a range of situations. Constructs in task-based language assessment are cast in the behavioral perspective, phrased in terms of features of examinee performances in tasks with specific kinds of features (Mislevy, 2018; Norris *et al.,* 1998). An interactionist approach views language activities as discursive meaning co-construction practices that all participants engage in, using a conglomeration of linguistic, interactional, and identity resources (Young, 2008). This last approach is compatible with a socio-cognitive perspective on language development, seeing a holistic, adaptive process that integrates the social, the physical, and the cognitive (Atkinson, 2002; O'Sullivan, 2012; Weir, 2005). A construct of interest in this approach can be phrased in terms of resources that enable the language event.

A construct in an assessment is the assessor's inevitably simplified conception of examinee capabilities, chosen to suit the purpose of a given assessment. We can think of examinee capabilities in terms of resource employment during a language event and describe them in terms of the extra-personal linguistic, cultural, and substantive patterns (LCS patterns; Mislevy, 2016, 2018; Mislevy and Yin, 2009) learners can usefully engage with in various contexts (Larsen-Freeman and Cameron, 2008). Approaches to construct definition can all be understood in terms of resource assembly, of LCS patterns that relate to language use. From an interactionist perspective, we see peoples' capabilities, contexts, and performances as intertwined in a constantly evolving process (Chalhoub-Deville, 2003; Chapelle, 1998). Any approach to construct definition necessarily involves all three, explicitly or implicitly. The assessor's intentions determine what the construct is, which in turn holds implications for design decisions in regard to tasks, evaluation procedures, and the measurement models used to aggregate evidence.

The central role of context

Use of language is enmeshed with the context in which a language event takes place. Language use patterns emerge as humans "soft assemble" resources in response to a communicative demand and are transformed through further usage (Larsen-Freeman and Cameron, 2008). Form-meaning-use mapping occurs as a result of ongoing interaction among the social, cultural, and psychological, as individuals become attuned to linguistic and cultural norms and conventions. A socio-cognitive view argues for a coherent coupling of and coordinated interaction between the individual and the context. Context as co-constructed by multiple participants in a communicative event evolves through ongoing meaning negotiation activities; through such interactions do individuals' capabilities develop (Douglas, 2000; Young, 2008).

Unobserved yet central to assessment is the examinee's internal perception of the context that presents externally as an assessment task (Douglas, 2000). We wish to examine the extent to which the examinee can assemble resources and activate internal patterns that go with the extra-personal LCS patterns required for an assessment task. The examinee's internal patterns structure and mediate her behaviors, which are evaluated by the standards of the external LCS patterns the assessor considers germane. The inference thus concerns the observed "mental context" (Council of Europe, 2001: 50) that combines unique features of the assessment situation with internal patterns at all levels she has built up from her past experiences. Claims and inferences are thus high-level descriptors of (1) the degree to which assessment tasks align with TLUs and (2) the extent to which targeted patterns are activated across assessment tasks.

Current contributions and research

Language tests developed using ECD

In language testing, ECD has been applied in the Test of English as a Foreign Language internet-Based Test (TOFEL iBT; Pearlman, 2008), the Test of English for International Communication (TOEIC) Speaking and Writing test (Hines, 2010), the College Board's (n.d.) Advanced Proficiency (AP) Spanish Language Exam, and the WIDA consortium's Alternate English Language Learner Assessment (ALTELLA). The procedures Educational Testing Service used to develop test blueprints for the TOEFL iBT illustrate ECD principles in a large-scale assessment of English language learners (ELLs) (Chapelle *et al.,* 2008). The step of defining the construct to be measured and the claims and subclaims to be made about examinee abilities illustrates a proficiency paradigm from domain modeling. An example of a claim for the speaking test is that the examinee can "communicate effectively by speaking in English-speaking academic environments" (Pearlman, 2008: 232). Subsequent steps characterize observations for grounding them and identify real-world tasks to provide evidence, resulting in evidence paradigms and task paradigms from domain modeling. Based on the results of these steps, templates for generating items or task shells are constructed to incorporate essential elements of task models into the conceptual assessment framework.

Expansions to ECD

The advent of computer technology and the availability of rich data have brought forth myriad opportunities for assessment designers to apply and expand on ECD to develop

innovative complex learning and assessment systems (e.g., Arieli-Attali *et al*., 2019; Behrens *et al*., 2012; Mislevy *et al*., 2012). One way assessment designers expand on ECD is to utilize rich data generated from simulation and game-based assessment to inform scoring rules and game design (Conrad *et al*., 2014; Grover *et al*., 2017; Shute *et al*., 2010). Adaptations also include retrofitting extensions to learning systems, or "ECD for Learning" (ECDL); Arieli-Attali *et al*. (2019) and Feng *et al*. (2009), for example, extend ECD with pedagogical models to utilize data within existing learning products. More generally, the expanded ECD framework (e-ECD; Arieli-Attali *et al*., 2019) expands all three basic models – for students, evidence, and tasks – by adding an additional learning layer in each model to specify targeted learning objectives, to provide task-based learning support, and to capture evidence for changes in learning. Such adaptations and expansions point to important ways the ECD framework can be applied to coherently blend learning and assessment and simultaneously inform and guide learning in an era when the use of computer technology and digital devices has become pervasive and ubiquitous.

Main research methods

This section describes ways that ECD has been applied with processes and representations that vary in details and the foci with which the framework is operationalized.

Claims and evidence

The TOEFL iBT final blueprints and TOEIC speaking and writing tests show how fine-grained pieces of evidence are marshaled to support claims and subclaims (Hines, 2010; Pearlman, 2008). ECD principles guide the various stages and procedures employed in the design of the two tests. Table 17.4 summarizes the steps. Note the effective use of proficiency, evidence, and task paradigms. This work articulated claims and subclaims in the domain and laid out sources of evidence, tasks for eliciting them, and task features that discriminate among a range of capabilities. A particular focus was the more technical step of creating task shells for generating parallel items.

Luecht's (2003) design procedures similarly align a theoretical construct model, a task development model, and a psychometric scoring model. The theoretical construct model defines the knowledge, skills, and ability space to be assessed – in effect, a proficiency paradigm. He organized an illustrative theoretical model of language proficiency around the Interagency Language Roundtable (ILR) proficiency scale to define five ability dimensions: vocabulary, grammar, pronunciation, fluency/integrative, and sociolinguistic ability, each organized into five ordered levels. The two components of the task development model are task development specifications and test assembly specifications, both CAF objects from the ECD perspective. Task development specifications provide a detailed schema for stimulus selection, item writing, and expected responses within each proficiency level. The psychometric model specifies the scoring procedures and uses measurement models to relate observations to the underlying construct. Iterative use of these procedures thus aligns constructs with test development practices and psychometric models, to address the complexity of language proficiency assessment.

The PADI tools and representations for working through an evidence-centered design are described in Riconscente *et al*. (2015). The illustrations are in the domain of science, but the approach is applied in the same manner for language testing.

Table 17.4 Steps taken to redesign TOEFL iBT and TOEIC speaking and writing tests and guided by layers in evidence-centered design

Steps in TOEFL iBT and TOEIC redesign		*Components from evidence-centered design*	*Layers in evidence-centered design*
Task design analysis	Reviewing prior theory and research	Domain documents	Domain analysis
	Articulating claims and subclaims about language proficiency in all domains	Proficiency paradigm	Domain modeling
	Listing sources of evidence for each claim and subclaim	Evidence paradigm	
	Identifying real world tasks that provide evidence	Task paradigm	
	Identifying task characteristics affecting task difficulty	Task paradigm	
	Identifying criteria for evaluating performance on the tasks	Task paradigm	
Constructing task shell	Reiterating claim and subclaims	Proficiency paradigm	From domain modeling layer
	Defining aspects of responses that are the focus of evaluation	Measurement model	Conceptual assessment framework
	Specifying fixed elements of tasks	Task model	
	Specifying variable elements of tasks	Task model	
	Composing rubric	Evaluation component	
	Describing example tasks	Task model	

Source: Adapted from Hines (2010)

Reusability and modularity

Taking advantage of the common conceptual language and representational forms of ECD can enhance efficiency as well as validity by exploiting the principles of reusability, modularity, and interoperability. Whenever design and implementation are required, the designer seeks to express the results in forms that are explicit and can be more readily adapted in further work in the same or other assessments.

Laying out assessment arguments and creating design patterns for hard-to-assess aspects of learners' capabilities help make explicit the design choices that have shaped an assessment and supports those choices (Mislevy *et al*., 2003). This is re-use of the conceptual structures that underlie assessments. Developing fragments of measurement models that express recurring evidentiary relationships allows psychometricians and content specialists to work together to design many new tasks with the same underlying measurement models (Mislevy *et al.,* 2002). This modular approach illustrates the concept of re-use for the evidentiary foundation of complex tests: complex tasks that differ on the surface but share the same evidentiary structure can use the same pieces of scoring procedures and measurement

models. Creating models for tasks and data structures for implementing and delivering them thus reduces programming and testing time for computer-based tests (Luecht, 2003). For existing assessments, understanding a test in terms of explicit arguments and modular elements makes it easier to adapt and reassemble elements for new purposes (Fulcher and Davidson, 2009). The projects mentioned in this chapter have used ECD in ways that make explicit the underlying arguments and the role the elements played in the operational test and expressed the results in forms that support further development.

Assessment use arguments

Assessment use argument (AUA; Bachman, 2005; Bachman and Palmer, 2010) extends the argument-based approach beyond the ECD design phase to assessment use and validation. Examinees' performance in the assessment situation is intended to characterize their capabilities for specific purposes in social practices and institutions (Mislevy, 2012) and predict their performance in some other situations, known as criterion situations. AUA addresses the social context in which an assessment is meant to be used. Essential elements of an AUA again follow the same Toulmin argument schema. Claims in an AUA concern intended uses of an assessment and are justified in terms of beneficence of consequences and equitability and value sensitivity of decisions made based on assessment results. For example, results of foreign language proficiency assessment can be an indication of whether language benchmarks have been met and, consequently, can be used for purposes such as evaluating outcomes of instruction and program effectiveness. Three kinds of data in AUA are summary statements about student performances in the assessment situation, data concerning assessment use situation, and other information concerning student vis-à-vis use situation (Mislevy, 2018). A coherent chain of reasoning connects essential elements in the assessment design argument and AUA. Examples of assessment use arguments for actual assessments can be found in Llosa (2008) and the Chapelle *et al*. (2008) volume on the TOEFL redesign.

Recommendations for practice

A developer wishing to use ECD to develop a language test should browse both the more conceptual discursions and the illustrative applications cited in this chapter. The conceptual references provide a common grounding of the essential elements of ECD, which underlie the variations seen in the applications. The applications show the ideas at work. Adapting the forms and the processes of a prototypical example to one's own project is often a useful way to begin. In all cases, the up-front attention paid to domain modeling more than pays for the time invested.

Claims, framed in terms of constructs, reflect the assessor's psychological grounding and the purpose or intended use of the assessment. As Messick indicated, attending to the nature of claims that are needed to serve an assessment's purpose is the starting point in design. For example, discrete-point grammar or vocabulary tests are designed to identify students' mastery of focused components of a language curriculum; results indicate whether instructional objectives have been met and guide instruction to address learning needs. Claims in general language proficiency tests are phrased in generic terms, such as examinees being able to speak or write at particular levels, with minimal descriptions of context. The purposes these tests are intended for are broad and span any number of TLUs. In contrast, tests specifying context of use are intended for more focused purposes. For instance, the Occupational English Test specifies examinee occupation as context of use and assesses examinee

competencies to perform tasks typical of a specific occupation, such as business or medicine (McNamara, 1996). Claims and inferences are contextualized and relate directly to the context for which they are intended. These considerations should guide test developers as they work through domain modeling, regardless of the particular ECD representations they employ (e.g., paradigms, Toulmin diagrams, design patterns).

Advances in language science inform and transform conceptions of language learning and language use. They call attention to aspects of context and language use that are inextricably connected and integral to language development. An interactionist perspective highlights the importance of context for language assessment and of the alignment between assessment tasks and TLUs needed to justify claims about the examinee performance in those TLUs. Focused analysis of features characterizing TLUs, of configurations of knowledge, skills, and competencies that underlie performance with reference to particular TLUs, is essential to task design because it is the representation of TLU features in assessment tasks that elicits an examinee's resources for the targeted LCS patterns. Task design requires principled patterning of features that invoke stable states and regularities, amid ongoing variation and change within and across individuals. An important aspect of these features is the "resource-directing dimension," defined as increasing demands on learners' cognitive resources while potentially directing them to aspects of language code to be used to complete the tasks (Robinson, 2009: 199).

The inescapable involvement of LCS patterns at all levels in any language event raises a tradeoff in tasks design as to how much and what kinds of context to include in tasks. Minimizing context reduces demands for necessary but unrelated knowledge and skills. In argument terminology, such demands introduce alternative explanations for poor performances; in measurement terms, they are possible sources of construct irrelevant variance. But decontextualized situations lack fidelity; students can perform well here, yet not apply what would seem to be the same skills in TLU situations (see Ross, this volume). Worse, instruction that emphasizes performances on decontextualized tasks draws attention from learning to using language in real-world situations. An alternative strategy that increases fidelity while minimizing alternative explanations is to use contexts that are known to be familiar to examinees. Assessing verb usage in a role-playing scenario of a familiar local festival is an example with younger students. Assessing reading proficiency with medical texts illustrates the idea with prospective healthcare professionals (Douglas, 2000).

Future directions

Scientific understanding of how people acquire and use language has advanced radically, and new technologies are coming online continually to provide interactive task environments and capture the resulting evidence. The challenge is harnessing this new knowledge for practical testing purposes – effectively, validly, affordably. Evidence-centered design provides a coherent conceptual framework that makes explicit the principles and interacting structures of argumentation, analysis, and operation that underlie assessment. It provides language and concepts to help developers design and coordinate the elements of familiar assessments as well as new ones. It motivates definitions of data structures and operational processes that can be used and re-used flexibly, across projects and organizations.

The work described in this chapter represents current practices in developing and using ECD. Future work will fill the framework out more fully, provide additional worked examples, and develop software tools and knowledge representations that make it easy for test developers to put the ideas to use. In the domain modeling layer, design patterns are needed

to provide and support performance assessments and simulation-based assessments that focus on interactive capabilities. In the CAF layer, we need task design frameworks, adaptable modules for measurement models, and automated or rubric-based evaluation procedures, all explicated in terms of their relationship to assessment arguments. Recent work mentioned earlier shows how researchers are extending the ECD framework to incorporate learning with assessment and to detail the more complex design and analysis interplay in game- and simulation-based assessments. In the implementation and delivery layers, we need further representations and data structures that exploit computer capabilities to build, administer, and evaluate tests. These are ambitious and broad undertakings, spanning years or decades. At the same time, however, the most fundamental ideas of ECD – carefully thinking through what capabilities one wants to assess and the kinds of things we need to see learners do in what kinds of situations – can be put into action immediately by every teacher and every test designer in the field.

Further reading

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B. and von Davier, A. A. (2019). The expanded evidence-centered-design (e-ECD) for learning and assessment systems: A framework to incorporating learning goals and processes within assessment design. *Frontiers in Psychology* 10: 853.
This article extends the ECD framework to learning. A model for learning is used to provide coordinated extensions to the ECD student, task, and evidence models. It is applicable when learning occurs both with and without explicit instruction.

Chapelle, C., Enright, M. K. and Jamieson, J. M. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language ™*. London: Routledge.
This volume presents an in-depth discussion of the application of some of the ideas of evidence-centered design and assessment use argumentation to a TOEFL redesign.

Luecht, R. M. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology* 14: 1–38.
The assessment engineering framework is compatible with ECD but provides additional structural support for task design, scoring, and psychometric modeling schemas.

Mislevy, R. J. (2018). *Sociocognitive Foundations of Educational Measurement*. New York & London: Routledge.
This monograph is an in-depth interdisciplinary treatise of foundational issues in assessment and measurement as re-examined from a socio-cognitive perspective.

Mislevy, R. J., Almond, R. G. and Lukas, J. (2004). *A Brief Introduction to Evidence-Centered Design*. CSE Technical Report 632. The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA. www.cse.ucla.edu/products/reports/r632.pdf.
This is a reader-friendly introduction to evidence-centered design. It is the best place for a new reader to start.

Mislevy, R. J., Steinberg, L. S. and Almond, R. A. (2002). Design and analysis in task-based language assessment. *Language Testing* 19: 477–96.
This is a more technical discussion of the application of evidence-centered design to task-based language tests. It addresses issues of construct definition, multivariate student models, and complex tasks.

Mislevy, R. J. and Yin, C. (2009). If language is a complex system, what is language assessment? *Language Learning* 59 (Supplement 1): 249–67.
This article discusses how to build and interpret assessment arguments from an interactionist/complex socio-cogntive systems perspective on language.

Riconscente, M. M., Mislevy, R. J. and Corrigan, S. (2015). Evidence-centered design. In S. Lane, T. M. Haladyna and M. Raymond (eds.), *Handbook of Test Development*. London: Routledge, 40–63.
This discussion of ECD is aimed at test developers. It uses representations and terminology from the Principled Assessment Design for Inquiry (PADI) project.

Shute, V. J., Masduki, I. and Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. *Technology, Instruction, Cognition & Learning* 8: 137–61.
In a series of articles and projects, Shute and her colleagues have used ECD in designing and implementing unobtrusive assessment in the context of games and simulations. This article explains and illustrates the approach.

Wilson, M. (2005). *Constructing Measures*. Mahwah, NJ: Lawrence Erlbaum.
This monograph is consistent with the ECD framework but goes more deeply into the definition, task modeling, and Rasch psychometric modeling of constructs. It is well suited for assessment built on learning progressions.

References

Almond, R., Steinberg, L. and Mislevy, R. (2002). Enhancing the design and delivery of assessment systems: A four-process architecture. *The Journal of Technology, Learning and Assessment* 1. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1671.

Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B. and von Davier, A. A. (2019). The expanded evidence-centered design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology* 10. https://doi.org/10.3389/fpsyg.2019.00853.

Atkinson, D. (2002). Toward a sociocognitive approach to second language acquisition. *The Modern Language Journal* 86: 525–45. https://doi.org/10.1111/1540-4781.00159.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly* 2: 1–34. https://doi.org/10.1207/s15434311laq0201_1.

———. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche and D. Bayliss (eds.), *Language Testing Reconsidered*. Ottowa, ON: University of Ottawa Press.

Bachman, L. F. and Cohen, A. D. (eds.). (1998). *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Bachman, L. F. and Palmer, A. S. (2010). *Language Assessment Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford: Oxford University Press.

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E. and Levy, R. (2012). Evidence centered design for learning and assessment in the digital world. In C. M. Michael, J. Clarke-Midura and D. H. Robinson (eds.), *Technology-Based Assessments for 21st Century Skills*. Charlotte, NC: Information Age Publishing.

Bloom, B. S. (ed.). (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I, Cognitive Domain*. London: Longman.

Bormuth, J. R. (1970). *On the Theory of Achievement Test Items*. Chicago: University of Chicago Press.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing* 20: 369–83. https://doi.org/10.1191/0265532203lt264oa.

Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research*. Cambridge: Cambridge University Press.

Chapelle, C., Enright, M. K. and Jamieson, J. M. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language ™*. London: Routledge.

College Board. (n.d.). *AP Course and Exam Redesign*. Retrieved from https://aphighered.collegeboard.org/courses-exams/course-exam-redesign.

Conrad, S., Clarke-Midura, J. and Klopfer, E. (2014). A framework for structuring learning assessment in an educational massively multiplayer online educational game – experiment centered design. *International Journal of Game-Based Learning* 4: 37–59. https://doi.org/10.4018/IJGBL.2014010103.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Cambridge University Press. Retrieved from www.coe.int/t/dg4/linguistic/source/Framework_EN.pdf.

Davidson, F. and Lynch, B. K. (2001). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT: Yale University Press.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

Embretson, S. E. (ed.). (1985). *Test Design: Developments in Psychology and Psychometrics*. New York: Academic Press.

———. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods* 3: 380–96. https://doi.org/10.1037/1082-989X.3.3.380.

Feng, M., Hansen, E. G. and Zapata-Rivera, D. (2009, April 13–17). *Using Evidence Centered Design for Learning (ECDL) to Examine the Assistments System*. Paper presentation. American Educational Research Association Annual Meeting, San Diego, CA.

Fulcher, G. and Davidson, F. (2009). Test architecture: Test retrofit. *Language Testing* 26: 123–44. https://doi.org/10.1177/0265532208097339.

Gierl, M. J., Zhou, J. and Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning, and Assessment* 7. Retrieved from http://escholarship.bc.edu/cgi/viewcontent.cgi?article=1137&context=jtla.

Grover, S., Bienkowski, M., Basu, S., Eagle, M., Diana, N. and Stamper, J. (2017). *A Framework for Hypothesis-Driven Approaches to Support Data-Driven Learning Analytics in Measuring Computational Thinking in Block-Based Programming*. Seventh International Learning Analytics & Knowledge Conference (ACM), New York, NY.

Haladyna, T. M. and Shindoll, R. R. (1989). Shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions* 12: 97–104. https://doi.org/10.1177/016327878901200106.

Higgins, D., Xi, X., Zechner, K. and Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech & Language* 25: 282–306. https://doi.org/10.1016/j.csl.2010.06.001.

Hines, S. (2010). *Evidence-Centered Design: The TOEIC® Speaking and Writing Tests*. Educational Testing Service. www.ets.org/research/policy_research_reports/publications/report/2010/itjx.

Hively, W., Patterson, H. L. and Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement* 5: 275–90. Retrieved from www.jstor.org/stable/1433778.

Irvine, S. H. and Kyllonen, P. C. (eds.). (2002). *Item Generation for Test Development*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kane, M. (2006). Validation. In R. J. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: Praeger.

———. (2013). Validating the interpretation and use of test scores. *Journal of Educational Measurement* 50: 1–73. https://doi.org/10.1111/jedm.12000.

Ke, F., Shute, V., Clark, K. M. and Erlebacher, G. (2019). *Interdisciplinary Design of Game-Based Learning Platforms*. Heidelberg, Germany: Springer Nature.

Larsen-Freeman, D. and Cameron, L. (2008). *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.

Lee, Y. W. and Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly* 6: 172–189. https://doi.org/10.1080/15434300902985108.

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educational Measurement: Issues and Practice* 27: 32–42. https://doi.org/10.1111/j.1745-3992.2008.00126.x.

Long, M. H. (2015). *Second Language Acquisition and Task-Based Language Teaching*. Oxford: Wiley-Blackwell.

Luecht, R. M. (2003). Multistage complexity in language proficiency assessment: A framework for aligning theoretical perspectives, test development, and psychometrics. *Foreign Language Annals* 36: 527–35. https://doi.org/10.1111/j.1944-9720.2003.tb02142.x.

———. (2013). Assessment engineering task model maps, task models and templates as a new way to develop and implement test specifications. *Journal of Applied Testing Technology* 14: 1–38. http://jattjournal.com/index.php/atp/article/view/45254/36645.

McNamara, T. F. (1996). *Measuring Second Language Performance*. Harlow: Addison Wesley Longman.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 32: 13–23. https://doi.org/10.3102/0013189X023002013.

Mislevy, R. J. (2012). Modeling language for assessment. In C. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. Hoboken, NJ: Wiley-Blackwell.

———. (2016). How developments in psychology and technology challenge validity argumentation. *Journal of Educational Measurement* 53: 265–92. https://doi.org/10.1111/jedm.12117.

———. (2018). *Sociocognitive Foundations of Educational Measurement*. London: Routledge.

Mislevy, R. J., Behrens, J. T., DiCerbo, K. E. and Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining* 4: 11–48. https://doi.org/10.5281/zenodo.3554641.

Mislevy, R. J., Hamel, L., Fried, R. G., Gaffney, T., Haertel, G., Hafter, A. . . . Wenk, A. (2003). *Design Patterns for Assessing Science Inquiry (PADI Technical Report 1)*. SRI International. Retrieved from http://padi.sri.com/downloads/TR1_Design_Patterns.pdf.

Mislevy, R. J. and Rahman, T. (2009). *A Design Pattern for Assessing Cause and Effect Reasoning in Reading Comprehension (PADI Technical Report 20)*. Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S. and Almond, R. A. (2002). Design and analysis in task-based language assessment. *Language Testing* 19: 477–96. https://doi.org/10.1191/0265532202lt241oa.

———. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives* 1: 3–67. https://doi.org/10.1191/0265532202lt241oa.

Mislevy, R. J. and Yin, C. (2009). If language is a complex system, what is language assessment? *Language Learning* 59 (Supplement 1): 249–67. https://doi.org/10.1111/j.1467-9922.2009.00543.x.

Norris, J. M., Brown, J. D., Hudson, T. D. and Yoshioka, J. K. (1998). *Designing Second Language Performance Assessment*. Honolulu: University of Hawai'i Press.

O'Sullivan, B. (2012). Assessment issues in languages for specific purposes. *The Modern Language Journal* 96: 71–88. https://doi.org/10.1111/j.1540-4781.2012.01298.x.

Pearlman, M. (2008). Finalizing the test blueprint. In C. Chapelle, M. K. Enright and J. M. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language ™*. London: Routledge.

Riconscente, M. M., Mislevy, R. J. and Corrigan, S. (2015). Evidence-centered design. In S. Lane, M. R. Raymond and T. M. Haladyna (eds.), *Handbook of Test Development*, 2nd edn. London: Routledge.

Robinson, P. (2009). Task complexity, cognitive resources, and syllabus design. In K. V. den Branden, M. Bygate and J. M. Norris (eds.), *Task-Based Language Teaching: A Reader*. Amsterdam: John Benjamins.

Shute, V. J., Masduki, I. and Donmez, O. (2010). Conceptual framework for modeling, assessing and supporting competencies within game environments. *Technology, Instruction, Cognition & Learning* 8: 137–61.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Weir, C. J. (ed.). (2005). *Language Testing and Validation: An Evidence-Based Approach. Research and Practice in Applied Linguistics*. Basingstoke: Palgrave Macmillan.

Wilson, M. (2005). *Constructing Measures*. Mahwah, NJ: Lawrence Erlbaum.

Young, R. F. (2008). *Language and Interaction: An Advanced Resource Book*. London: Routledge.

18

Accommodations and universal design

Jamal Abedi

Introduction

Accommodations make assessments more accessible for ELL students. To help reduce the impact of construct-irrelevant sources in the assessment of ELL students, accommodations should (1) be effective in making assessments more accessible for ELL students and (2) provide assessment outcomes that are valid: i.e., comparable with those of non-ELLs with no accommodations provided. If these conditions are met, then these accommodations can help develop a universally designed assessment. Universal learning applies diverse teaching methods in order to incorporate "student diversity within the classroom context" (Capp, 2017). Since educators are searching for improved ways to teach diverse students, it is essential to develop assessments that can be readily available and applicable to various groups of students. For example, one of the frequently used accommodations for ELL students is linguistic modification of assessments. In this assessment, unnecessary linguistic complexities are identified, and the level of complexities is reduced. Based on the literature, this accommodation does not alter the focal construct and makes assessments more universally applicable to students. Therefore, the concept of accommodation can be extended to include accommodations that do not alter the construct being measured and can be used for everyone, including ELLs and students with disabilities. In other words, these accommodations can improve the performance of all students and make assessments more usable and universal, given that the accommodations do not alter the focal construct.

In this chapter, we will discuss the commonly used accommodations and determine if any of these accommodations can be used to make assessments universally designed to be accessible to all students, including ELLs.

Historical perspectives

Many test providers currently use accommodations for ELL students (Abedi, 2007; Abedi *et al*., 2004; Francis *et al*., 2006; Rivera *et al*., 2000; Sireci *et al*., 2003; Willner *et al*., 2008; Wolf *et al*., 2008). The widespread use of accommodations, however, raises a number of

 DOI: 10.4324/9781003220756-24

issues and questions. Does the use of accommodations provide valid inferences about ELL students' content knowledge? Which students are eligible to receive accommodation(s), and what criteria are used to decide their eligibility? Do accommodations provide an unfair advantage to the recipients? Is it meaningful to aggregate accommodated and non-accommodated scores for all students? When we look for answers to these questions throughout studies of content area assessments, we are confronted with a striking lack of empirical research.

The following two questions can help us determine the *accessibility* of an assessment.

1 How effective are the accommodations in making assessments more representative of ELL student content knowledge (effectiveness)?
2 Are the accommodated assessments measuring the same construct as the non-accommodated assessments (validity)?

To shed light on these questions, one must examine these issues using studies based on sound research methodology. Among the necessary conditions for such experiments are random assignment of ELL and non-ELL students to different accommodation conditions.

Accommodations that are effective and valid can be used to develop universally designed assessments that are accessible for all students. For example, Goegan *et al*. (2018) explain that universally designed accommodations should take into account the following areas: setting, timing, presentation of the test, and the students' method of response. For example, the most commonly used accommodations for ELL students in the United States are extended time (42 of the 50 states); use of a glossary (26 states), an English dictionary (33 states), or a bilingual dictionary (22 states); and linguistically simplified test items (12 states) (Rivera and Collum, 2005). Willner *et al*. (2008) distinguished between accommodations that are relevant for ELLs and those that may be irrelevant for these students. Research results are available only for a few forms of accommodations while others have yet to be examined. For example, research has shown that providing an English dictionary (Abedi *et al*., 2003a) and extra time (Abedi *et al*., 2004; Hafner, 2001; Thurlow, 2001) affects the performance of all students, including those for whom the accommodations are not intended (see also Maihoff, 2002; Thurlow and Liu, 2001). To aggregate the results of accommodated and non-accommodated assessments, one must have enough evidence on the validity of the accommodations.

Main research methods

The summary of literature presented earlier clearly shows mixed results on the effectiveness and validity of many forms of the accommodations currently used for ELL students. These mixed results could be attributed to several issues with the design of these studies. First are the issues regarding samples of students for these studies. If the sample size is small or the sample is not representative of the population in question, then inconsistent results could emerge. The inconsistencies in the outcomes of the studies described earlier may also be due to the impact of extraneous variables on the study that are considered threats to the internal and external validity of the study that may not have been properly controlled (Gay, 1981). To help understand issues in the use and interpretation of accommodations, we will discuss the methodological issues concerning accommodations that were reflected in the two research questions presented earlier.

Question 1: examining the effectiveness of accommodations for ELL students

The concept of accommodation was developed and applied mainly in the assessment of students with disabilities, and some of the accommodations that are used for ELL students were originally designed and implemented for students with disabilities (Willner *et al*., 2008). Therefore, they may be irrelevant for ELLs and may not be effective at improving ELL student performance. For example, among the 73 accommodations used for ELL students (Rivera and Collum, 2005), only 11 (15 percent) of them were found to be relevant for ELL students (Abedi, 2006). The rest of these accommodations (62) did not seem to be relevant or did not even have face validity for ELL students.

An accommodation is effective for ELL students if it helps reduce the sources of construct-irrelevant variance and makes the assessment more accessible to these students. Therefore, it is imperative to assess ELL students' needs and the ways that accommodations could help them when selecting effective accommodations. Unlike students with disabilities, who are grouped under several different categories with different types of disabilities and different needs, ELL students share a common characteristic: they all need linguistic assistance. Therefore, accommodations that can directly address their linguistic needs will be the most relevant accommodations for this group.

Effective accommodations for ELL students should help provide assessments that are more accessible to these students by reducing unnecessary linguistic complexity. The effectiveness of accommodations for ELL students can be examined through an analysis of performance differences between ELL students assessed under the accommodation conditions (the treatment group) and ELL students tested under the standard testing condition with no accommodations provided (the control group). This comparison should be conducted through a randomized experiment in which students are assigned randomly to the treatment (accommodated) and control (non-accommodated) groups. Such random assignment (assuming a large enough sample size) would control for many sources of threats to the internal validity of the design for testing effectiveness of accommodations. By assigning students randomly to the treatment and control groups, students' initial differences can be controlled.

Unfortunately, the results of studies on the effectiveness of accommodations have illustrated that some of the accommodations that are shown to be effective may alter the focal construct: the construct being measured (Abedi *et al*., 2000; Sireci *et al*., 2003). Therefore, it is essential to interpret the evidence on the effectiveness of accommodations in light of the validity evidence for the accommodations.

Research design for examining effectiveness of accommodations for ELLs

ELL students should be randomly assigned either to a treatment group in which they receive accommodations or to a control group in which they are tested under the standard testing condition in which no accommodation is provided. Evidence of improvement in the performance of ELL students under the accommodated condition may be an indication of the effectiveness of the accommodation used. It is also important to control for any other sources of variation that could affect the outcome of the effectiveness study. For example, studies on accommodations are often conducted in intact classrooms. There are some

expected sources of variability due to classroom characteristics (e.g., teacher background, average socioeconomic status, percentage of ELLs, percentage of students with disabilities) and school and district characteristics and sources that can be controlled statistically with a hierarchical linear modeling approach and other statistical methodologies.

Question 2: examining the validity of accommodations for ELL students

Accommodations are provided to ELL students to reduce construct-irrelevant variance from the assessment outcomes, thus making assessments more accessible to ELL students. A major source of construct-irrelevant variance is the language factors in the assessments of ELL students. Unnecessary linguistic complexity in the assessment introduces an additional dimension into the assessment model, which may be unrelated to the focal construct. Therefore, one approach in examining the validity of accommodations is to compare the performance of non-ELL students under the accommodated and non-accommodated conditions. If the performance of non-ELL students, for whom accommodations are not intended, changes significantly under the accommodated condition, then the accommodation is believed to do more than it is intended to do. The accommodated outcomes under such situation may not be valid. Another approach in examining the validity of accommodations is to compare the structural relationship of the assessment outcomes with a set of valid external criteria under accommodated and non-accommodated conditions.

While there is not enough evidence to judge the validity of all the accommodations used for ELL students, results of existing research suggest that some of the accommodations used for ELL students and students with disabilities may alter the construct being measured (Abedi, 2007; Bielinski *et al*., 2001; Meloy *et al*., 2000). For example, *read-aloud* accommodation has been shown to be effective for many students with reading disabilities, but there are concerns over the validity of this accommodation. Košak-Babuder *et al*. (2019) studied the effect of reading aloud on the text comprehension of dyslexic and non-dyslexic English language learners. Results indicated that the use of this accommodation may increase the comprehension scores of young dyslexic L2 learners when reading difficult texts. The use of the read-aloud accommodation might assist dyslexic students to demonstrate their text comprehension abilities and could be regarded as a helpful tool, especially if the text is difficult. Furthermore, the use of read-aloud accommodation by students with reading difficulties might be particularly useful because it improves student performance and participation and can potentially increase engagement with L2 texts and L2 learning motivation. The authors also indicated that offering read-aloud accommodation to dyslexic language learners might also be helpful as it can offset some of the negative impacts of their disability. However, the validity of scores produced when this accommodation is provided is a concern. Abedi and Ewers (2013) discussed several studies that challenge the validity of this accommodation. Therefore, Košak-Babuder *et al*. (2019) stressed that if the assessment has high stakes, students' score reports would need to include a disclaimer about the application of read-aloud assistance.

For others, however, there is not much evidence to judge the validity of assessments under those accommodations. Such lack of evidence puts states in a difficult position since they may have to use accommodations without having evidence of their validity and comparability.

Research design for examining the validity of accommodations for ELLs

Both the effectiveness and the validity of accommodations depend, to a large extent, on the content being measured. Accommodations that may be effective and valid for one content area may not be so for other content areas. Therefore, it is imperative to examine the validity of accommodations for each particular content and testing situation. To provide more valid and consistent results on the validity of accommodations, an experimental study can be conducted in which large samples of students are randomly assigned to the accommodated and non-accommodated conditions, and sources of extraneous variables that affect the validity of the outcomes are controlled to the extent possible. Then the validity of the accommodations can be examined using two different yet complementary approaches: (1) performance difference and (2) the criterion-related validity approach.

Performance difference

If accommodations do not alter the focal construct, then performance of non-ELL students – for whom accommodations are *not* intended – should remain the same under both the accommodated and non-accommodated assessments. Thus, a significant performance change under accommodated assessment for the non-ELL students may suggest that accommodations are doing more than they are supposed to do. The null hypothesis on the validity of the accommodated assessments can be examined separately for each accommodation. To test this hypothesis, the mean scores of non-ELL students are compared across the accommodation conditions (accommodated/non-accommodated). If the null hypothesis of no performance difference between the accommodated and non-accommodated assessments is rejected, then one can conclude that the accommodations are affecting the focal constructs and that the validity of those accommodations may have been compromised.

Criterion-related validation approach in examining validity of accommodations

Some researchers argue that increased performance of non-ELL students under accommodations may not be enough reason to suggest that the accommodated assessment outcome is invalid, especially when the ELL group benefits more from the accommodations (the interaction effect). Therefore, examining the validity of accommodated assessments using the criterion-related validity approach may provide more convincing evidence of the validity of accommodated assessments. It would be helpful to apply both approaches in the study of the validity of accommodations.

To assess the validity of accommodation through an external criterion approach, additional information from the participating students would be needed. Students' scores on external tests, as well as some relevant background variables, may be used as the criterion variables. These variables can then be correlated with the accommodated assessment scores using a latent-variable modeling approach. Using a multiple-group confirmatory factor analysis approach, the structural relationships across the accommodated and non-accommodated assessments can be compared (Abedi, 2002). Hypotheses of differences between the two groups (non-ELL accommodated and non-ELL non-accommodated students) can be

examined by testing the invariance of the relationships between the accommodated assessment outcomes and the external criteria.

The following hypotheses of invariance can then be tested:

1 Factor loading of item parcels with the latent variables for a math test would be the same across the accommodated and non-accommodated assessments.
2 Correlations between the math test and the state assessment latent variable (math) would be the same across the accommodated and non-accommodated assessments.
3 The fit indices would be similar between the treatment and the control groups.

If these hypotheses of invariance are not rejected, then the comparability of accommodated and non-accommodated assessment outcomes can be established.

In summary, providing accommodations for ELL students not only helps make assessments more accessible for these students but also provides more valid assessments in general. However, in selecting accommodations for ELL students, several major criteria must guide us. The purpose of accommodations is not to give an unfair advantage to the recipients but to provide equal assessment opportunities to everyone; therefore, accommodations must be able to play this important role. In selecting accommodations, one must clearly consider the purpose and use those accommodations that are relevant for such purpose.

Results of research on the assessment and accommodations of ELL students have indicated that some accommodations are not effective in making assessments more accessible to ELL students while others have been shown to have a more positive impact on the assessment outcomes for ELL students. Among them are linguistic modification of assessments, English dictionary/glossary and bilingual dictionary/glossary, and testing in the student's native language. In the next section of this chapter, we will present a summary of research on some of the most commonly used accommodations in the nation and will then briefly present methodological issues for examining major characteristics of these accommodations based on the two research questions presented earlier.

Current contributions and research

As indicated earlier, accommodations are meant to assist ELL students in the area in which they need the most help – i.e., language – in order to "level the playing field" for these students. Accommodations are strategies intended to reduce threats to the validity of test scores. Students' performance on content-based assessments, such as mathematics and science, can be confounded by language, which is considered to be irrelevant to the construct being measured. In other words, the test may unintentionally measure students' language abilities, rather than their content knowledge in areas such as math or science. Accommodations can help ELL students demonstrate their content knowledge by reducing the confounding of language with the content knowledge. On the other hand, accommodations are not intended to give ELL students an unfair advantage over students not receiving accommodations. Therefore, research must be conducted to determine not only if an accommodation helps ELL students score better on an assessment than without it but also that the accommodation *does not* alter the focal construct. The following is a brief presentation of research findings for the most commonly used accommodations for ELLs. Any of these accommodations that improve performance of ELL students without

providing unfair advantage to the recipients can be used to develop universally designed assessments, which can be used for all students.

- Extended time
- English dictionary
- English glossary
- Bilingual dictionary or glossary
- Customized dictionary
- Native language testing
- Linguistically modified test

Extended time

This accommodation provides students with additional (or extended) time to complete an assessment. This is one of the most commonly used accommodations for ELL students and students in different categories of disabilities (Rivera *et al*., 2000), possibly due to its ease of administration relative to other accommodations. Some studies found extended time to be an effective accommodation for ELLs and students with disabilities (Abedi *et al*., 2000; Hafner, 2001; Thurlow, 2001; Chiu and Pearson, 1999).

However, other studies did not show extended time to be effective for ELLs and students with disabilities (Munger and Loyd, 1991; Fuchs *et al*., 2000; Marquart, 2000). A major issue is that extended time may affect the performance of non-ELL students positively as well, therefore making the validity of an assessment using this accommodation suspect (Abedi *et al*., 2000, 2004; Wolf *et al*., 2008).

Recommendation: This accommodation does not alter the focal construct; therefore, all students can benefit from this accommodation if they need additional time. Thus, extra time can be considered as an accessibility/universal design feature.

English dictionary

Providing a published English dictionary for students to use during a test is another commonly used accommodation for ELL students (Abedi *et al*., 2000; Sireci *et al*., 2003; Willner *et al*., 2008). However, a dictionary must be provided in combination with extended time in order to avoid the problem of information overload. Some studies have found that the dictionary and extended time accommodation affects the performance of all students (Abedi *et al*., 2000; Hafner, 2001; Thurlow, 2001; Maihoff, 2002; Thurlow and Liu, 2001; Wolf *et al*., 2008), casting doubt on the validity of the assessment outcomes under this accommodation.

Furthermore, by having access to the definitions of content-related terms, recipients of a dictionary may be advantaged over those who do not have access to dictionaries, which may compromise the validity of assessment. Consequently, the results of accommodated and non-accommodated assessment may not be aggregated under such conditions. Providing a uniform dictionary is also difficult and burdensome for test administrators, which makes it less feasible to implement (Abedi *et al*., 2001).

Recommendation: Since this accommodation has the potential of altering the focal construct, it cannot be included in the list of accessibility features. It is also important to

note that the assessment outcomes for this accommodation cannot be aggregated with the results for all students.

English glossary

An English-language glossary can be provided as an alternative to a commercial dictionary. In this case, only terminology that appears in the test appears in the glossary, making a glossary less burdensome than providing a dictionary. In some studies, the English glossary with extended time raised the performance of both ELL and non-ELL students. For example, the results of a study by Abedi *et al*. (1998, 2000) showed ELL students' performances increased by 13 percent when they were given a glossary with extended time. While this seemed promising, the non-ELL students' performances increased by 16 percent (Abedi *et al*., 2004), indicating a change in construct from the non-accommodated test version. That is, rather than decreasing the performance gap between ELLs and non-ELLs, this accommodation actually increased the performance gap between the two groups. Thus, the outcomes of assessments under this accommodation may not be aggregated with the non-accommodated outcomes.

Recommendation: Since both ELL and non-ELL students can benefit from this accommodation, it may be included in the list of accessibility features, given the content-related terminologies are not included in the glossary.

Bilingual dictionary/glossary

Bilingual dictionaries and glossaries are also among widely used accommodations for ELL students. As with an English dictionary, by having access to definitions of content-related terms, recipients of a published bilingual dictionary may be advantaged over those who do not have access to one (Abedi *et al*., 2000, 2003b). However, a benefit to providing a bilingual glossary is that only students who want to utilize the accommodation can use the glossary feature. If the student does not feel like he or she needs the accommodation, then the student can simply answer the question based on the original English version of the test (Solano-Flores, 2012). A major limitation with a bilingual dictionary is the content equity issue. Different published bilingual dictionaries present a substantial range of content coverage. Students may also be unaccustomed to using these accommodations in the classroom, making them unfamiliar and therefore not useful. As later explained in the native language testing accommodation, there are various aspects of translating an assessment that make this area of accommodations complicated to implement in classroom assessments. Furthermore, students who speak another language at home may not be literate or fully literate in their home language and would not find a bilingual dictionary helpful.

Recommendation: This accommodation has the potential of altering the focal construct by providing content-based information. If the content-based terms are excluded and only the language unrelated to the focal construct is included in the glossary, then this accommodation has the potential to improve the accessibility of the assessment.

Customized dictionary

In one study, a customized English dictionary was introduced as a more valid alternative to English/bilingual dictionaries (Abedi *et al*., 2001). The customized dictionary was a literal

cut-and-paste of the actual dictionary entries, but only included terms that were (1) in the test and (2) non-content related. Results of studies suggested that this is a highly effective and valid accommodation for ELL students.

Recommendation: This accommodation does not alter the focal construct since it does not provide content-based information. Therefore, this accommodation has the potential to improve the accessibility of the assessment for all students and can safely be put on the list of accessibility/universally designed features.

Native language testing

Translating tests into a student's native language is an accommodation used by educational institutions and examination agencies (Abedi *et al*., 2001; Rivera *et al*., 2000). However, translating a test can make the instrument easier or harder in another language, and some cultural phrases and idioms can be difficult to translate (Hambleton, 2001). Translating assessment tools into a student's native language may not produce desirable results if the language of instruction and assessment are not aligned (Abedi *et al*., 2000). As discussed by Solano-Flores (2012), a limitation of having an assessment translated into a student's native language is that there is the assumption that the student is proficient in their native language. Furthermore, students may only be proficient in *speaking* their home language, not in reading it, particularly if the students have not received formal instruction in the home language. Given the large number of languages spoken in public schools and the multiple steps in translating tests, this accommodation raises many feasibility concerns due to the resources needed (Solano-Flores, 2012).

Recommendation: This accommodation is likely to alter the focal construct by issues in the translation process and language equivalence. Another major issue with this accommodation is the language of instruction. If the language of instruction is not consistent with the language of assessment, results cannot be properly discussed. Because of all these problems, including this accommodation as an accessible feature is not recommended.

Linguistically modified test

Some accommodations help ELL students with their English language needs without compromising the validity of the assessment. Studies suggest that a linguistically modified version of the test items is an effective and valid accommodation for ELL students (Abedi *et al*., 2000; Maihoff, 2002; Rivera and Stansfield, 2001; Sato *et al*., 2010). We will elaborate more on this form of accommodation than the other accommodations for two reasons. First, as indicated earlier, a linguistically modified version of a test is a good example of an effective, valid, and relevant accommodation for ELL students. Second, the impact of linguistic factors on the assessment of ELL students is clearly being demonstrated by this accommodation.

Linguistic modification involves modifying the linguistic features on existing assessments of any item that may include unnecessary linguistic complexities. For example, something as simple as switching from passive voice to active voice is considered a linguistic modification. Some research refers to this as linguistic simplification, but it is not

necessarily simplifying a test, but rather reducing linguistic complexity that is not related to the focal construct.

To provide more contextual framework for this accommodation, we start with the concept of language as a source of construct-irrelevant variance for content-based assessment in which language is not the primary construct being measured. Research has identified 14 major categories of linguistic features that appear to contribute to the difficulty of comprehending text (Abedi, 2006, 2007, 2010; Abedi *et al*., 1997; Shaftel *et al*., 2006). These features may slow the reader down, increase the likelihood of misinterpretation, or add to the reader's cognitive load, thus interfering with concurrent tasks. Indexes used to measure language difficulty include unfamiliar (or less commonly used) vocabulary, complex grammatical structures, and discourse that include extra material, abstractions, and passive voice (Abedi *et al*., 1997).

In the linguistic modification approach, sources of unnecessary linguistic complexity, such as those mentioned here, are identified. A team of content, linguistic, and measurement experts determine which linguistic features, among the 14 major categories that were identified in research (Abedi, 2010), are present in the text and how to reduce their impacts. First, each item in the test is examined for the presence and intensity of each linguistic feature by rating these features in the item. A five-point Likert-type rating rubric (with 1 indicating little or no complexity of the item on that feature and 5 suggesting that the item was seriously impacted by the feature) is used to generate ratings on all items. Items that were rated high (rating of 3 or higher) can then be marked for either deletion or revision. The same group of experts who identified these features would then guide the revision process, changing or eliminating items when necessary. The goal would be for all items to have a rating of 2 or below on the linguistic features of the test.

This accommodation may have great potential to be effective and help provide valid assessment outcomes for ELL students. It appears to be quite relevant for ELL students, since it provides the type of assistance (linguistic assistance) that ELL students need, is based on consistent results from experimentally controlled experiments, and does not alter the construct being measured (Abedi, 2010).

Recommendation: This accommodation does not alter the focal construct. Therefore, this accommodation has the potential to improve the accessibility of the assessment for all students and can safely be put on the list of accessibility/universally designed features.

Computer testing

Research findings suggest that computer testing is an effective and valid accommodation for ELL students. In a study by Abedi (Abedi, 2009; Abedi *et al*., 2003a), ELL students showed higher levels of motivation on the assessment when it was administered by computer. The study used a computer to implement a customized English "pop-up" glossary, which was also found to be effective and valid for ELL students. The pop-up glossary aspect of the system was set up to include terms that were unrelated to the content being measured (mathematics in this study). The system kept track of how often the glossary terms were utilized and how much time each student spent on the glossary terms. In the second version of Abedi's study, several additional accommodation features were added to the system. These new features include variation in font size, read-aloud test items and adaptive testing based

on the students' level of English proficiency, which is tested at the start of the session. In a later study (Roohr and Sireci, 2017), researchers found through computerized testing that even though accommodations are provided for students, it does not guarantee that students will utilize them. Specifically, computerized testing allowed the researchers to track how much time students spent answering each item on the test. This additional information is easily attained through computer assessments and may be beneficial in determining which accommodations are most used by students, which then determines whether the accommodation shapes student test scores. It may be even more useful to track how much time was spent answering the question but also how long the accommodation is used (Roohr and Sireci, 2017). For example, with pop-up glossaries or dictionaries, the computer could track how long the student kept the accommodation open while answering the question. It is also important to note that students should be accustomed to using the computerized accommodation. This is important because they should not spend time learning how to use the accommodations instead of focusing on the context of the assessment (Solano-Flores, 2012).

Recommendation: This accommodation provides opportunities to improve assessments in general and those of ELLs in particular. This accommodation facilitates implementation of some of the well-known accommodations for ELLs. For example, the customized dictionary, which is shown to be effective and valid, would be operationally challenging with the paper-and-pencil approach because students have to go to the test appendix section for a glossary and return to the testing session. In computer-based assessment, glossaries are presented in a pop-up format. Therefore, this accommodation has the potential to improve the accessibility of assessment for all students and can safely be put on the list of accessibility/universally designed features. However, there are several issues with this accommodation. First and most importantly, students' computer background has a major impact on the success of this form of testing. Students with higher levels of access to computers are likely to perform better, regardless of the content difficulty of the assessment. Therefore, it is essential to make sure that all students who are tested under this accommodation have the necessary computer skills to be tested on a computer system. Second, due to the many accessibility features that are incorporated into the system, the assessment becomes more complex and time consuming. Students should be given enough time to deal with the issues presented by the computer-based accommodations.

Future developments

One of the most important principles in content-based assessment is validity: the assessment measures what it is intended to measure. Issues concerning the use and validity of accommodations are gaining international momentum. Given the growth in the population of immigrants in the world and considering the education issues of ELLs, accommodations are among the important topics in the international education community. For example, it is likely that the United Kingdom will go down a route akin to that of the United States in coming years, given the growing numbers of ELLs in state schools and, similarly, Germany, with its Turkish and Kurdish communities. Educational researchers and policy makers have given their highest attention to these issues since accommodation issues are of major concern due to the increasing population of immigrant students. There is much to be learned from the research amassed on accommodations in the United States, as well as that from other countries facing the same situation.

Most importantly, if the outcome of the research on accommodations determines that some accommodations can safely be used to improve that assessment quality for everyone, then we can build comprehensive accessibility features that could help us with developing universally designed assessments.

Since accommodations are likely to become a major component of curriculum, instruction, and assessment in the future world's educational system, examining their validity and effectiveness should have a high priority on each nation's educational agenda. It is important to understand that ELLs have the knowledge and ability to learn the content knowledge at the same level as their native English-speaking peers if language disparity can be resolved. The use of some language-related accommodations can "level the playing field": i.e., give ELLs the opportunity to learn and the opportunity to be as fairly assessed as their peers while learning academic content in a language in which they may not be proficient.

The international community can share and benefit from research on the accommodations for ELL students. It is therefore imperative to have an effective communication channel to collaborate with the research in accommodations and share the outcomes of such studies.

Acknowledgements

The author is grateful to Professor Glenn Fulcher for his careful reading of this chapter and for providing very helpful suggestions and advice. I also wish to acknowledge the help of Cecilia Sanchez in reviewing and editing drafts of this chapter.

Further reading

Abedi, J., Zhang, Y., Rowe, S. E. and Lee, H. (2020). Examining effectiveness and validity of accommodations for English language learners in mathematics: An evidence-based computer accommodation decision system. *Educational Measurement* 39: 41–52. https://doi.org/10.1111/emip.12328.

Studies on the accommodations for ELLs have identified at least 73 different types of accommodations that are used by different states, many of which are adopted from the pool of accommodations created and used for students with disabilities and may not be relevant for this group of students. ELL students need assistance in the language of instruction and assessment to assimilate successfully into the mainstream instruction and assessment; therefore, language-based accommodations would be most relevant for these students. On the other hand, not all language-based accommodations can be used for ELLs because some of these accommodations may provide unfair advantage to the recipients and can invalidate the accommodated assessment outcomes. For example, a commercial dictionary is one of the most commonly used language-based accommodations in the nation. While a dictionary can help overcome ELL students' language barriers, it provides content definitions as well, which is unfair to non-recipients of this accommodation.

Based on the review of literature, we identified five different language-based accommodations that do not provide unfair advantage to the recipients and may be effective in making assessments more linguistically accessible to ELLs while at the same time not altering the focal construct (mathematics in this study). We selected mathematics as the content because the focal construct is mathematics, not language.

The accommodations used in this study were (1) a linguistically modified version of the mathematics test, (2) an English glossary of non-content terms, (3) read-aloud test items, (4) a Spanish version of the test, and (5) a Spanish glossary for the test.

To examine the effectiveness of the accommodations used in this study, we compared the performance of ELL students under accommodation with their performance in non-accommodated groups. Results indicated that while students under some accommodation, such as linguistically modified

accommodation, consistently performed better under the accommodated assessments, the difference did not reach a statistically significant level.

To examine validity of the accommodations used, we compared the performance of non-ELL students under the accommodated and non-accommodation conditions. The results of the analyses did not show any significant difference between non-ELLs who were accommodated and those tested under standard testing conditions with the original English version. These results confirm that none of the accommodations used in this study altered the focal construct.

Abedi, J. (2007). Utilizing accommodations in the assessment of English language learners. In N. H. Hornberger (ed.), *Encyclopedia of Language and Education: Vol. 7: Language Testing and Assessment*. Heidelberg, Germany: Springer, 331–48.
Concerns over the validity of accommodations for ELL students have surfaced as the international education community moves toward inclusion of all students (including ELLs and SDs) in national and local assessments. The results of accommodated and non-accommodated assessments cannot be aggregated if the validity of accommodations has not been established. This paper summarizes the results of research on the effectiveness and validity of accommodations for ELLs and examines research findings to determine if there is sufficient evidence to inform decision makers on the provision of accommodations and reporting of accommodated assessments. While the focus of this paper is on accommodation issues for ELL students, some references have been made to accommodations for students with disabilities since ELL accommodation practices are highly affected by accommodation policies and practices for SDs.

Abedi, J., Hofstetter, C. and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based research. *Review of Educational Research* 74: 1–28.
Decisions about which accommodations to use, for whom, and under what conditions are based on limited empirical evidence of their effectiveness and validity. Given the potential consequences of test results, it is important that policymakers and educators understand the empirical base underlying their use. This article reviews test accommodation strategies for ELLs, derived from "scientifically based research." The results caution against a one-size-fits-all approach. The more promising approaches include modified English and customized dictionaries, which can be used for all students, not just ELLs.

Francis, D., Rivera, M., Lesaux, N., Kieffer, M. and Rivera, H. (2006). *Practical Guidelines for the Education of English Language Learners: Research-Based Recommendations for the Use of Accommodations in Large-Scale Assessments*. Portsmouth, NH: RMC Research Corporation, Center on Instruction. www.centeroninstruction.org/files/ELL3-Assessments.pdf.
Language is a key component of individual student success, yet ELLs are often unprepared for the rigors of academic language in content areas. This meta-analysis examines the effectiveness and validity of selected accommodations, familiar to the student through daily instructional use and applied during assessments. The results suggest the importance of linguistically appropriate accommodations, ones that have been used effectively in the classroom, and are appropriately selected to match individual student needs. Although no single accommodation has been shown to level the playing field, the most effective ones may be the use of a dictionary or glossary with extra time, provided that the student has previous classroom experience with dictionary or glossary use. Ensuring the opportunity to learn during instruction combined with accommodations that are familiar and useful to individual students in the classroom can increase the chances of academic success for all students.

References

Abedi, J. (2002). Assessment and accommodations of English language learners: Issues, concerns, and recommendations. *Journal of School Improvement* 3: 83–89.

———. (2006). *Accommodations for English Language Learners That May Alter the Construct Being Measured*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

———. (2007). English language learners with disabilities. In C. Cahlan-Laitusis and L. Cook (eds.), *Accommodating Students with Disabilities on State Assessments: What Works?* Arlington, VA: Council for Exceptional Children.

———. (2009). Computer testing as a form of accommodation for English language learners. *Educational Assessment* 14: 195–211. https://doi.org/10.1080/10627190903448851.

———. (2010). Linguistic factors in the assessment of English language learners. In G. Walford, E. Tucker and M. Viswanathan (eds.), *The Sage Handbook of Measurement*. Oxford, UK: Sage Publications.

Abedi, J. and Ewers, N. (2013). *Smarter Balanced Assessment Consortium: Accommodations for English language learners and students with disabilities: A research-based decision algorithm*. Olympia, WA: Smarter Balanced Assessment Consortium. Retrieved from http://www.smarterbalanced.org/wordpress/wp-content/uploads/2012/08/Accomodations-for-under-represented-students.pdf

Abedi, J. and Herman, J. L. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teachers College Record* 112: 723–46.

Abedi, J., Lord, C. and Plummer, J. (1997). *Language Background as a Variable in NAEP Mathematics Performance*. CSE Technical Report No. 429. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., Lord, C. and Baker, E. (1998). *NAEP Math Performance and Test Accommodations: Interactions with Student Language Background*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., Hofstetter, C. and Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice* 19: 16–26. https://doi.org/10.1111/j.1745-3992.2000.tb00034.x.

Abedi, J., Courtney, M., Mirocha, J., Leon, S. and Goldberg, J. (2001). *Language Accommodation for Large- Scale Assessment in Science*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Courtney, M. and Leon, S. (2003a). *Effectiveness and Validity of Accommodations for English Language Learners in Large-Scale Assessments*. CSE Technical Report No. 608. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Leon, S. and Mirocha, J. (2003b). *Impact of Students' Language Background on Content-Based Assessment: Analyses of Extant Data*. Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C. H. and Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research* 74: 1–28. https://doi.org/10.3102/00346543074001001.

Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J. and Freidebach, M. (2001). *Read-Aloud Accommodation: Effects on Multiple-Choice Reading and Math Items*. Technical Report. Minneapolis, MN: National Center on Educational Outcomes.

Capp, M. J. (2017). The effectiveness of universal design for learning: A metaanalysis of literature between 2013 and 2016. *International Journal of Inclusive Education* 21: 791–807. https://doi.org/10.1080/13603116.2017.1325074.

Chiu, C. W. T. and Pearson, D. (1999). *Synthesizing the Effects of Test Accommodations for Special Education and Limited English Proficient Students*. Paper presented at the National Conference on Large Scale Assessment, Snowbird, UT.

Francis, D., Lesaux, N., Kieffer, M. and Rivera, H. (2006). *Research-Based Recommendations for the Use of Accommodations in Large-Scale Assessments*. Houston, TX: Center on Instruction.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C., Binkley, E. and Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children* 67: 67–81. https://doi.org/10.1177/001440290006700105.

Goegan, L. D., Radil, A. I. and Daniels, L. M. (2018). Accessibility in questionnaire research: Integrating universal design to increase the participation of individuals with learning disabilities. *Learning Disabilities: A Contemporary Journal* 16: 177–90. https://doi.org/10.7939/r3-cmkq-1c82.

Gay, L. R. (1981). *Educational Research: Competencies for Analysis and Application*, 2nd edn. Columbus, OH: Charles E. Merrill.

Hafner, A. L. (2001). *Evaluating the Impact of Test Accommodations on Test Scores of LEP Students & Non-LEP Students*. Los Angeles, CA: California State University.

Hambleton, R. H. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (ed.), *Setting Performance Standards: Concepts, Methods, and Perspectives*. Mahwah, NJ: Lawrence Erlbaum, 89–116.

Islam, C. and Park, M. (2015). Preparing teachers to promote culturally relevant teaching: Helping English language learners in the classroom. *Multicultural Education* 23: 38–44.

Košak-Babuder, M., Kormos, J., Ratajczak, M. and Pižorn, K. (2019). The effect of read-aloud assistance on the text comprehension of dyslexic and non-dyslexic English language learners. *Language Testing* 36: 51–75. https://doi.org/10.1177/0265532218756946.

Maihoff, N. A. (2002, June). *Using Delaware Data in Making Decisions Regarding the Education of LEP Students*. Paper presented at the Council of Chief State School Officers 32nd Annual National Conference on Large-Scale Assessment, Palm Desert, CA.

Marquart, A. (2000). *The Use of Extended Time as an Accommodation on a Standardized Mathematics Test: An Investigation of Effects on Scores and Perceived Consequences for Students of Various Skill Levels*. Paper presented at the annual meeting of the Council of Chief State School Officers, Snowbird, UT.

Meloy, L. L., Deville, C. and Frisbie, D. (2000, April 24–28). *The Effect of a Reading Accommodation on Standardized Test Scores of Learning Disabled and Non Learning Disabled Students*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. Retrieved from http://eric.ed.gov/PDFS/ED441008.pdf.

Munger, G. F. and Loyd, B. H. (1991). Effect of speediness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research* 85: 53–57. https://doi.org/10.1080/00220671.1991.10702812.

Munro, J., Abbott, M. and Rossiter, M. (2013). Mentoring for success: Accommodation strategies for ELLs. *Canadian Journal of Action Research* 14: 22–38. https://doi.org/10.33524/cjar.v14i2.83.

Rivera, C. and Collum, E. (2005). *State Assessment Policies for English Language Learners: A National Perspective*. New York: Routledge.

Rivera, C. and Stansfield, C. W. (2001, April). *The Effects of Linguistic Simplification of Science Test Items on Performance of Limited English Proficient and Monolingual English-Speaking Students*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Rivera, C., Stansfield, C. W., Scialdone, L. and Sharkey, M. (2000, April 12). *An Analysis of State Policies for the Inclusion and Accommodation of English Language Learners in State Assessment Programs During 1998–99*. Final Report. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA. Retrieved from http://deepwater.org/bioteacher/12-Review%20and%20SOLs/SOL-Delaware-simplification.pdf.

Roohr, K. and Sireci, S. (2017). Evaluating computer-based test accommodations for English learners. *Educational Assessment* 22: 35–53. doi:10.1080/10627197.2016.1271704

Roohr, K. C. and Stephan, G. (2017). Evaluating computer-based test accommodations for English learners. *Educational Assessment* 22: 35–53. https://doi.org/10.1080/10627197.2016.1271704.

Sato, E., Rabinowitz, S., Gallagher, C. and Huang, C. W. (2010). *Accommodations for English Language Learner Students: The Effect of Linguistic Modification of Math Test Item Sets*. NCEE 2009-4079. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shaftel, J., Belton-Kocher, E., Glasnapp, D. and Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment* 11: 105–26. https://doi.org/10.1207/s15326977ea1102_2.

Sireci, S. G., Li, S. and Scarpati, S. (2003). *The Effects of Test Accommodation on Test Performance: A Review of the Literature*. Center for Educational Assessment Research Report No. 485. Amherst, MA: School of Education, University of Massachusetts.

Solano-Flores, G. (2012). *Translation Accommodations Framework for Testing English Language Learners in Mathematics*. Submitted to Smarter Balanced Assessment Consortium, University of Colorado, Boulder.

Thurlow, M. L. (2001, April). *The Effects of a Simplified-English Dictionary Accommodation for LEP Students Who Are Not Literate in Their First Language*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Thurlow, M. and Liu, K. (2001). *State and District Assessments as an Avenue to Equity and Excellence for English Language Learners with Disabilities*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

U.S. House of Representatives. (2001). *No Child Left Behind Act of 2001*. 107th Congress 1st Session, Report 107-334. Washington, DC: Author.

Willner, L. S., Rivera, C. and Acosta, B. D. (2008). *Descriptive Study of State Assessment Policies for Accommodating English Language Learners*. Arlington, VA: George Washington University Center for Equity and Excellence in Education.

Wolf, M. K., Herman, J. L., Kim, J., Abedi, J., Leon, S., Griffin, N. . . . Shin, H. W. (2008). *Providing Validity Evidence to Improve the Assessment of English Language Learners*. CRESST Report 738. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from https://cresst.org/publications/cresst-publication-3109/.

Wolf, M. K., Kim, J. and Kao, J. (2012). The effects of glossary and read-aloud accommodations on English language learners' performance on a mathematics assessment. *Applied Measurement in Education* 25: 347–74. https://doi.org/10.1080/08957347.2012.714693.

19

Rater and interlocutor training

Larry Davis

Introduction

In performance assessment, raters perform the key role of transforming evidence of language ability into a score. Raters therefore have considerable impact on the extent to which scores reflect the intended construct, are reliable, and are fair for different types of individuals. Producing scores that are accurate, reliable, and fair is no easy task, and various aids such as scoring rubrics and exemplar performances are typically used to guide the decision-making process. However, such aids can in themselves be challenging to use (Lumley, 2005), and it is generally recognized that some sort of training is necessary for raters to appropriately apply the scoring criteria as identified by the test designers (Fulcher, 2003; Weigle, 2002).

The goal of rater training is to help raters focus their perceptions on relevant aspects of performance and then value those aspects in accordance with the rating scale. The training procedures actually used in a given situation will vary depending on the stakes associated with the test, the complexity of the scoring decisions to be made, and the context of training. Greater effort in training is required when the assessment stakes are high, when applying multiple criteria, or when scoring complex responses. The types and focus of training activities may also vary in different training and assessment settings. For example, initial training of new raters will necessarily involve orienting raters to the assessment content and scoring criteria, whereas ongoing training of experienced raters will likely focus more on providing feedback on scores (e.g., Knoch, 2011). The modality of training, such as face-to-face in a meeting room or self-access in an online environment, will also dictate the feasibility of training activities such as group discussions or the provision of individual feedback to raters. While the goal of training is always to ensure that scores are valid, reliable, and fair, the way in which training is carried out must be adapted to meet the needs of each situation.

While this chapter will focus primarily on scoring, in some speaking tests, the rater also serves as an interlocutor and participates in the creation of the examinee's response. This creates a special challenge: the examiner must elicit the spoken interaction that serves as evidence of ability while also evaluating this evidence. This task requires skill in managing the interaction as well as the cognitive capacity to simultaneously make accurate scoring decisions. It has long been documented that in dialogic assessment tasks, a particular

 DOI: 10.4324/9781003220756-25

interlocutor's contributions may have considerable impact on the response produced by the test taker (Lazaraton, 1996; Ross, 1992), thereby influencing the evidence available upon which to base a score. Despite the potential for interlocutor behavior to influence test results, Brown (2012) observed in the first edition of this volume that interlocutor training has received relatively little attention in the research literature. Unfortunately, this situation still remains the case today. Consequently, this chapter will also consider the specific challenges posed by interlocutor behavior and different approaches to interlocutor training.

Historical perspectives

It has long been known that raters may differ in the scores they award; indeed, the issue of rater variability was the topic of one of the earliest statistical investigations of an assessment (Edgeworth, 1890). Over the years, a succession of studies has demonstrated that variability in scores can be especially pronounced when raters are not given any formal guidance but rather are asked to judge as they see fit. An oft-cited example is the study by Diederich *et al*. (1961), in which 53 individuals from a variety of professional backgrounds each scored a group of 300 essays produced by first-year college students. Judgments were made using a scale of one to nine based on "whatever hunches, intuitions, or preferences you normally use in deciding that one paper is better than another" (p. i). The result was that 94 percent of papers obtained scores from different raters that spread across at least seven of the nine score categories. More recently, O'Hagan and Wigglesworth (2015) reported similar results from a study in which a group of faculty evaluated essays produced in university management courses. While all individuals had taught the subject in question and were experienced in evaluating such essays, there were several instances in which grades for the same essay spread across all six possible levels. An extensive body of research has also shown that raters may differ in the features of performance considered in scoring or the process used to reach a decision (e.g., Brown, 2007; Cumming *et al*., 2002; Milanovic *et al*., 1996; Orr, 2002; Papajohn, 2002).

While it was clear that examiners might vary in their perceptions, in the earliest years, relatively little attention was paid to rater standardization. For example, in the Cambridge Certificate of Proficiency in English (CPE), initiated in 1913, measurement of language ability focused on the construct being measured and less on technical issues of reliability (Weir, 2005). Examiners were given broad latitude to evaluate language performance and, in the case of speaking, to engage the examinee in conversation to elicit a sample of language (Weir, 2003). The validity and reliability of the results was assumed and based ultimately on the taste and discernment of the examiners, who were drawn from elite levels of society (Weir, 2005).

In the United States, efforts to develop language tests began in earnest in the years following World War Two, and in keeping with several decades of practice in educational and psychological testing, much more emphasis was placed on test reliability (Spolsky, 1995). Efforts to standardize scoring were also more in evidence, particularly in the development of language tests by the Foreign Service Institute (FSI) of the US government. A major purpose for creating the influential FSI scale was to promote consistency in grading across examiners and languages (Sollenberger, 1978), and thorough training was seen as a key contributor to consistent results in the FSI oral proficiency interview (Wilds, 1975). New examiners reviewed various materials, including exemplar interviews, as well as commentaries intended to clarify scoring criteria and suggest interviewing techniques. Raters also learned from other more experienced examiners and monitored each other. Wilds describes

a situation that bears strong resemblance to the more recent concept of a "community of practice" (see Lave and Wenger, 1991), where examiners worked together to share knowledge and develop understanding:

> It is, however, very much an in-house system which depends heavily on having all interviewers under one roof, able to consult with each other and share training advances in techniques or solutions to problems of testing as they are developed and subject to periodic monitoring.
>
> *(1975: 35)*

The focus on building a community of examiners became even more pronounced in the 1970s and 1980s, particularly with the development of holistic scoring practices for writing (White, 1984). A major feature of the holistic approach was to bring readers together in the same location, where an "interpretive community" could be established "that determines and enforces the standards of measurement" (White, 1984: 404). Training involved scoring batches of exemplar papers, followed by discussion to reach consensus on scores and to revise the scoring guidelines if necessary. In the words of White (1985), the goal of such training was

> not indoctrination into standards determined by those who know best (as it is too often imagined to be) but rather the formation of an assenting community that feels a sense of ownership of the standards and the process.
>
> *(p. 164)*

Many of the training practices described by White remain in use today. However, until recently, creating a community of practice where practitioners interact to share and improve their craft (Lave and Wenger, 1991) has necessarily entailed raters meeting together. In contrast, over the last 20 years, there has been an increasing trend in large-scale assessments toward distributed scoring, facilitated by new technologies that allow raters to work individually from a variety of locations. This separation makes it very challenging to establish an interactive community of the type envisioned by White (1985), and while online communication tools now exist to enable virtual meetings, such tools did not exist in the early 2000s when distributed scoring began to proliferate. In any case, the focus of training in online contexts has moved from creation of a shared consensus to mastery of pre-defined scoring criteria, for which much of the training is done on a self-access basis (e.g., Wolfe *et al*., 2010).

Critical issues and topics

Rater expertise as an organizing framework

Consideration of rater training ultimately leads to the issue of rater expertise; that is, any discussion of training requires consideration of the expertise that the training is intended to produce. In language testing research, rater expertise has often been defined in terms of language teaching background and experience in scoring (e.g., Barkaoui, 2010; Weigle, 1994). The use of background variables as an indicator of expertise is a relatively simple way to operationalize rater quality, but expertise is more properly defined in terms of ability and performance rather than background (Ericsson, 2006). Accordingly, Davis (2012) defined

rater expertise as "the knowledge, skills, and behaviors that enable a rater to produce consistent, valid, and accurate scores within a particular scoring context" (p. 9).

However, what are the knowledge, skills, and behaviors that characterize an "expert" rater? As a starting point, the broad psychological literature on expertise provides useful insights into the general nature of this construct (e.g., Feltovich *et al*., 2006; Mislevy, 2010; Shanteau, 1992). Some key points from this literature are:

1 Expertise involves the ability to focus on relevant information, ignore unimportant information, and notice details that might be missed by others.
2 Expertise involves the ability to recognize patterns and organize information in chunks. This allows more information to be held in memory and new information to be quickly organized.
3 Expertise involves automation of basic tasks, gained through extensive practice. Automated basic cognition then supports higher-level reasoning and integration of information.
4 Reflection is important for expertise; experts monitor themselves to avoid mistakes and to adjust to unusual situations.
5 The highest levels of expertise cannot be achieved by experience alone but also require focused practice, feedback, and reflection.
6 Expertise does not exist solely within the expert's head but is socially mediated. Cognition may be distributed and require interactions with others and is also supported by skilled use of artifacts and knowledge representations such as scoring rubrics (Mislevy, 2010).
7 Expertise is context specific, and excellent performance in one context does not necessarily transfer to a different context.

These findings are helpful for interpreting language assessment research as well as for guiding training practices. For example, in the context of writing assessment, Zhang (2016) reported that more accurate raters were more consistent in classifying the gravity of various errors, better at abstracting and integrating various features of each essay, and frequently monitored their scoring process to ensure a well-reasoned decision. These findings exemplify characteristics of expertise mentioned in items one through four on this list. Similarly, Wolfe (2006) observed that highly proficient raters of essays tended to read the entire essay before considering a score, while less proficient raters were much more likely to reach a scoring decision earlier. Wolfe suggested that early decision making reduces cognitive load for less proficient individuals by breaking the decision into smaller parts, while experts were more able to integrate the full qualities of the essay into a holistic judgment. While the language assessment literature has so far made little reference to studies of general expertise, these examples give a taste of how the behavior of raters might be more fully understood as a specific example of expert performance.

Validity frameworks and rater training

Rater expertise and training must also be understood in terms of their impact on the validity framework supporting a particular assessment use (see Chapelle and Lee, this volume). The clearest description of such impacts is provided in Knoch and Chapelle (2018), who explored how the rating process can be understood within an argument-based approach to validation. They systematically identified issues related to rater performance in all six major

inferences that commonly form the steps in a validity argument. For example, the first step (the evaluation inference) entails an assertion that raters are consistent in scoring responses. The assumption of inter-rater consistency supports the next step (the generalization inference), while the claim that rater cognition reflects the intended construct is part of the third step (the explanation inference). The fourth step (the extrapolation inference) rests partly on the assertion that raters employ scoring criteria in a way similar to individuals in the target language use domain. The fifth step (the decision inference) requires an assumption that the scale adequately differentiates test takers, which in turn relies on raters' ability to distinguish score categories. For the sixth and final step (the consequence inference), it is expected that raters themselves should benefit from the experience of scoring, particularly less experienced individuals who may improve their understanding of the construct.

Knoch and Chapelle argue that careful specification of a validity argument clarifies the type of validation research needed to support test uses and provides a common framework for understanding such research. In describing recent examples of studies referencing a validity argument, Knoch and Chapelle write:

> Despite the variety of tests and score uses represented by these examples, they share a clarity of research approach because each of the studies draws upon common elements of argument-based validity research. Such clarity allows researchers to communicate both the motivation for their research and the meaning of the results with reference to the validity of their test score interpretations, uses, and consequences.
>
> *(p. 2)*

Beyond facilitating research, validity frameworks also serve as a basis for guiding rater and interlocutor training by clarifying how different aspects of rater performance may influence the validity of an assessment use, which in turn helps identify the aspects of rater expertise that will be most important to address in training.

Rater variability as a strength

In the previous edition of this book, Brown (2012) made the point that while rater variability is typically seen as problem to be minimized through training and other means, there is another tradition in which rater variability is seen as a strength rather than a weakness. That is, the diverse perceptions of individual experts have the potential to combine into an evaluation that is more robust and nuanced than the result from a single individual. Moss (1994) termed this the "hermeneutic approach," in which holistic interpretations, based on multiple pieces of evidence, are made by a community of evaluators who are familiar with the assessment context. Under a hermeneutic approach, the richness of interpretation is valued. Common real-life examples include job interviews and dissertation defenses, in which a local decision is made by a committee of people specifically selected to represent a range of viewpoints. In contrast, the psychometric approach commonly applied in language testing is based on a view of raters as more or less interchangeable, and the emphasis is on statistical reliability achieved through repeated observations. As Brown (2012) noted, the hermeneutic approach tends to be associated with local assessments led by teachers and focused on classrooms, while the psychometric approach is dominant in the context of large-scale, high-stakes testing.

While rater training as carried out in the psychometric tradition has the goal of directing raters' perceptions to approximate a set standard, a hermeneutic view might suggest a

different focus. The procedures for holistic scoring described earlier (White, 1984, 1985) illustrate one example of a hermeneutic approach to rater training. Hamp-Lyons (2007) provides another example, wherein a part of the training was actually directed toward helping raters more efficiently interact in reaching a consensus. Raters were provided with guidelines describing the types of interactions to be encouraged during rater standardization sessions and were also encouraged to use "positive sharing," in which each reader shares positive qualities of the essay, and

> all readers accept that each reading of the essay is inherently "true," that each reader is in turn an expert, and that therefore a fair scoring of the essay must lie somewhere in the interaction of these truths.
>
> *(p. 6)*

While this chapter will focus on research and methods associated with a psychometric approach to scoring, it is worth remembering that other approaches exist and may be equally or more valid in certain contexts.

Yet even within a psychometric approach, an excessive focus on high levels of rater agreement may actually lead to a variety of problems (Eckes, 2011). In the extreme case, raters who literally produce identical scores violate the assumption of local independence of raters. Such duplicate ratings reduce the information that would be collected from truly independent raters, resulting in inflated estimates of test reliability, unless lack of independence is specifically accounted for when modeling rater behavior (Bock *et al*., 2002). More realistically, an over-emphasis on agreement may push raters to "play it safe" by selecting scores near the center of the scale or to base their decisions on surface-level phenomena that are easy to judge, resulting in construct underrepresentation. While raters "should be able to provide ratings based on the same point of view in terms of their understanding of the construct" (Eckes, 2011: 69), they are assumed to be independent experts who may vary somewhat in their opinions. Differences in severity, if consistent, can potentially be adjusted through means such as multi-facet Rasch measurement (MFRM).

Current contributions and research

Given the importance of the rater's role in performance assessment (see Galaczi and Lim, this volume), investigation of rater training and expertise has long been an active area of research. A consistent finding is that rater training is usually effective in standardizing raters; specifically, training of novice raters tends to reduce extremes in severity and increase agreement with reference scores (Davis, 2016; Furneaux and Rignall, 2007; Shaw, 2002; Weigle, 1998). Raters showing bias toward specific scoring categories may also improve following training (Fahim and Bijani, 2011). Additionally, studies employing verbal report methodology have found that training helped raters more accurately interpret the scoring criteria (Kim, 2015; Weigle, 1994) and balance their attention across different aspects of performance (Davis, 2019; Furneaux and Rignall, 2007).

However, a nearly ubiquitous finding is that raters continue to differ in severity regardless of training. Studies using MFRM commonly report a range of severity of 1 to 2 logits from the mean, even among pools of well-trained raters (e.g., Eckes, 2005; Myford and Wolfe, 2000; Nakatsuhara *et al*., 2020). As mentioned earlier, some degree of difference is expected if raters are considered "independent experts," and variation in scores will always be present due to random influences such as scoring conditions or rater fatigue (Eckes, 2011). Also,

training does not always eliminate systematic differences in scoring associated with rater background, although findings are mixed. Several studies of speaking assessments found no differences in scores from trained raters from different first language backgrounds (Brooks, 2013; Wei and Llosa, 2015; Xi and Mollaun, 2009) while other studies have reported that raters tended to give higher scores when they were familiar with the candidate's accent or shared the same first language (Carey *et al*., 2011; Winke *et al*., 2013). Trained raters who are native or non-native speakers of the language being assessed typically produce similar scores but may differ in the aspects of performance most frequently mentioned while scoring (Kim, 2009; Kuiken and Vedder, 2014; Zhang and Elder, 2014), and differences in severity have also been documented (Kim and Di Gennaro, 2012). A challenge in making sense of these findings is that the training regimens used may differ in content or length, and the impact of such variables remains poorly understood. However, these studies highlight the need to monitor rater performance; it cannot simply be assumed that training "worked."

Beyond initial training of new raters, the effectiveness of ongoing training and feedback has also been a focus of research. In an early study, Wigglesworth (1993) provided experienced speaking raters graphical feedback based on MFRM, along with a face-to-face retraining session and practice scoring; after training, several raters showed reduced bias toward analytic subscales. Shaw (2002) also saw progressive reductions in the range of severity and the frequency of bias among raters who completed repeated training sessions over a period of two months. Conversely, other studies in which individualized feedback was provided have observed minimal or inconsistent impacts on severity or bias (Elder *et al*., 2007; Knoch, 2011; O'Sullivan and Rignall, 2007). These investigations generally included representative samples of raters, and an important unanswered question is the impact of feedback or retraining of raters whose performance fails to meet acceptable standards. This situation would arguably be the most important use case for feedback, and limited circumstantial evidence suggests that training may have the greatest effect on individuals who are most different from their peers (e.g., Fahim and Bijani, 2011; Weigle, 1998).

The impact of the delivery format of training is also an area of interest, with comparison of online versus face-to-face formats being a particular focus, given the increasing use of online training in large testing programs. Generally, studies have found that comparable training delivered by either method produced similar results. For example, Wolfe *et al*. (2010) compared training delivered face to face in a group setting, on a self-access basis in a computer lab, and online with participants working from their own locations. No differences were seen in trainee success rates in qualifying to score operationally, in the quality of ratings, or in rater perceptions of training, although individuals completed the self-access training more quickly. Similarly, little difference in rater performance was observed when comparable training or feedback was delivered in person or online (Knoch *et al.,* 2018; Knoch *et al*., 2007). Overall, there appears to be no general reason why training in online and/or self-access environments cannot be as successful as face-to-face training, and it seems likely that the content of the training influences rater performance more than the delivery method.

Finally, beyond the impact of training, the influence of other contributors to scoring expertise has been an area of active research. A full exploration is beyond the scope of this chapter, but some additional themes and typical findings are listed next.

- **The linguistic sophistication of raters.** Raters with linguistic training are more likely to reference technical terms or be more analytical in describing performance phenomena, although this does not necessarily impact scores compared to linguistically naïve

raters. Recent studies focus primarily on perceptions of oral comprehensibility and include Hsieh (2011), Isaacs and Thomson (2013), and Saito *et al*. (2017).

- **The professional background of the rater.** Language professionals tend to focus their attention on language use while subject experts attend more to task completion. Studies on this topic come primarily from speaking assessment and include Douglas and Myers (2000), Duijm *et al*. (2018), and Knoch (2009).
- **The level of experience in scoring the assessment.** Increased experience in scoring a given assessment tends to be associated with better scoring performance. Recent studies in writing assessment include Choi and Wolfe (2020), Lim (2011), Şahan and Razı (2020) and, for speaking, Davis (2016) and Kim (2015).

Main research methods

Investigations of rater expertise have broadly focused on (1) how different variables influence the scores raters produce and (2) how raters arrive at their scoring decisions. As might be expected, quantitative analyses are typically used to investigate patterns in scores while qualitative methods have been used to understand the cognition underlying scoring decisions.

Typical quantitative techniques used to study scoring performance include analyses based on correlation/agreement, many-facet Rasch measurement (MFRM), and generalizability theory (G-Theory). Rater reliability has traditionally been evaluated in terms of correlations between or within raters and agreement in scores, either between raters or in terms of agreement with a previously confirmed score. More sophisticated techniques such as G-Theory and MFRM make it possible to model rater performance alongside multiple other sources of variability present in the assessment context. Analyses using G-Theory employ techniques based on analysis of variance to evaluate the impact of rater variability on final scores (see Fan and Yan, this volume), whereas MFRM is an approach developed from item response theory to model the severity or difficulty associated with different variables (see Ockey, this volume). It has been argued that the two approaches provide complementary information, with generalizability analyses being useful for high-level investigation of how raters, along with other test variables, contribute to score variability and MFRM analyses being useful for understanding the scoring patterns of individual raters (Bachman *et al*., 1995; Lynch and McNamara, 1998).

Qualitative techniques used to understand rater perceptions and decision making typically involve some form of self-report. Written comments made by raters during or after scoring have been commonly gathered (e.g., Kim, 2009; Zhang and Elder, 2011); such reports are relatively easy to integrate into existing scoring procedures but are limited in the degree to which they can capture cognitive processes. Verbal protocol analysis (Ericsson and Simon, 1993) arguably provides more direct access to the actual decision process. In writing assessment, raters may be asked to think out loud during the scoring process (e.g., Barkaoui, 2010) or retrospectively (e.g., Weigle, 1994). In speaking assessment, the difficulty of verbalizing while simultaneously listening to a response typically dictates the use of retrospective techniques such as stimulated recall (Mackey and Gass, 2016). While verbal report remains the standard for eliciting evidence of cognition, the resulting protocols are typically time consuming to analyze. Survey techniques may be more efficient where specific information is of interest and have been used to understand raters' decision-making style (Baker, 2012) or the emphasis raters place on different scoring criteria (Eckes, 2008, 2012). Scoring criteria have also been evaluated by asking raters to score hypothetical test takers with different

performance profiles (Cai, 2015; Randall and Engelhard, 2009). A weakness of these types of direct questioning is that they collect raters' perceptions of what is valued, which may not always be an accurate or full reflection of actual decision making.

Ultimately, understanding the links between cognition and scores often dictates a mixed-methods approach, with qualitative methods used to study psychological factors such as a rater's scoring criteria or decision process, and quantitative methods used to study patterns in scoring (e.g., Weigle, 1994). Recently, studies have been framed within specific mixed-methods research designs (e.g., Baker, 2012; Kim, 2009), as frameworks for conducting mixed-methods studies have become popularized in applied linguistics research.

Recommendations for practice

It is accepted professional practice that those who score examinations should receive appropriate training. As mentioned earlier, training and monitoring activities used in specific situations will vary; descriptions of training procedures have been reported for both small-scale assessments (e.g., Elder *et al.,* 2007; Lim, 2011) and larger-scale tests (e.g., Furneaux and Rignall, 2007; Taylor and Galaczi, 2011; Xi and Mollaun, 2009; Everson and Hines, 2010).

The process starts with rater selection. While the evidence shows that native-speaker status is not a requirement for proficient scoring, some sort of relevant background is typically considered desirable for prospective raters. Besides proficiency in the language being assessed, major testing organizations typically require rater candidates to have earned a bachelor's degree, with teaching experience or teaching certification being desirable or required. During initial training, scoring procedures are described, and raters are introduced to the language tasks to be scored, the scoring rubric, and the type of language expected to be elicited by each task. This is followed by review of example responses that have been selected to represent the different levels of the scale, along with commentaries that explain the rationale for the score awarded to each response. Additional responses may also be included to further illustrate the variety of performance typical for a given score band or to illustrate borderline cases. This familiarization process is followed by practice scoring, in which trainees attempt to score responses and then receive feedback. Practice ratings may be discussed if training is conducted in a group and/or led by a trainer (e.g., Furneaux and Rignall, 2007). If the training is self-access, as with many online scoring programs, then a predetermined score for the response can be provided along with a scoring rationale (e.g., Everson and Hines, 2010). Initial scoring practice may be followed by additional practice and review of scoring materials until the individual feels ready to attempt certification. In certification, rater candidates score a set of pre-scored responses, and individuals who achieve acceptable levels of accuracy then become eligible to participate in operational scoring.

It should be noted that rater training in itself does not ensure the quality of scores. Beyond training, robust quality control entails ongoing monitoring and support of raters. Such monitoring may range from something as simple as raters discussing their scores in cases of disagreement, an approach that might suffice for local and low-stakes assessments, to complex systems with multiple forms of monitoring, feedback, and retraining, which are often necessary in large-scale testing programs. As an example of the latter, testing programs at Educational Testing Service typically require raters to complete a calibration exercise at the start of some or all scoring sessions, in which a sample of responses must be scored with acceptable accuracy before the rater is allowed to begin work. During scoring, small groups of raters are monitored in real time by a supervisor who rescores a sample of responses and

can answer raters' questions or give informal feedback if any problems are observed. Rater performance is also monitored using pre-marked responses randomly distributed through scoring sessions, which provide data for post-hoc analysis of trends in rater performance. Raters who consistently show undesirable scoring patterns can then be directed to additional training. In other organizations, many-facet Rasch measurement has been used for such post-hoc analyses and to generate detailed individualized feedback on scoring performance, such as bias toward particular scoring criteria (e.g., Knoch, 2011; O'Sullivan and Rignall, 2007). Although a full discussion of quality control in human scoring is outside the scope of this chapter, it is important to recognize that rater training is only one piece in a larger effort (see Saville and McElwee, this volume).

In speaking assessments for which the examiner serves as both interlocutor and scorer, training must necessarily be expanded to cover elicitation techniques and management of the testing session. It has been well documented that the interlocutor may influence candidate's performance (Brown, 2005; Lazaraton, 1996), so care is taken to ensure that raters learn how to consistently interact with test takers in the manner specified for the test. Examiner training in the Cambridge ESOL exams, as described by Taylor and Galaczsi (2011), happens in face-to-face sessions that start with familiarizing trainees with the format of test, including the focus of each task, the type of prompt, task timing, and the interaction patterns expected in the resulting discourse. Examiners are also trained in how to conduct the test, including proper handling of test materials and strategies for managing and supporting candidate's interactions as needed. Taylor and Galaczi report that two of the most common challenges are maintaining the timing for the test and making appropriate use of the interlocutor frame (or script), used to ensure consistency in the interactions between examiners and test takers. Raters must follow the script while still speaking naturally and "preserving the human element, which is, after all, one of the distinguishing features of direct tests of speaking" (p. 219). Trainees practice administering the test with each other, followed by practice with volunteer test candidates, during which trainers observe and give feedback. Separate certifications are held for test administration and scoring. Once certified, examiners are observed at least once every two years; several tests are viewed and feedback given.

The ACTFL OPI represents a somewhat different training challenge, in that examiners have much greater latitude in directing the course of the interaction. In this case, considerable attention must be paid to elicitation techniques, and training is an extended process that can potentially take up to a year (ACTFL, 2020). Training starts with a face-to-face workshop of two or four days, which introduces the ACTFL scale, the structure of the interview, and techniques for interviewing and scoring. Following the workshop, optional rating practice is available online. An extended certification process follows, starting with demonstrating an adequate level of language proficiency through an OPI interview and showing the ability to score accurately through scoring recorded OPIs using an online system. This is followed by two "guided interviews" (p. 10), in which the novice examiner interviews two volunteer test takers at different levels, and a trainer observes in real time and gives feedback. A set of four "independent interviews" comes next, which are recorded and submitted for review. In the final certification round, the examiner candidate is asked to submit up to four additional interviews, as needed, to confirm that the candidate's interviewing technique meets the requirements at each level of language proficiency that the examiner will be certified to assess.

It is interesting that in both examples, certification of interviewing skills and scoring accuracy appear to be separated to some extent; that is, examiner candidates are not required to interview and report a score on the spot. This approach simplifies the logistics

of certification but leaves a degree of uncertainty regarding the examiner's ability to handle the cognitive load required to simultaneously interact with the test taker and make scoring decisions.

Future directions

Issues related to rater training and expertise continue to be active areas of research. However, as Brown (2012) noted in the previous edition of this volume, the relative effectiveness of different training methods remains unclear. More broadly, much remains to be done to understand how different factors interact in the development of rater expertise, including background characteristics, rater training, experience in scoring, and the nature of the decision-making task. For example, it has been speculated that exemplars and scoring practice may play different roles in developing the skill of deciding the "what" and "how much" of a scoring decision; that is, exemplars may direct rater perceptions to specific performance phenomena while practice scoring aligns quantitative judgments to the rating scale (Davis, 2019). A major challenge in teasing out such effects is that much of the research has been conducted within operational testing contexts, and controlled studies that rigorously test interactions between different factors are difficult and expensive to carry out.

Beyond interactions of variables known to influence scoring performance, intriguing basic questions also remain unanswered. For example, are there specific cognitive abilities or personality traits that contribute to an aptitude for scoring? White (1985) noted the following:

> The ability to read essays accurately and quickly demands a peculiar skill not necessarily associated with other academic skills. Some excellent teachers never get the hang of it, while the least likely individuals often turn out to be amazingly talented as readers.
>
> *(pp. 156–57)*

More recently, Attali (2016) found that some individuals, with no relevant qualifications and with minimal training, were nonetheless able to pass a certification exercise and then score short essays with accuracy and reliability similar to expert raters. One possible interpretation of these results is that the minority of participants (40 percent) who passed certification possessed the type of "talent" mentioned by White. However, it remains unclear if such an aptitude for scoring actually exists, and if so, what specific knowledge or abilities might contribute to it.

In addition, new and more objective insights into rater cognition are becoming possible due to methodological advances associated with innovative technologies. Eye-tracking technology in particular shows promise for increasing our understanding of how raters interact with responses and scoring materials, illuminating the decision-making process (e.g., Ballard, 2017; Winke and Lim, 2015). As eye-tracking equipment becomes more widely available, we can expect such research to play a larger role in understanding rater cognition. Improved technology in the field of natural language processing has also made it easier to analyze the linguistic features of test taker responses, such as lexical sophistication, syntactic complexity, or phonological features. Such linguistic measures can then be compared to raters' judgments as another window into the criteria raters actually employ in decision making. This type of approach has been widely used in studies of accent and comprehensibility of speech (e.g., Kang, 2012; Saito *et al*., 2017), but advancements in natural language processing increasingly make possible analyses using a broad array of linguistic measures.

One example is Wolfe *et al*. (2016), in which measures of language phenomena obtained from an automated scoring system were used to identify characteristics of responses that were difficult to score: i.e., where raters tended to disagree.

It is clear that the development of scoring expertise varies with many factors, including the prior knowledge and experiences individuals bring to the scoring task and the specific training activities and materials used to develop scoring ability. While research to date has clarified the individual impact of various factors on scoring performance, in future research, it is to be hoped that new techniques, combined with frameworks for thinking about rater performance, will finally lead to a more complete understanding of how to optimize rater training for specific contexts.

Further reading

Brown, A. (2005). *Interviewer Variability in Language Proficiency Interviews*. Frankfurt am Main: Peter Lang.
Brown's study of a language proficiency interview documents the impact of the interviewer on scores, how interviewer behaviors influenced discourse with the test taker, and how raters compensated (or didn't) for the interviewer's performance. This study remains a model for how a variety of techniques (MFRM, conversation analysis, verbal report) can be brought to bear to understand the impact of interlocutor behavior.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement*. Frankfurt am Main: Peter Lang.
This concise and accessible volume describes the use of MFRM to analyze language performance assessments. The book also describes various types of rater behavior and how such behaviors can be detected using MFRM.

Knoch, U. and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing* 35: 477–99.
This article provides a detailed exploration of how various aspects of the rating process impact the claims in a validity argument. Examples are provided that illustrate how a validity argument might be used to frame research into raters and scoring.

Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective*. Frankfurt am Main: Peter Lang.
Lumley's book provides a detailed view of how raters in a writing assessment dealt with the challenge of interpreting scoring materials, which typically provide only minimalist descriptions of performance. Lumley also explores the usefulness and limitations of think-aloud protocols as a method for understanding rater cognition. The study provides another useful example of how quantitative and qualitative data can be combined to understand the scoring process.

References

American Council on the Teaching of Foreign Language (ACTFL). (2020). *ACTFL OPI Tester Certification Information Packet*. ACTFL. Retrieved from www.actfl.org/sites/default/files/assessments/ACTFL%20OPI%20Tester%20Certification%20Packet%202020.pdf.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing* 33: 99–115. https://doi.org/10.1177/0265532215582283.

Bachman, L. F., Lynch, B. K. and Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12: 238–57. https://doi.org/10.1177/026553229501200206.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly* 9: 225–48. https://doi.org/10.1080/15434303.2011.637262.

Ballard, L. (2017). *The Effects of Primacy on Rater Cognition: An Eye Tracking Study*. Doctoral dissertation. Michigan State University. MSU Electronic Theses and Dissertations. https://doi.org/10.25335/M5SR15.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly* 7: 54–74. https://doi.org/10.1080/15434300903464418.

Bock, R. D., Brennan, R. L. and Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement* 26: 364–75. https://doi.org/10.1177/014662102237794.

Brooks, R. L. (2013). *Comparing Native and Non-Native Raters of US Federal Government Speaking Tests*. Doctoral dissertation. Georgetown University. Georgetown University Graduate Theses and Dissertations. Retrieved from http://hdl.handle.net/10822/559500.

Brown, A. (2005). *Interviewer Variability in Language Proficiency Interviews*. Frankfurt am Main: Peter Lang.

———. (2007). An investigation of the rating process in the IELTS oral interview. In L. Taylor and P. Falvey (eds.), *IELTS Collected Papers*. Cambridge: Cambridge University Press, 98–139.

———. (2012). Interlocutor and rater training. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. London: Routledge, 413–25.

Cai, H. (2015). Weight-based classification of raters and rater cognition in an EFL speaking test. *Language Assessment Quarterly* 12: 262–82. https://doi.org/10.1080/15434303.2015.1053134.

Carey, M. D., Mannell, R. H. and Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing* 28: 201–19. https://doi.org/10.1177/0265532210393704.

Choi, I. and Wolfe, E. W. (2020). The impact of operational scoring experience and additional mentored training on raters' essay scoring accuracy. *Applied Measurement in Education* 33: 210–22. https://doi.org/10.1080/08957347.2020.1750404.

Cumming, A., Kantor, R. and Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal* 86: 67–96. https://doi.org/10.1111/1540-4781.00137.

Davis, L. E. (2012). *Rater Expertise in a Second Language Speaking Assessment: The Influence of Training and Experience*. Doctoral dissertation. University of Hawai'i, Manoa. ScholarSpace, University of Hawaii at Manoa. Retrieved from https://scholarspace.manoa.hawaii.edu/handle/10125/100897.

———. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing* 33: 117–35. https://doi.org/10.1177/0265532215582282.

———. (2019). Rater training in a speaking assessment: Impact on more- and less-proficient raters. In S. Papageorgiou and K. Bailey (eds.), *Global Perspectives on Language Assessment*. London: Routledge, 18–31.

Diederich, P. B., French, J. W. and Carlton, S. T. (1961). *Factors in Judgements of Writing Ability*. Research Bulletin, RB-61-15. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1961.tb00286.x.

Douglas, D. and Myers, R. (2000). Assessing the communication skills of veterinary students: Whose criteria? In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge: Cambridge University Press, 60–81.

Duijm, K., Schoonen, R. and Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing* 35: 501–27. https://doi.org/10.1177/0265532217712553.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly* 2: 197–221. https://doi.org/10.1207/s15434311laq0203_2.

———. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing* 25: 155–85. https://doi.org/10.1177/0265532207086780.

———. (2011). *Introduction to Many-Facet Rasch Measurement*. Frankfurt am Main: Peter Lang.

———. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly* 9: 270–92. https://doi.org/10.1080/15434303.2011.649381.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society* 53: 644–63. Retrieved from www.jstor.org/stable/2979547.

Elder, C., Barkhuizen, G., Knoch, U. and Von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing* 24: 37–64. https://doi.org/10.1177/0265532207071511.

Ericsson, K. A. (2006). An introduction to the Cambridge handbook of expertise and expert performance: Its development, organization, and content. In K. A. Ericsson, N. Charness, P. J. Feltovich and R. R. Hoffman (eds.), *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press, 3–19.

Ericsson, K. A. and Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*, 2nd edn. Cambridge, MA: MIT Press.

Everson, P. and Hines, S. (2010). How ETS scores the TOEIC® speaking and writing test responses. In D. E. Powers (ed.), *The Research Foundation for the TOEIC® Tests: A Compendium of Studies*, vol. II. Princeton, NJ: Educational Testing Service, 8.1–8.9.

Fahim, M. and Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. *Iranian Journal of Language Testing* 1: 1–16.

Feltovich, P. J., Prietula, M. J. and Ericsson, K. A. (2006). Studies of expertise from psychological perspectives. In K. A. Ericsson, N. Charness, P. J. Feltovich and R. R. Hoffman (eds.), *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge: Cambridge University Press, 41–67.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Longman.

Furneaux, C. and Rignall, M. (2007). The effect of standardization – training on rater judgements for the IELTS writing module. In L. Taylor and P. Falvey (eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment*. Cambridge: Cambridge University Press, 422–45.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing* 1: 1–9. https://doi.org/10.1016/j.asw.2007.05.002.

Hsieh, C. N. (2011). *Rater Effects in ITA Testing: ESL Teachers' Versus American Undergraduates' Judgments of Accentedness, Comprehensibility, and Oral Proficiency*. Doctoral dissertation. Michigan State University. MSU Electronic Theses and Dissertations. Retrieved from https://d.lib.msu.edu/etd/1282.

Isaacs, T. and Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly* 10: 135–59. https://doi.org/10.1080/15434303.2013.769545.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly* 9: 249–69. https://doi.org/10.1080/15434303.2011.642631.

Kim, A. Y. and Gennaro, K. D. (2012). Scoring behavior of native vs. non-native speaker raters of writing exams. *Language Research* 48: 319–42. Retrieved from http://hdl.handle.net/10371/86486.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly* 12: 239–61. https://doi.org/10.1080/15434303.2015.1049353.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing* 26: 187–217. https://doi.org/10.1177/0265532208101010.

Knoch, U. (2009). Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale. *Spaan Fellow Working Papers in Second or Foreign Language Assessment* 7: 21–46.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing* 28: 179–200. https://doi.org/10.1177/0265532210384252.

Knoch, U. and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing* 35: 477–99. https://doi.org/10.1177/0265532217710049.

Knoch, U., Fairbairn, J., Myford, C. and Huisman, A. (2018). Evaluating the relative effectiveness of online and face-to-face training for new writing raters. *Papers in Language Testing and Assessment* 7: 61–86.

Knoch, U., Read, J. and von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing writing* 12: 26–43. https://doi.org/10.1016/j.asw.2007.04.001.

Kuiken, F. and Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing* 31: 329–48. https://doi.org/10.1177/0265532214526174.

Lave, J. and Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing* 13: 151–72. https://doi.org/10.1177/026553229601300202.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing* 28: 543–60. https://doi.org/10.1177/0265532211406422.

Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective*. Frankfurt: Peter Lang.

Lynch, B. K. and McNamara, T. F. (1998). Using G-Theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing* 15: 158–80. https://doi.org/10.1177/026553229801500202.

Mackey, A. and Gass, S. M. (2016). *Stimulated Recall Methodology in Applied Linguistics and L2 Research*. London: Routledge.

Milanovic, M., Saville, N. and Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic and N. Saville (eds.), *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem*. Cambridge: Cambridge University Press, 92–114.

Mislevy, R. J. (2010). Some implications of expertise research for educational assessment. *Research Papers in Education* 25: 253–70. https://doi.org/10.1080/02671522.2010.498142.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher* 23: 5–12. https://doi.org/10.3102/0013189X023002005.

Myford, C. M. and Wolfe, E. W. (2000). *Monitoring Sources of Variability Within the Test of Spoken English Assessment System*. ETS Research Report Series. Report No. RR-00-6. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01829.x.

Nakatsuhara, F., Inoue, C. and Taylor, L. (2020). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*. Advance online publication. https://doi.org/10.1080/15434303.2020.1799222.

O'Hagan, S. R. and Wigglesworth, G. (2015). Who's marking my essay? The assessment of non-native-speaker and native-speaker undergraduate essays in an Australian higher education context. *Studies in Higher Education* 40: 1729–47. https://doi.org/10.1080/03075079.2014.896890.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System* 30: 143–54. https://doi.org/10.1016/S0346-251X(02)00002-7.

O'Sullivan, B. and Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor and P. Falvey (eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment*. Cambridge: Cambridge University Press, 446–78.

Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly* 36: 219–33. https://doi.org/10.2307/3588333.

Randall, J. and Engelhard Jr, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement* 46: 1–18. https://doi.org/10.1111/j.1745-3984.2009.01066.x.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing* 9: 173–85. https://doi.org/10.1177/026553229200900205.

Şahan, Ö. and Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing* 37: 311–32. https://doi.org/10.1177/0265532219900228.

Saito, K., Trofimovich, P., Isaacs, T. and Webb, S. (2017). Re-examining phonological and lexical correlates of second language comprehensibility: The role of rater experience. In T. Isaacs and P. Trofimovich (eds.), *Second Language Pronunciation Assessment: Interdisciplinary Perspectives*. Multilingual Matters, 141–56. Retrieved from www.jstor.org/stable/10.21832/j.ctt1xp3wcc.12.

Shanteau, J. (1992). The psychology of experts: An alternative view. In G. Wright and F. Bolger (eds.), *Expertise and Decision Support*. New York: Plenum, 11–23.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes* 8: 13–17. www.cambridgeenglish.org/Images/23120-research-notes-08.pdf.

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (ed.), *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service, 1–12.

Spolsky, B. (1995). *Measured Words*. Oxford: Oxford University Press.

Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), *Examining Speaking: Studies in Language Testing 30*. Cambridge: Cambridge University Press, 171–233.

Wei, J. and Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly* 12: 283–304. https://doi.org/10.1080/15434303.2015.1037446.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing* 11: 197–223. https://doi.org/10.1177/026553229401100206.

———. (1998). Using FACETS to model rater training effects. *Language testing* 15: 263–87. https://doi.org/10.1177/026553229801500205.

———. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2003). A survey of the history of the certificate of proficiency in English (CPE) in the twentieth century. In C. J. Weir and M. Milanovic (eds.), *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*. Cambridge: Cambridge University Press, 1–56.

———. (2005). *Language Testing and Validation*. Houndgrave: Palgrave Macmillan.

White, E. M. (1984). Holisticism. *College Composition and Communication* 35: 400–9. https://doi.org/10.2307/357792.

———. (1985). *Teaching and Assessing Writing*. San Francisco: Jossey-Bass.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10: 305–19. https://doi.org/10.1177/026553229301000306.

Wilds, C. P. (1975). The oral interview test. In R. L. Jones and B. Spolsky (eds.), *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 29–44.

Winke, P., Gass, S. and Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing* 30: 231–52. https://doi.org/10.1177/0265532212456968.

Winke, P. and Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs *et al*. rubric: An eye-movement study. *Assessing Writing* 25: 37–53. https://doi.org/10.1016/j.asw.2015.05.002.

Wolfe, E. W. (2006). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment* 2: 37–56.

Wolfe, E. W., Matthews, S. and Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *Journal of Technology, Learning, and Assessment* 10. Retrieved from https://ejournals.bc.edu/index.php/jtla/article/view/1601.

Wolfe, E. W., Song, T. and Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing* 27: 1–10. https://doi.org/10.1016/j.asw.2015.06.002.

Xi, X. and Mollaun, P. (2009). *How Do Raters from India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?*. Research Report No. RR-09-31. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02188.x.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing* 27: 37–53. https://doi.org/10.1016/j.asw.2015.11.001.

Zhang, Y. and Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing* 28: 31–50. https://doi.org/10.1177/0265532209360671.

———. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the college English test-spoken English test (CET-SET). *Assessment in Education: Principles, Policy & Practice* 21: 306–25. https://doi.org/10.1080/0969594X.2013.845547.

Section 6
Writing items and tasks


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

20

Item writing and item writers

Dongil Shin

Introduction

In the last few decades, positivistic worldviews and related scientific testing techniques have influenced the development and evaluation of second and foreign language tests (McNamara and Roever, 2006). Relatedly, research using psychometric approaches to scored/rated data has been encouraged and effectively linked to professional language testing literature. However, more effort has been given to rater reliability and rater characteristics than to item construction issues related to validity.

The general procedures for item writing have been repeated and elaborated in language testing (Alderson *et al*., 1995; Bachman and Palmer, 1996; Hughes, 2003; Spaan, 2006, 2007) and educational measurement literature (Downing and Haladyna, 2006). Some researchers (Davidson and Lynch, 2002; Fulcher and Davidson, 2007) approach item writing as consisting of the process of writing test specifications (or "test specs") and building validity arguments as a system. However, both the actual process through which items are developed from test specs and the people who write test items remain hidden areas of active inquiry. Research and development on the item writing process and item writer training (for example, Peirce, 1992) have not been properly introduced to testing communities while issues related to rating and rater training have often appeared in language testing literature.

The purpose of this chapter is to summarize the critical issues and research topics surrounding item writing and item writers by reviewing different historical views of item writing. However, this chapter does not focus on the scoring dimension of item writing, despite the intricate relationship between the two. When test items are developed to be objectively scored, the scoring is strictly a matter of item creation; for instance, if it is a multiple-choice formatted item, then the item writer's responsibilities include the crafting of good distractors linked to the item. Items scored by raters in English-speaking tests or admissions-related portfolios examined by experts involve complicated steps beyond the item creation process. These types of items require the development of some kind of scoring rubric or training system (see Galaczi and Lim, this volume; Davis, this volume). In these instances, answering the following questions remains complicated: Where does the item writer's responsibility end? When does responsibility shift to the province of test administration, scale calibration,

DOI: 10.4324/9781003220756-27

rater training, and so forth? Although these are interesting and important questions, addressing them lies beyond the scope of this chapter. Instead, I will focus on item writing in and of itself.

Historical perspectives

Many existing books focus on item development for second and foreign language testing. Some of these books detail the item types and sample items that item writers can use to test for vocabulary (Read, 2000), grammar (Purpura, 2004), listening (Buck, 2001), reading (Alderson, 2000), speaking (Fulcher, 2003; Luoma, 2004), writing (Cushing, 2002), and language for specific purposes (Douglas, 2000). These resource books often list a number of different item types to operationalize the construct of the target test.

In this section, I summarize four different perspectives (psychometric, authentic, systematic, and critical) to elucidate both past and present approaches to item writing. It should be noted, however, that all four approaches are not mutually exclusive. Additionally, more recent perspectives have not necessarily made older ones obsolete. For instance, while authentic test items are pervasive now, the psychometric tradition remains strong. The systematic approach has facilitated advances in the professionalization of the item writing process, but research confirming its usefulness remains scarce. Critical approaches (e.g., Foucauldian frames) to testing have been explored in language education, but micro-level issues in item design have received little coverage in literature when compared to macro-level arguments against test misuse.

The psychometric approach or "Reproduce similar items in the multiple-choice format. Do it objectively."

Dominant paradigms, assessment cultures, and the sets of assumptions behind social practices always influence the item creation process (see Lynch and Shaw, 2005 for examples). In the second half of the twentieth century, the language testing industry grew under a positivistic paradigm, which had a quantitative, objective, and outcome-oriented view of testing. This paradigm maintained that language testing and testers can, and should, be objective. Testers preferred context- and time-free inferences from well-standardized test items.

The multiple-choice format was preferred to test language proficiency in general as well as achievement in classroom contexts. True-or-false and matching item types were also welcomed. These formats were publicized as scientific. Much research was conducted to enhance test reliability in relation to testers' biases in written items (see McNamara and Roever, 2006: Chapter 4). The multiple-choice format was also powered by a tradition of discrete point testing. The psychometric approach to item development has continued to evolve not only to measure separate points of knowledge (e.g., grammar, vocabulary, and pronunciation) but also to tactically infer advanced levels in the four macro-skills of language proficiency (listening, reading, speaking, and writing) in standardized formats. This tradition remains foundational in assessing language learning processes all over the world, and resources (books, courses, multimedia products) for test preparation are conventionalized using the same categories. Taxonomies of multiple-choice item writing guidelines (principles or technologies in other names) are widely available in the field of educational measurement (Haladyna and Downing, 1989; Haladyna *et al*., 2002; Roid and Haladyna, 1982).

During this period, the key principle of item writing was to "reproduce similar items in multiple-choice format. Do it objectively." As elaborated by McNamara and Roever (2006: 1), this was an era in which "psychometrics became the substrate discipline, prescribing the rules of measurement, and language was virtually poured into these preexisting psychometric forms"; the testing industry had poured separate points of language knowledge into a psychometrically efficient mold: the multiple-choice format of testing.

The authentic approach or "Search for authentic items. Avoid the multiple-choice" format

While the psychometric approach has remained influential, alternative paradigms have emerged (Hamp-Lyons and Lynch, 1998), and diverse formats of language test items (e.g., authentic items in self- or peer assessment, simulations, and diaries) are now associated with qualitative, subjective, and process-oriented views of testing (Brown and Hudson, 1998). The key principle behind item writing under this approach was "Search for authentic items. Avoid the multiple-choice format." Item writers tried to incorporate real-world value and practical language skills, and they chose items that appeared integrative, authentic, communicative, and pragmatic from real-world language-use situations.

Because multiple-choice, discrete point testing had been criticized for placing too much emphasis on students' knowledge of the linguistic system, integrative test items, in the format of direct tests of speaking and writing (e.g., oral proficiency interviews) became highly favored, along with tests targeting comprehension of extended discourse. This approach was exposed to repeated criticism (mostly from psychometric testers) for having subjective scales, unreliable raters, and impractical test administration (e.g., time-consuming procedures for test taking, scoring, and expert training). However, the proficiency movement in the United States continued promoting this type of testing to assess productive skills for real-world communication. In the 1970s, the cloze test was a popular alternative format in which test takers were asked to fill in the missing words in an extended text. The cloze test was researched and considered a substitute for integrative tests (Oller, 1972; Stubbs and Tucker, 1974), but its use is questionable and controvertible (McNamara, 2000: 15). There is, for example, disagreement regarding what constructs are related to test design and score interpretation, as well as how and whether specific language skills can be meaningfully studied for cloze test activities. The communicative movement, theoretically guided by Hymes's theory of communicative competence, also impacted the way authentic items were developed. Meanwhile, assessments of second language pragmatics, such as discourse completion tests, also emerged (McNamara and Roever, 2006: Chapter 3).

The expansion of anti-psychometric approaches to test items and the growing value placed on the multiple-constructed realities of inquiries have been accompanied by humanism, hermeneutics, constructivism, and postmodernism. These approaches do not seek one correct "true" answer for each test item. Instead, this type of item creation requires the adoption of context-driven and value-bound approaches to item development and response interpretation processes. All the items that have appeared in the name of authentic assessment or performance assessment can be considered separate in quality and character from the multiple-choice format.

The systematic approach or "See the whole picture of test development. Do it systematically."

In the past decade, many language testers have avoided identifying as either psychometric or authentic purists. They have realized that no specific item type can function perfectly to make inferences about test constructs. All items

> have their particular strengths and weaknesses, and tend to engage different skills: some assess one set of skills and some another; some advantage one group of test takers, or another. By using a variety of different task types, the test is far more likely to provide a balanced assessment.
>
> *(Buck, 2001: 153)*

For test items to be comprehensively balanced, they need to be characterized, or coded, for specific features (for an example of a literature review of the testing of oral proficiency, see Buck, 2001: Chapter 5; Fulcher, 2003: Chapter 3). Bachman and Palmer's non-test "target language use" (TLU) checklist is one of the most frequently referenced frameworks for comparing test tasks in non-test settings (e.g., in the volumes of the Cambridge University Press Language Assessment series). It was believed that the situational authenticity of TLU-ness in test tasks was positively related to the validity argument.

Following Fulcher (2000, 2003), Fulcher and Davidson (2007) challenged this emerging tradition. They argued that "authenticity, even conceived of as matching test method facets to facets of similar tasks in the real world (Bachman and Palmer, 1996), does not make test tasks automatically valid through directness" (63). Instead, they tried to design test items as "the 'response as evidence' to the construct as language knowledge, or as the ability for use within specified parameters laid out in a framework and test specification" (Fulcher and Davidson, 2007: 63). They endorsed a systematic framework called evidence-centered design (ECD), which has been well received by the testing community because of its solid chain of reasoning (Mislevy *et al*., 2002, 2003; see also Yin and Mislevy, this volume). What makes ECD unique is that test items are specified, and the anticipated responses are systematically planned to claim "a priori validity" arguments. Kim's (2008) empirical research is a good example of how ECD is utilized in test development and validation contexts. When designing an ESL diagnostic reading-to-write test in a university context, Kim conceptualized the desirable effects of test items and considered how these effects could be achieved by using ECD as an assessment framework.

Because ECD functions as an initial platform in an overall system, test specifications can provide detailed information about the individual design of different test items (Fulcher and Davidson, 2007: 70). It has been widely accepted that clear documentation of the test specification process contributes to formulating and evaluating the validity argument and that test items need to be well discussed and clearly characterized at the stage of test specification creation. Unlike those written in the 1970s and 1980s, many language testing books published in recent decades have included more room for test specifications (see Alderson *et al*., 1995; Bachman and Palmer, 1996; Fulcher, 2003). Davidson and Lynch drew on previous literature and conceived of an "efficient generative blueprint by which many similar instances of the same assessment task can be generated" (2002: 4). Examples of research focusing on the value of test specifications continue to appear (Fulcher and Davidson, 2009; Harsch, 2018; Kim *et al*., 2010; Walters, 2010; Zandi *et al*., 2014). The key principle of item writing in relation to this approach is explicit: "See the whole

picture of test development. Do it systematically." By taking a pragmatic stance, Fulcher and Davidson (2007) and Davidson and Lynch (2002: 114) approached test design and validation as a system. A pragmatic approach does not accept determinism outright but instead conceives of testing practices as both contingent on and inextricable from complex variables. Additionally, it is seen as open to different methods of item creation and validation while it does not accept any unchanging, universal formula for test item development. Davidson and Lynch stated that their "approach to test development is intended to be inclusive, open, and reflective" and that any test specifications "will fit all settings very rarely – or more likely never" (2002: 2).

The critical approach or "Create items to help test takers. Think socio-politically."

Item writing was once judged to be "an immature science" (Cronbach, 1970), and researchers reacted and ultimately circumscribed the item writing process with a tightly organized set of guidelines or validity arguments. However, some scholars (e.g., Lynch, 2001; Pennycook, 2001; Spolsky, 1995) have tried to situate testing practices within a much larger sociopolitical picture. In articulating "critical language testing," for example, Shohamy (2001: 131–32) framed tests as political tools in the context of power struggles. This approach challenges dominant traditions for test use and encourages an active, critical response from test takers. McNamara (2008), who has echoed critical perspectives regarding the misuse of language tests, subscribed to the views of many scholars who made major contributions to the socio-political dimensions of language testing. Acknowledging the political characteristics embedded in the process of test design and use, Hamp-Lyons (1997, 2000) called for ethical awareness in language testing and referred to "ethical language testing" when discussing the notion of "responsibility."

From the perspective of critical language testing, objective test formats, such as multiple choice or true-false, are the features that grant language tests power. The assumption behind multiple-choice test items is that they have "one correct answer, i.e., one 'truth,' one interpretation" (Shohamy, 2001: 24); thus, different approaches to item design are needed to allow for the different meanings and interpretations of different test takers (see Shohamy, 2001: 27, 111). This criticism is different from the way other approaches to item writing have recognized multiple-choice items as either meaningful or not meaningful. The psychometric approach sees multiple-choice items as scientific entities. The authentic approach distances itself from multiple-choice items with a preference for utilizing real-world authentic tasks, and the systematic approach focuses on how items support inferences from the response evidence to the claims testers make about the test construct.

It should be noted that while the systematic approach to test design is well connected to test use and its validation argument, the critical approach focuses on test misuse – that is, the negative impacts of test use. It may be that each approach has its own strengths for professionalizing test designs and uses as theory, and the critical perspective may be functioning not as "a replacement for current testing practices, but as an additional approach to making decisions and judgments about individual language ability" (Lynch, 2001: 361). Critical language testers, however, need to pay more attention to the dimensions of item design that are connected with intended or unintended misuses of tests. Even though the general features of critical language testing (e.g., a democratic system, an empowerment model, and multiple assessment procedures) have appeared in thought-provoking literature (Shohamy, 2001), critical theorists have not yet suggested an alternative item creation protocol.

What would a critically influenced test item look like? How would we develop critical language test items? Certain rules of standardized test administration (e.g., the same questions and test-taking steps) may be removed. The creation of a test could be heavily influenced by test takers and other stakeholders. For example, drawing on postmodernist notions of multiple interpretations, Tunks (2001) and Yu (2007) invited test takers to contribute to the creation of items and scoring templates. Lynch and Shaw's (2005) portfolio assessments also illustrated how students actively participated in the assessment process and reflected on their identities within a school context. Even if Lynch and Shaw (2005) used different terms, "alternative assessments operationalized from an interpretivist perspective," their Foucauldian framework of portfolio assessment provides a meaningful procedure for considering a critical approach as part of the larger picture of item design and validity argument. In their discussions of portfolio design and use, they explicitly linked power relations to the ethics and validity of assessment inferences.

Critical issues

Who are the item writers?

If the recruitment and education of item writers are important stages for professional testing organizations, in-service teacher training programs, and testing courses, who are the individuals who develop test items? Who trains these individuals? Although language testing research is awash with studies regarding the development and administration of diverse tests and often concerns the scrutiny of test items, "who" produces test items remains a fairly under-researched area (Ryan and Brunfaut, 2016). The voices of item writers who raise their own concerns (see Peirce, 1992, for an item writer's experiences at the Educational Testing Service) have been even rarer. Unlike other language testing professionals (such as test designers, raters, rater trainers, data analysts, and validity evaluators), item writers seem to be viewed as quasi-professional, and item writing is viewed as a kind of abstract skill to be acquired in the testing world.

Davidson and Lynch highlighted the importance of the people who generate and utilize test items and specifications, claiming, for example, "test development is most efficient when human resources are invested" (2002: 121). What images flash through our minds when we think of item writers? Native speakers of the target language or full-time professional testers in collared shirts who drive to their offices each morning? As argued by Green and Hawkey, the field of language testing is interested in "exerting more direct control over item writing processes to achieve a transparent and consistent relationship between item design and the measurement properties of the resulting items. From this perspective, the identity of writers should, ideally, be a matter of indifference" (2011: 110).

Kim *et al*. (2010) examined the individual characteristics of item writers, including their previous experience (prior training in language testing and ESL/FL teaching/working experience), language background (nativeness/foreignness), cultural background, personality, and topic preferences to determine how these characteristics affected their item writing processes and reduced ambiguity. They discovered that feedback and revision processes are related to individual item writers' personalities (e.g., "affable," "harmonious but willing to state their opinion," "dogmatic and heavily opinionated," "introverted") and their social identities (e.g., native/non-native speaker). They therefore argued that personality and language statuses do not represent "absolute qualification[s]." However, an awareness

of individual characteristics is obviously necessary to group, train, and facilitate feedback exchange among item writers.

Their characteristics, however, are often dichotomized or categorized into somewhat conventional categories, such as whether they are native speakers, have teaching experience, or have professional degrees. The narratively constructed characteristics of item writers and how these characteristics relate to test standardization processes have not yet been fully described. Because many journal articles and books have described and narrated stories of learners and teachers, it should be noted that item writers also live their own complicated lives.

What do they do?

The job routines that item writers go through are also under-researched. Item writing is acclaimed as a "creative art," but it is often a tedious process that involves "just cutting it and cutting it and cutting it, deciding which information you can target and which bits of the text will be suitable for particular task types" (Green and Hawkey, 2011: 118). Salisbury elaborated three stages that the item writers in her study of listening comprehension test items passed through – namely, "an *Exploratory Phase* (searching through possible texts), a *Concerted Phase* (working in an intensive way to prepare text and items for first submission) to a *Refining Phase* (after either self-, peer- or editor-review, polishing/improving the test paper)" (2005: 75).

In their study of item-writing processes for an academic reading test, Green and Hawkey (2011) summarized the specific steps taken by item writers in greater detail. These specific details are as follows (115): (1) "text sourcing"; (2) "careful reading"; (3) "typing up and making slight amendments"; (4) "length adjustment"; (5) "work on first (most obvious) task type"; (6) "mark up further areas of text for suitable items, get an idea of how many items are possible for each section"; (7) "work on tasks, amending texts as I go"; (8) "print off and attempt tasks, cutting text that might give double keys and inserting more distractions if necessary"; (9) "write answer key;, (10) "check text for length and prune if necessary"; and (11) "review and proof read." Green and Hawkey also listed the following edits for textual authenticity that item writers make: deletion, consolidation, expansion, permutation, moving, substitution, and insertion. A detailed examination of how item writers select and adapt texts to fulfill target test requirements remains neglected. Researchers have generally underestimated the complexity of item-writing decision-making processes. Further research should offer insights into the intensive individual/collective works used in the tests. Debates should occur over how source texts are edited for testing purposes; these discussions should also consider the usefulness of the revision process in aiding test validity to elucidate these areas.

There has been considerable interest in "scientifically grounded, algorithmic laws of item writing to counter traditional guidelines . . . empirical research into the validity of item writing rules . . . efforts to at least partially automate text selection" (Green and Hawkey, 2011: 110). In this vein, language testers who take a systematic approach have accentuated the usefulness of explicit item development procedures, item construction guidelines, or test specifications to guide item writers to successfully complete the linear (or iterative) processes of source text selection, editing, item writing, and revising (Davidson and Lynch, 2002; Fulcher and Davidson, 2007). Digital technology has also allowed for many item writing processes to become fully automatized. However, it should be noted that comparisons of automated item generation (AIG), in which algorithms create multiple-choice items,

with traditional methods (e.g., Pugh *et al*., 2020) indicate that AIG has the potential to dramatically change systematically conventionalized item development processes or eliminate the role of experienced item writers in the future testing industry.

Some empirical studies focusing on the use of item specifications, according to item writers, have appeared, but how test items are "'really' developed in the real world" is still an area of research that has received too little attention. Testing experts may need to be paired up with target language informants (Ryan and Brunfaut, 2016). It should be admitted that (experienced) item writers' predictions regarding item difficulty are often inaccurate (Sydorenko, 2011). When item writing according to test specifications is encouraged as part of the larger picture of test design, the spec-to-item process must become a reality for stakeholders. Is it true that better items can be produced using test specifications? Is the spec-to-item flow applicable and feasible? Or do hidden sociocultural factors impede the item-writing process based on test specs?

A case study by Kim *et al*. (2010: 163) found that item writers had negative feelings and attitudes toward test specifications, which they described with words such as "abstract," "mysterious," "difficult," "less transparent," "not helpful," and "heuristics." Why is this the case? When the systematic design of the spec-item fit was considered, why did the process still seem to fail on some levels? Kim *et al*. also found that item writers used their own internal guidelines for item-writing processes, such as their teaching experiences, other item-writing practices, face-to-face meetings, and oral communications with item coordinators. Thus, different forms of test specifications or "organic guidelines" are needed to utilize spec-to-item workflows. This is just one worthwhile lesson from real people who document the item-writing process, and more questions should be proposed to challenge the taken-for-granted reality that the spec-item process always functions successfully as quality assurance. In other words: Do item writers feel comfortable with the conventional format and content in test specs? Do they really need multiple (or specific) forms of test specifications? Different stakeholders, purposes, and cultures of testing can lead to different ways of thinking.

Is item writing a social practice?

Item writing has been considered both an individualized and a collectively structured procedure (Salisbury, 2005). Kim *et al*. (2010) found that the item writing process involves group dynamics and that group formation, collaborative frames, and members' hierarchies influence the effectiveness of teamwork, which includes the satisfaction and responsibilities of item writers. A more thorough investigation, however, needs to be conducted in a much broader social context (e.g., for tests in less-commonly taught languages, reported by Ryan and Brunfaut, 2016) or beyond a system of behaviors or psychological processes within and between groups of item writers.

In taking a more socially oriented view, researchers may find the intentions and effects in item writing to be inappropriate in the context of test item creation. This might impede effect-driven assessment or bias the appropriate use of tests. Item writing practices often involve factors beyond cognitively driven decision-making processes; power, culture, and monetary issues are of importance. Experience, education, and beliefs about how items should/can be constructed also matter. When item writers are underpaid, when parts of item-writing jobs (e.g., searching for relevant passages on the internet or editing/proofreading item drafts) are unethically foisted upon part-time or freelance writers, when some cultures of communication do not allow group dynamics to be horizontal, and when job coordinators

or peers take too much power in decision making, then even good test specifications or well-qualified item writers cannot be guaranteed to generate good items for the testing topics. In this vein, item writing needs to be understood as a social practice.

Both the images and job descriptions of item writers are naively idealized, and little research has elaborated upon their socially constructed identities as (professional) testers. Literature in which item writing is seen as a social practice within group dynamics or in relation to overarching power structures remains sparse. In testing companies and even school-based testing contexts, it is highly likely that testers are weakly positioned within their social structures; thus, they are often tempted to leave their jobs, change career paths, or make ethically inappropriate decisions, like plagiarizing and releasing test materials unlawfully. In the Korean context, for example, I often see newspaper articles exposing testing stakeholders, including item writers and reviewers, for cheating. Few studies have examined why these stakeholders break the law.

Shohamy (2001) re-examined language testing policies and practices through a Foucauldian lens as a way to empower test takers. Through this lens, she conceptualized test takers and testers as mismatched in their access to power. In Chapter 2 of her book, Shohamy (2001) described all the complaints, chagrin, and fears of test takers while showing how testers are viewed as oppressive. However, this dichotomy needs to be carefully considered. For example, Bachman (2000: 17) was skeptical of the "perpetrator-victim" characterization. This reductive dichotomy neglects the multifaceted and changing life of testing and tested individuals, and early critical theories continue to fail in adequately addressing this complexity. If test-taking students are categorized as the suppressed group of the testing industry, who are the suppressive individuals? Where do the item writers belong? Item writers are both oppressed (underpaid, casual contracts, little recognition of their skills) and oppressors (selecting and deciding what test takers will be tested on and how). More research on the relationship between item writers and their social world from poststructuralist perspectives needs to be conducted (Morgan, 2007; Norton, 2000).

Work in language testing can be divided into two levels: macro-level studies that consider theories and systems of testing and micro-level studies of specific facets of test development and analysis. Item writing is micro-level work, and item writers need to use their professional experience to publish their own item-writing guidelines or (in)validate current conventions introduced in language testing resources or taught in workshops or training programs (see Haladyna *et al*., 2002). The social approach to item writing may be outside the usual scope of item writers' responsibilities, but arguably, the main job of an item writer is to write items based on what has been given in the form of test specifications and other testing documentation. This raises the question: Are item writers test designers too? A door should now be opened for item writers to be educated on these topics, and related researchers should be invited into macro-level discussions (Constantinou *et al*., 2018). Combined analyses of the micro-level practices that item writers engage in and their sociocultural contexts needs to be conducted.

Current contributions and research

Untrained item writers are presumably less effective than those trained through formalized training processes. If this is the case, what kind of curricula, workshops, course work, and materials will improve item writing for the specific context of test development and validation? How can item writing training be standardized based on activities such as reviewing item-writing guidelines, creating sample items, sharing feedback, and editing? Research

offering any insights into the standardization of training programs remains scarce. Cambridge ESOL, for example, developed the RITCME (recruitment, induction, training, co-ordination, monitoring, and evaluation) framework to train and develop external partners, including item writers (Ingram, 2008). In the first half (RIT), potential writers are recruited and are then are allowed to take (pre-)training in item-writing skills and related information. They are often assigned to item-writing teams for team-specific trainings, which include three stages: the "adaptation of written sources, feedback on live test performance, and selection of texts" (p. 7). In the second half (CME), item writers receive feedback on their materials and item-writing performance. They are also monitored to ensure satisfactory item acceptance rates.

Any specific framework or, indeed, any planning, coordination, certification, or evaluation process needs to be empirically validated with adherence to the language testing literature. Different kinds of workshops, such as time-intensive weeklong programs, prolonged multiple sessions, or online short courses of varying lengths over extended periods of time, may have differential impact on the quality of test items produced. Are the item statistics obtained from either classical item analysis or item response theory used as feedback to improve item writing during the training process (Spaan, 2007)? Related research can provide key evidence for building validity arguments for test development and administration (Downing and Haladyna, 1997).

Non-native speakers proficient in the target language from countries where many students prepare for tests used across international contexts need to be invited to participate in item-writing research and development teams. How do they handle challenges related to the following issues: maintaining authenticity in source text searches and edits, understanding culture-specific and unfamiliar content, and utilizing collaborative strategies? Interestingly, native speakers of English are also marginalized in international testing contexts. In Korea, for example, a majority of Korean educators take full responsibility for developing foreign language tests. To facilitate discussions of empowerment, collaboration, and democracy, the opportunities to access the decision-making process need to be investigated in relation to different contexts of item development. One study demonstrating how non-native speakers of English are positioned in the spec-to-item mechanism is Shin and Kim (2010). They collected data from test writers' and reviewers' meetings to illustrate the positive role of test specifications in item writing and revision processes. They then categorized and matched all the frequently used words and phrases from the meetings with the language used in the test specs, finding a mismatch between what was documented in the specs and what was actually discussed in the meetings.

As previously discussed, a thorough documentation of item writers' voices does not yet exist, so caution is still required in evaluating the systematic feasibility of the item-writing processes described in the language testing literature. In the field of language testing, item writing is one of the most poorly articulated job descriptions. Item writers are viewed as important, but their specific roles are not qualitatively elaborated on and valued in case studies, action research, or narrative inquiries. It is critical for future studies to address this topic. Similarly, shifting perspectives to see item writing as a social activity is timely and meaningful. A social approach to item development may allow for an expanded and more inclusive range of language test development and validation theories. However, distinguishing the social practices or ideological dimensions from the scattered and muddled experiences of item writing is not easy.

Item-writing processes involve constructing test specifications, draft writing, revisions, peer reviews, pre- and post-test reviews, and test specification modifications, and it seems obvious that external and sociocultural forces affect these processes. To make arguments

about item writing processes as social activities, researchers should conduct ethnographic research or critical discourse analyses to elaborate the reasons that item writing needs to take a social approach. Theoretical or empirical arguments need to be presented to demonstrate the limitations of the psychometric view of item writing or to show what might emerge from viewing item writers' experience as a socio-political activity. The mapping of a sociocultural framework for item writing needs to be feasible and workable. Item writer narratives detailing experiences of being an item writer should be welcomed. Organizing these accounts in coherent sequences or using thick description will illustrate item writers' lived experiences, group dynamics, power structures, and collaborative and iterative processes; for example, "from [a] sociocultural theoretical lens" and "through (auto)ethnographic data" (Elwood and Murphy, 2015; Johnson *et al*., 2017; Ngo, 2016). It should be noted, however, that conducting such research will always be constrained by issues relating to security and confidentiality. Receiving approval within target organizations for research designs focused on item writers' roles and power relations may prove difficult.

Future directions

The language testing literature contains an overabundance of information about test item formats and content types, but it seems likely that language testers will continue to explore item writing in new domains such as tests of pragmatics, integrated skills, process-oriented performance, and language for specific purposes (e.g., English proficiency for the aviation industry). The use of classroom-based assessment items is likely to be the subject of much additional research. High-stakes test items continue to be used to measure general proficiency or a wide range of expected knowledge, and small-scale testing is still necessary to make direct connections to everyday learning. Additionally, different stakeholders will likely be empowered to magnify the meaningfulness of different testing purposes at school.

With the outbreak of the COVID-19 epidemic in 2020, item development and validation in online environments should be seriously revisited as a future direction of research. It has previously been argued that test items embedded in multimedia environments may not function as authentic or communicative, except in some limited situations (see Chalhoub-Deville, 1999; Bachman, 2000: 9). However, the global pandemic disrupted previously long-held views about what constituted "authentic" language use. The way in which language has continued to be acquired, learned, and taught "online," spontaneously and interactively, throughout the pandemic seems to align well with the original tenets of the communicative approach. Talking, teaching, and learning through online technology is frequent in real life now. As a result, item writers and researchers need to examine previously held views about the meaningfulness and authenticity of test items created and administered in virtual spaces. Similarly, the rapid rise of artificial intelligence (AI) "assistants" (Siri, Alexa, etc.) and the increasing prevalence of speech-to-text technology mean that we are increasingly speaking to non-human interlocutors. This situation brings into question arguments made previously regarding the "fantasy" of technological affordances in language testing, such as the argument put forward in McNamara (2000):

> In the dazzle of technological advance, we may need a continuing reminder of the nature of communication as a shared human activity, and that the idea that one of the participants can be replaced by a machine (in speaking test, for example) is really a technological fantasy.
>
> *(p. 85)*

What about the multiple-choice format, the controversial item type criticized for mostly pursuing reliability and efficiency at school or in the testing industry? The practice of testing language abilities with multiple-choice items has evolved in newly developed tests such as the TOEFL iBT. All over the world, studies of multiple-choice tests have appeared in educational measurement and language testing literature, most of which have neutrally analyzed the features of the format (see Shizuka *et al*., 2006 to review the effects of reducing the number of options per test item). Critical language testers harbor antipathy toward overexposed formats and content due to power maintenance. Investigative journalists have also voiced concerns regarding how standardized testing emerged and how the multiple-choice format survived in politically charged marketplaces. Kohn insisted that "the most damaging testing programs are characterized by certain readily identifiable features, beginning with the use of exams that are mostly multiple choice" (2000: 11). However, AIG technology may emerge as a "relief pitcher" that will help sustain the value of efficiency in the multiple-choice format. The fact that the items developed by AIG demonstrate comparable quality (Pugh *et al*., 2020) means that heated discussions about related ethical and social issues are likely to occur in the near future.

Can we teach and learn meaningfully without pursuing an efficiency-oriented item-writing paradigm? A variety of avenues of research are awaiting inquiry from the integrative/alternative, systematic, and critical testing paradigms. If we guard against the efficiency paradigms behind the multiple-choice format, then principle-based item creation (Shohamy, 2001) needs to be fully discussed. In her discussion of alternative and democratic testing, Shohamy emphasizes that "such methods must be more time consuming and costly. However, democratic practices are not chosen for their efficiency or costs; rather they are chosen because of their principles" (2001: 139). Can we imagine different testing worlds without multiple-choice items? How can we take responsibility for diverse educational needs by using item formats beyond multiple choice?

Different perspectives need to be collected to establish a philosophical foundation for item development and validation. Why are standardized items still necessary? How do they function in society? Do we take a view of the testing industry, focused on efficiency, that is optimistic, pessimistic, or somewhere in between? Do we take a utilitarian/pragmatic or critical/emancipatory worldview of testing? Can we reconcile theories of neoliberalism, pragmatism, globalization, poststructuralism, and Freirean pedagogy? Different ontological, epistemological, and methodological stances should be elaborated to trace the bases of each worldview of testing and to drive new test items teleologically (Lynch, 2001; Lynch and Shaw, 2005). Because the world of language testing is generally regarded as still open and incomplete, language testing researchers can play a role in creating and negotiating multiple realities or paradigms. Language testing has often been considered a science, and a disconnect exists between the practice of item development and related philosophical debates. I hope that future directions for item writing will be related to macro-level discussions, ultimately leading to broader conversations about test validity.

Further reading

Alderson, J. C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

Chapters 2 and 3 of this volume have useful information on test specs, item types, and item writing. Their practical tips include (1) specs need to be adjusted to different audiences for different contexts

and (2) item writing is best practiced in groups. Their guidelines remain effective for and relevant to recent discussions of the item-writing process.

Bachman, L. F. and Palmer, A. S. (2010). *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford, UK: Oxford University Press.
This book's strength is that it elaborates all the nuts and bolts of the assessment use argument (AUA) framework for the entire test development and validation process. The authors have identified a great deal of useful information by conceptualizing test design and score interpretation through AUA and by illustrating specific examples for different contexts.

Davidson, F. and Lynch, B. K. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT: Yale University Press.
This concisely written book focuses on applying test specs to the creation of valid test items. Item writers' experience and wisdom, in the forms of principles or guidelines, are integrated to create a unified system for the specs-to-item writing process. This idea has been expanded in a recent publication on test retrofit (Fulcher and Davidson, 2009), much of which is metaphorically explained using architecture.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York, NY: Routledge. The practice of item development requires analytic and integrated language testing skills.
This book is well designed to teach practitioners these skills and includes extensive discussion and application exercises. The notion of effect-driven testing is useful for understanding how item writing, as an iterative process, needs to be transparently regarded within the full picture of test development and validation.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses and Consequences of Language Tests*. London, UK: Longman.
Although this book may appear disconnected from work containing guidelines/systems for item development, it effectively explains why item writing should not be viewed only as a science. Stohamy's claims that intended effects can lead to unintended and unexpected consequences deserve thoughtful attention in all areas of test design.

References

Alderson, J. C. (2000). *Assessing Reading*. Cambridge, UK: Cambridge University Press.

Alderson, J. C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing* 17: 1–42. https://doi.org/10.1177/026553220001700101.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

Brown, J. D. and Hudson, T. D. (1998). The alternatives in language assessment. *TESOL Quarterly* 32: 653–75. https://doi.org/10.2307/3587999.

Buck, G. (2001). *Assessing Listening*. Cambridge, UK: Cambridge University Press.

Chalhoub-Deville, M. (ed.). (1999). *Issues in Computer-Adaptive Testing of Reading Proficiency*. Cambridge, UK: Cambridge University Press.

Constantinou, F., Crisp, V. and Johnson, M. (2018). Multiple voices in tests: Toward a macro theory of test writing. *Cambridge Journal of Education* 48: 411–26. https://doi.org/10.1080/0305764X.2017.1337723.

Cronbach, L. J. (1970). Book review. [Review of the book on *The Theory of Achievement Test Items*, by J. R. Bormuth] *Psychometrika* 35: 509–11.

Davidson, F. and Lynch, B. K. (2002). *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. New Haven, CT: Yale University Press.

Douglas, D. (2000). *Assessing Language for Specific Purposes*. Cambridge, UK: Cambridge University Press.

Downing, S. M. and Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education* 10: 61–82. https://doi.org/10.1207/s15324818ame1001_4.

Downing, S. M. and Haladyna, T. M. (eds.). (2006). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Elwood, J. and Murphy, P. (2015). Assessment systems as cultural scripts: A sociocultural theoretical lens on assessment practice and products. *Assessment in Education: Principles, Policy & Practice* 22: 182–92. https://doi.org/10.1080/0969594X.2015.1021568.

Fulcher, G. (2000). The "communicative" legacy in language testing. *System* 28: 483–97. https://doi.org/10.1016/S0346-251X(00)00033-6.

———. (2003). *Testing Second Language Speaking*. London, UK: Longman.

———. (2009). Test use and political philosophy. *Annual Review of Applied Linguistics* 29: 3–20. https://doi.org/10.1017/S0267190509090023.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London, UK: Routledge.

———. (2009). Test architecture: Test retrofit. *Language Testing* 26: 123–44. https://doi.org/10.1177/0265532208097339.

Green, A. and Hawkey, R. (2011). Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing* 29: 109–29. https://doi.org/10.1177/0265532211413445.

Haladyna, T. M. and Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education* 1: 51–78. https://doi.org/10.1207/s15324818ame0201_4.

Haladyna, T. M., Downing, S. M. and Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education* 15: 309–34. https://doi.org/10.1207/S15324818AME1503_5.

Hamp-Lyons, L. (1997). Ethics in language testing. In C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education, Vol. 7. Language Testing and Assessment*. Dordrecht, Netherlands: Kluwer Academic Publishers, 323–33.

———. (2000). Social, professional and individual responsibility in language testing. *System* 28: 579–91. https://doi.org/10.1016/S0346-251X(00)00039-7.

Hamp-Lyons, L. and Lynch, B. K. (1998). Perspectives on validity: A historical analysis of language testing conference abstracts. In A. J. Kunnan (ed.), *Validation in Language Assessment: Selected Papers from the 17th Language Testing Research Colloquium, Long Beach*. Mahwah, NJ: Lawrence Erlbaum, 253–76.

Harsch, C. (2018). Developing test specifications for listening. In J. I. Liontas, T. International Association and M. DelliCarpini (eds.), *The TESOL Encyclopedia of English Language Teaching*. https://doi.org/10.1002/9781118784235.eelt0619.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge, UK: Cambridge University Press.

Ingram, K. (2008). The Cambridge ESOL approach to item writer training: The case of ICFE listening. *Research Note* 32: 5–9.

Johnson, M., Constantinou, F. and Crisp, V. (2017). How do question writers compose external examination questions? Question writing as a socio-cognitive process. *British Educational Research Journal* 43: 700–19. https://doi.org/10.1002/berj.3281.

Kim, J. (2008). *Development and Validation of an ESL Diagnostic Reading-to-Write Test: An Effect-Driven Approach*. PhD thesis, University of Illinois at Urbana-Champaign.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H. and Roullion, V. (2010). A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writer's characteristics. *Language Assessment Quarterly* 7: 160–74. https://doi.org/10.1080/15434300903473989.

Kohn, A. (2000). *The Case Against Standardized Testing: Raising the Scores, Ruining the Schools*. Portsmouth, NH: Heinemann.

Luoma, S. (2004). *Assessing Speaking*. Cambridge, UK: Cambridge University Press.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing* 18: 351–72. https://doi.org/10.1177/026553220101800403.

Lynch, B. K. and Shaw, P. (2005). Portfolio, power, and ethics. *TESOL Quarterly* 39: 263–97. https://doi.org/10.2307/3588311.

McNamara, T. (2000). *Language Testing*. Oxford, UK: Oxford University Press.

———. (2008). The socio-political and power dimensions of tests. In E. Shohamy and N. H. Hornberger (eds.), *Encyclopedia of Language and Education, Vol 7. Language Testing and Assessment*, 2nd edn. New York, NY: Springer, 415–27.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. London, UK: Blackwell Publishing.

Mislevy, R. J., Steinberg, L. S. and Almond, R. G. (2002). Design and analysis in task-based language assessment. *Language Testing* 19, 477–96. https://doi.org/10.1191/0265532202lt241oa.

——— (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives* 1: 3–62. https://doi.org/10.1207/S15366359MEA0101_02.

Morgan, B. (2007). Poststructuralism and applied linguistics: Complementary approaches to identity and culture in ELT. In J. Cummins and C. Davison (eds.), *Springer International Handbook of Education, Vol. 15. International Handbook of English Language Teaching*. Norwell, MA: Springer, 949–68.

Ngo, X. M. (2016, October). *Demystifying Item Writing: The Need for a Theoretical Framework*. Paper presented at the 4th British Council New Directions in English Language Assessment, Conference, Hanoi, Vietnam.

Norton, B. (2000). *Identity and Language Learning: Gender, Ethnicity, and Educational Change*. London, UK: Longman.

Oller, J. W. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal* 56: 151–58. https://doi.org/10.2307/324037.

Peirce, B. N. (1992). Demystifying the TOEFL reading test. *TESOL Quarterly* 26: 665–91. https://doi.org/10.2307/3586868.

Pennycook, A. (2001). *Critical Applied Linguistics: A Critical Approach*. Mahwah, NJ: Lawrence Erlbaum.

Pugh, D., De Champlain, A., Gierl, M., Lai, H. and Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *RPTEL* 15. https://doi.org/10.1186/s41039-020-00134-8.

Purpura, J. E. (2004). *Assessing Grammar*. Cambridge, UK: Cambridge University Press.

Read, J. (2000). *Assessing Vocabulary*. Cambridge, UK: Cambridge University Press.

Roid, G. H. and Haladyna, T. M. (1982). *A Technology for Test-item Writing*. New York, NY: Academic Press.

Ryan, È. and Brunfaut, T. (2016). When the test developer does not speak the target language: The use of language informants in the test development process. *Language Assessment Quarterly* 13: 393–408. https://doi.org/10.1080/15434303.2016.1236110.

Salisbury, K. (2005). *The Edge of Expertise: Towards an Understanding of Listening Test Item Writing as Professional Practice*. PhD thesis, Kings College London.

Shin, D. and Kim, N. (2010). Munhang Kaebal Kwajungeseo Shihumjaksung Sepukyehwukseo YukwhalyeonKu [Investigating the role of test specifications in item development process]. *Foreign Languages Education* 17: 257–79.

Shizuka, T., Takeuchi, O., Yashima, T. and Yoshizawa, K. (2006). A comparison of three- and four-option English tests for university entrance selection purposes in Japan. *Language Testing* 23: 35–57. https://doi.org/10.1191/0265532206lt319oa.

Shohamy, E. (2001). *The Power of Tests: A Critical Perspective on the Uses and Consequences of Language Tests*. London, UK: Longman.

Spaan, M. (2007). Evolution of a test item. *Language Assessment Quarterly* 4: 279–93. https://doi.org/10.1080/15434300701462937.

———. (2006). Test and item specifications development. *Language Assessment Quarterly* 3: 71–79. https://doi.org/10.1207/s15434311laq0301_5.

Spolsky, B. (1995). *Measured Words*. Oxford, UK: Oxford University Press.

Stubbs, J. B. and Tucker, G. R. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal* 58: 239–41. https://doi.org/10.2307/325020.

Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly* 8: 34–52. https://doi.org/10.1080/15434303.2010.536924.

Tunks, J. (2001). The effect of training in test item writing on test performance of junior high students. *Educational Studies* 27: 129–42. https://doi.org/10.1080/03055690120050374.

Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language test specification reverse engineering. *Language Assessment Quarterly* 10: 317–42. https://doi.org/10.1080/15434303.2010.516042.

Weigle, S. W. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing* 24: 539–72. https://doi.org/10.1177/0265532207080780.

Zandi, H., Kaivanpanah, S. and Alavi, S. (2014). The effect of test specifications review on improving the quality of a test. *Iranian Journal of Language Teaching Research* 2: 1–14.

21
Writing integrated tasks

Lia Plakans

Introduction

In developing assessments that reflect language use in context, integration of the four skills has become an inviting direction for task design. Treating language as holistic or as an amalgamation of reading, writing, listening, and speaking, rather than as a set of isolated abilities, may allow for different and more substantive interpretations of performances that are transferrable to real-world domains. While integrating skills presents challenges for test developers and users, such tasks provide measures that may overcome some gaps left by independent skills assessment as well as introducing new directions for assessing language (Knoch and Sitajalabhorn, 2013; Plakans, 2008; Read, 1990; Weigle, 2004; Yang, 2014).

The term *integrated* has been applied fairly broadly to tasks that require more than one skill for completion. For example, an integrated task could entail reading a short text on a topic followed by writing a summary, or it may involve listening to a conversation and commenting on the speakers' opinions. Knoch and Sitajalabhorn (2013) contended that integrated writing tasks should (1) include source material that is linguistic (i.e., not graphs or images), (2) require the source material be transformed into the written performance, (3) be rated on a scale that includes skill integration, and (4) be grounded in actual performances on an integrated task. The following example illustrates an integrated writing task for an advanced level English language placement test.

> Reading-to-write task
> Read the following passage on an issue in education. Then consider your position on this topic. *Write an essay agreeing* or *disagreeing with the author*. Include the main points from the article in your essay and your opinion on each point. Do not copy exact phases from the article. Your essay will be evaluated on the following:
>
> - explanation of key ideas from the reading
> - organization
> - development of ideas
> - clarity in grammar and vocabulary use

DOI: 10.4324/9781003220756-28

Evaluating teacher quality

Sometimes it seems that school reform is an annual event. Politicians and administrators continually seek to improve education through reform. One topic that appears in this area is teacher quality: in other words, how to make sure our children have the best teachers leading them in successful learning. A proposal suggested recently is to put parents in charge of judging teacher quality. This idea has several merits.

Parents, it may be argued, know what is best for their children. They see their young ones every day and can observe their learning and improvement. Parents know when their kids are excited about school and when they reach certain milestones in learning. It is also true that parents have the greatest stake in teacher quality. This should lead them to evaluate teachers with high motivation for strong results – school improvement. Unlike politicians, parents have their kids' best interests in mind, which should transfer into good assessment of teacher quality.

While this proposal might be controversial, it simply brings evaluation of teacher quality to the people who are impacted most, parents and children. Through their eyes, we might find new ideas and better ways to address teacher quality.

Integrated assessment is not an entirely new idea but one that has regained momentum with current language teaching approaches, such as task-based language teaching (TBLT) and content and language integrated learning (CLIL), which view language holistically or functionally, rather than as four separate skills. Outside the classroom, large-scale language assessments include integrated assessment tasks with independent tasks to better represent the constructs of speaking and writing. While benefits exist for integrating skills in assessment, numerous challenges surface that warrant attention when developing and using such test tasks. This chapter will first discuss past work in integrative assessment, a precursor to integrated skills assessment, and then focus on integrated task research, leading into issues and recommended practice in writing integrated tasks.

Historical perspectives

Conceptually, the argument for assessing language as an integrated whole has a fairly extended history. The definition of language as four skills can be found in debates from the 1960s regarding testing language competence and use versus language knowledge (Cooper, 1965). Evolving from this debate, several hypotheses related to skill integration took shape: (1) language can be defined as an overarching ability, rather than as separate components; (2) skill integration is important for authentic language use; and (3) skills can be combined because they share processes. These perspectives could be found in the language testing literature since the 1960s and are very much alive today in arguments for integrated skills assessment.

The first hypothesis operationalized the tenet that one could measure global language proficiency, rather than dividing it into components. This notion was written about extensively by Oller (1979) and termed "integrative." Research by Oller (1983) and others supported the hypothesis that language proficiency was a unitary construct because all components or measures of language components share variance, and this global proficiency or unity construct was called "expectancy grammar." He claimed, "There are two kinds of empirical data which strongly suggest the possibility that all language skills are based on a rather unitary proficiency factor or internalized grammar" (Oller, 1983: 3). Examples of integrative

test items include the cloze test, a kind of gap-filling item, and dictation tests. In relating expectancy grammar to integrated skills, it is important to note that assessing general language ability through "integrative tests" emerged from a response to discrete item testing of vocabulary and grammar (Farhady, 1979; Oller, 1983; Lewkowicz, 1997) rather than to encourage skill integration. The idea of expectancy grammar eventually met with diminished enthusiasm because of questions regarding the item types, such as the cloze, and critiques of the analysis used to support it (Chalhoub-Deville and Deville, 2005; Farhady, 1983); nevertheless, the contention that language should be assessed as a whole rather than as separated parts has contributed to the field of language testing and to integrated skills assessment.

The second hypothesis, rooted in authenticity, focused on including integration in assessment because it occurs in real language use situations. In presenting fundamentals in language testing, Carroll (1961) recognized the usefulness of assessing areas of language knowledge, such as knowledge of structure, knowledge of general-use lexicon, auditory discrimination, oral production, phonemic reading, and writing (spelling). However, he advocated for integrated testing of rate and accuracy in four domains: listening comprehension, interactive speaking, reading comprehension, and writing comprehension. The rationale for integration in these areas was to facilitate transference to real language use, which Carroll (1961) articulated in the following:

> [A]n ideal English language proficiency test should make it possible to differentiate, to the greatest possible extent, levels of performance in those dimensions of performance which are relevant to the kinds of situations in which the examinees will find themselves after being selected on the basis of the test.
>
> *(Carroll, 1961: 319)*

A third perspective on integration of skills appears somewhat later and may be attributed to literature in first language (L1) domains such as educational and cognitive psychology (Fitzgerald and Shanahan, 2000), rather than in language testing specifically. This is a view that many processes involved in language skills are shared. In an empirical study, Shanahan and Lomax (1986) proposed three theoretical models of the reading and writing relationship and analyzed L1 data for each. The results of their study supported an interactive model, which holds that writing can influence reading development, and reading can impact writing development.

While these three positions regarding integration of skills were initiated over 40 years ago, they are still relevant today and appear in research seeking to elucidate the construct elicited by integrated tasks (e.g., Cumming, 2013; Grabe and Cui, 2016; Huang *et al*., 2018; Yang and Plakans, 2012; Yang, 2014; Sawaki *et al*., 2009; Sawaki *et al*., 2013).

Research on integrated skills tasks

Research on integrated assessment can be categorized into overlapping themes: comparison of tasks, task characteristics, performances features across score levels, and investigations of construct.

Comparing integrated and independent skills tasks

As interest in integrated assessment tasks increases, administrators, teachers, and language testing researchers have sought to understand the differences and similarities between these

tasks and those that treat skills independently. Studies have compared scores and used correlations to determine if test takers differ in their performance on the two kinds of tasks, confirming that a relationship exists (Brown *et al*., 1991; Gebril, 2009; Lee, 2006; Lewkowicz, 1994; Huang and Hung, 2013; Huang *et al*., 2018). However, when delving more deeply into the features of the products from writing-only and integrated writing tasks, differences have emerged. One area of divergence is in development. Cumming *et al*. (2005) found that independent tasks lead to longer essays. In contrast, Lewkowicz (1994) discovered that while there was no significant difference in word count (length), more points were introduced by writers on integrated tasks than on the independent tasks. Cumming *et al*. (2005) did not find significant differences in grammatical accuracy between the two tasks but did find the independent task essays scored lower on a number of rhetorical features of persuasive writing.

While these studies have focused on comparing the products of independent and integrated assessment tasks, a few studies have looked at how test takers respond in relation to the two types of tasks. For example, Plakans (2008) found that one of the main differences was in planning, with independent tasks requiring more pre-writing planning, and the integrated tasks leading to more planning while writing. A pair of studies of independent and integrated (speaking) tasks investigated topic knowledge, and anxiety impacted performance differentially (Huang and Hung, 2013; Huang *et al*., 2018). The researchers found that for one topic, related to business, anxiety had comparable impact on both task types, while on the science-related topic, anxiety had a greater impact on integrated task performance. Huang and Hung suggest that introducing reading or listening on a topic that a test taker is not familiar with might result in anxiety about comprehending the source material in addition to anxiety regarding performance.

Despite the varied results in comparing integrated and independent tasks, when test takers have been asked their opinions, a preference for integration is resounding (Gebril and Plakans, 2009; Huang and Hung, 2013). Stakeholders report feeling more knowledgeable on the topic after reading or listening to content as well as finding the source texts useful for ideas, vocabulary, and organization.

Task characteristics

While a good deal of variation exists in integrated tasks, a few studies have looked at the differences across them. Researchers have explored the genre of writing, topic/audience, and, most recently, the use of source materials in integrated writing. In a study that included test takers' process and genre, Ascención (2005) compared summary writing to response writing, finding that the latter required more planning and monitoring than summarizing, which could be explained by the need to create one's own voice in response writing. In a study of the Test of English as a Foreign Language (TOEFL) speaking tasks, Swain *et al*. (2009) found differences in strategies used to complete integrated tasks that included two or three skills. They found an increase in skills related to an increase in strategy use, with tasks involving three skills eliciting more strategies than those with two skills and either integrated task revealing more strategies than the independent task. Cho and Choi (2018) explored the impact that came with defining an audience for writers in an integrated task by comparing a task to write for fellow classmates with a task that had no audience specification. Analyzing the writing, the researchers found that summaries differed in giving contextual background, communicating ideas accurately, and providing more complete summaries when writing for an audience.

The topic selected for a task has emerged as an issue for integrated tasks (Huang *et al*., 2018). Cumming *et al*. (2005) discovered differences in source text use between two task topics, one regarding political science and the other on cinema. Lee and Anderson's (2007) study investigated the interplay in integrated writing assessment performance between task topic, test taker academic background, and general English language proficiency. The results showed that the model best fit for the data entailed main effects for topic and language proficiency. Despite the judgments of equivalency in difficulty, complexity, and cognitive demands, topic still impacted test takers' scores. This struggle has existed with independent skills tasks as well, but given the potential for compounding background knowledge with reading or listening to source texts, integrated tasks may require additional care in topic selection.

Related to the issue of topic, the texts used in integrated tasks have generated increased interest recently. An early study on this topic (Yu, 2009) investigated source text characteristics in summary writing and found that they had a greater impact on score than writers' proficiency. He compared the writers' use of three source texts of different domains and genres: an expository piece on education, a narrative about rivers, and an argumentative text on work. The texts also differed in other features, such as having an introduction or no introduction and clear timeline or no detailed time frame. While the source texts significantly affected the summary task score, it is not clear which features differentiated the texts' impact or if a combination of certain features was affecting test takers. However, in a follow-up questionnaire and interview, the test takers reported that features such as organization, unknown words, length, and topic impacted their summarizing of a text. They felt that the education-related expository source text was easiest to summarize because of its clear organization and the low number of unfamiliar words. Similarly, Li (2014) analyzed two summary tasks: one on a narrative text and the other expository. The results show that it was more difficult to receive a high score on a summary written about the narrative text than a summary of the expository text. In contrast to the score analysis, test takers' perceptions were that the expository text was more difficult to summarize.

Delving more deeply into different text structures, Cho *et al*. (2013) studied the effect of variance in the source materials presented (listening and reading texts) related to how the texts presented content and how that content was organized. Their analysis consisted of two stages: first, a multilevel modeling of the quantitative data, and second, qualitative analysis of the prompts based on the modeling results. After language English ability, two source text characteristics that were significant in explaining the variation in score: passage difficulty and lecture distinctness. These studies are identifying the complexities in developing integrated tasks, and continued work in this area will help qualify the characteristics of task type, topic, and texts used in integrated assessment.

Performance features

Research on integrated assessment tasks has considered the interplay between performance and proficiency level. Attention to general second language or writing proficiency has generated studies comparing essays across test scores for certain language features. Consistent with second language writing research, higher-scoring integrated writing is longer, based on the metric of word count (Cumming *et al*., 2005; Gebril and Plakans, 2009; Plakans *et al*., 2019). Stronger integrated writing also has more accurate grammar and greater integration of sources than writing at lower levels. Interestingly, in a study by Gebril and Plakans (2009), closer comparison across three levels of scores revealed that linguistic features

such as grammatical accuracy were stronger at differentiating between the lowest-scoring group and the two higher-scoring groups than between the higher groups. In a similar study with a reading-listening-writing task, Plakans *et al*. (2019) found similar patterns in the contribution of fluency, accuracy, and complexity to integrated writing scores. Fluency was the strongest predictor, with accuracy the second; complexity was the least predictive. Baba (2009) investigated how test takers' lexical proficiency impacted their scores on summary writing to explore whether general language proficiency or L2 lexical proficiency had an impact. The impact of lexical proficiency on summary writing score was weak compared to that of language proficiency, with receptive vocabulary indicators of size and depth having little impact and productive indicators of lexical density. These results may indicate that integrated tasks tap test takers' ability to manipulate language.

The issues of source use and score level have appeared in a number of studies (Cumming *et* al., 2005; Gebril and Plakans, 2009; Plakans and Gebril, 2013). These studies on language proficiency and score in integrated tasks are contributing to a picture of test taker performance at different levels and may, eventually, impact the design of rating scales as well as the interpretations made from the scores.

Seeking the construct

Given the importance of defining a construct for test development, it is not surprising that scholars have undertaken studies targeting the underlying abilities involved in integrated assessment. One question being pursued is whether the resulting score from an integrated task is a measure of *all* the skills that are integrated. Sawaki *et al*. (2009) analyzed the internet-based TOEFL products to inform score interpretation. They found that scores on the integrated speaking and writing tasks were only slightly related to the less dominant skills in the tasks (listening, reading), and thus scores were better interpreted as measures of writing or speaking. The skill of reading in reading-writing tasks has been studied from several angles. Two researchers correlated integrated reading-writing scores with separate measures of reading and writing and, in both cases, determined that the variance in scores was determined more by writing and minimally explained by reading (Delaney, 2008). In fact, Watanabe (2001) suggested that the variance attributed to reading may be related to general language proficiency. However, studies that have investigated the processes of writers during reading and writing tasks have found that reading is an important process in completing the assessment (Esmaeili, 2002; Plakans, 2009a, Plakans and Gebril, 2012).

Another approach to defining the construct in integrated tasks investigates the abilities that underlie integration and the synthesis of the skills, rather than individual skills added together. In a study to investigate what processes were shared across reading and writing, researchers developed an integrated task that was iterative, involving multiple stages of integration (Plakans *et al.,* 2019). The processes that they identified as shared between reading and writing included (1) focusing at the word level, (2) drawing on background knowledge, (3) monitoring for comprehension, (4) rereading, and (5) summarizing. Ascención (2005) and Plakans (2009b) investigated the concept of "discourse synthesis" in second language integrated writing tasks using think-aloud verbal protocols. *Discourse synthesis* is a term used in English-first language writing theory to describe the constructive meaning-making process of reading for writing that includes the sub-processes of organizing, selecting, and connecting (Spivey, 1990). In both studies, the results of recording writers' online processes revealed the sub-processes of discourse synthesis did occur in composing integrated writing tasks. These sub-processes also appeared in a writing strategy survey used by Yang and

Plakans (2012) to investigate reading-listening-writing tasks: the study found that higher-scoring writers utilized organizing, selecting, and connecting strategies in their test-taking process. While the research so far has been ambitiously trying to define the construct of integrated assessment, more work is needed to identify and clarify the underlying abilities involved in the various kinds of integrated tasks and how scores reflect this.

Critical issues and guidelines for practice

In writing integrated tasks, several steps are required, most of which align with good test development practice in general; however, given their constructs and the complex processes they elicit, such tasks might require extra care. Several issues require consideration when writing or using integrated tasks in language assessment, including defining what is being measured; considering appropriate and feasible input/output dimensions; choosing task type and developing tasks; carefully constructing instructions; establishing scoring criteria and rating materials; and piloting repeatedly to test out the input, instructions, timing, and rating (see also Kremmel *et al*., this volume).

Needs analysis and defining the construct

In deciding whether to use integrated skills in assessment, the purpose of the test should be the first consideration, along with a needs analysis, to determine what the domain for language use entails. If a test will be used to identify language ability in contexts that require skill integration, then such tasks should appear in a test. However, the nature of the integration deserves careful consideration in the development phase. For example, how many texts do language users manage when integrating reading with writing or listening with speaking? What are the nature, length, and difficulty level of listening or reading texts? At what points in the process do the language users integrate skills? How fully are they integrated? Needs analysis at the beginning of the development process should attend closely to the nuances of integrated language use. Since a major argument for using such tasks is their authenticity, their fidelity claim needs verification.

Along with a needs analysis, test developers should consider what abilities underlie the integrated tasks being developed. Defining the construct is a major challenge for integrated assessment as it is less clearly definable than separated skills testing. It may be that the constructs for these tasks are not just a sum of skills, but synthesized matrices. Once the tasks have been developed, data may need to be collected for validation to defend, further define, or discard the construct. Continual scrutiny of construct validity as well as updated needs analyses should be a regular part of any testing program using integrated skills tasks.

Considering task type

Given the broad definition of integrated assessment, many tasks reside under this umbrella, with room for further innovation. This variety makes integrating skills promising and, at the same time, complex (Leki *et al*., 2008). The choice of integrated task type should be determined by the needs analysis and construct definition.

A common type of integration appears with content- or "text-responsible" tasks (Leki and Carson, 1997: 41), in which the test taker reads and/or listens to a text, followed by speaking or writing a response that includes some of this content. A summary task is a common integrated task that would fall into this task type; related to summaries are review

tasks, in which the test taker summarizes the text and adds his/her commentary on the topic. An example of a text-responsible task appears at the B2 level of the DELF French language proficiency test administered by Centre International D'étude Pédagogiques. To assess speaking, test takers read a short passage that is meant to create a reaction, and then they defend their reaction to or opinion of the text in an extended spoken response. In content-responsible tasks, test takers may be asked to comprehend and negotiate more than one text. Another example of an integrated task using multiple texts provides the test taker with two opinions on an issue and requires a persuasive response that is supported by the content (see example task that follows).

> Example of a multiple text content-responsible task
>
> Instructions: Read the two passages below presenting different opinions on children watching television. Then consider your viewpoint on this topic and write an essay giving reasons for your opinion. Use the readings appropriately to support your position.
>
> The danger of television for children
>
> Recent research has indicated that watching more than two hours a day of television can affect children's cognition and potentially lead to an increased chance of having attention deficit disorder. The way TV programs are designed for children, to be rapidly changing and exciting, may cause problems as they get older in terms of focusing on homework or extended tasks. Parents need to restrict the time that children watch as much as they attend to what is being watched. TV is not a babysitter.
>
> The healthy impact of children's television
>
> While there may be issues with watching extended hours of television, children can learn valuable lessons from educational programs. Recently, several programs, such as *Sesame Street*, have included healthy eating and exercise lessons to remind kids to get physical and not eat too many cookies. This kind of programming can have a great impact on what kids think about healthy choices, and open up this topic for discussion with their parents. If we lose this medium for influencing kids, they may not learn these valuable lessons.

Another task described as integrated is one in which input texts are provided for test takers but not required in the response. The content in the texts serves as an illustration or stimulus for the task topic. While the resulting score from this kind of task does not necessarily relate to the integration of skills, test takers will utilize more than one skill in their process. The example that follows shows an integrated stimulus task that provides the test taker with a conversation that introduces the topic for a short spoken response.

> Example of a listening/speaking task
>
> Instructions: Press the play button to listen to a conversation between two people waiting to buy a movie ticket. You will then have 30 seconds to plan a short response for the following: Recommend a movie that you like to these people, and explain why you suggest it. You will have 60 seconds to respond.

Audio transcript

Person 1: Which movie should we see? Do you feel like a comedy or something scary?

Person 2: Hmmmm. None of these looks very good to me. What do you think?

Person 1: Well, I know this movie is for kids, and it's all animation. I don't want to see that. And the other three are too serious. I'd really like something more upbeat, like that movie we saw last week with the two travellers who kept having funny adventures. The characters were hilarious and the special effects, amazing.

Person 2: Yeah, or I liked your idea about something scary too. I don't really care what kind of movie it is, as long as it's good. I think we should find a movie to watch online or rent something instead of watching any of these. It would be cheaper for sure.

Choosing texts

The role of texts and topics in integrated tasks is a critical issue that exists in language tasks generally but has an elevated importance in integrating skills. The possible variety of topics in texts and performances in integrated assessment leads to several concerns: different constructs elicited by different texts, the interaction of previous experience or background knowledge, and the magnification of topic effect. While only a limited amount of research has looked at these issues, good practices in writing integrated tasks can be outlined.

The first step is choosing a topic. This decision could be guided by a needs analysis. If specific content areas are likely to be familiar to students, these topics are best. In addition, choosing texts in test takers' content area may allow for better assessment of integration in their field. For example, in integrated speaking and writing tasks at the C1 and C2 levels of *Diplôme approfondi de langue française* (DALF), test takers may choose between two areas: humanities and social science or science. However, often information about test takers' domain of interest or expertise is not available to test developers, in which case more general topics may be selected. Once topics are chosen, texts can be developed or collected. Test developers may choose to find published texts on chosen topics but should carefully pilot these texts and consider revising for difficult structures and challenging vocabulary. There may also be issues with copyrighted information, and permissions should be sought on authentic texts used. The alternative is to develop new texts.

Once texts have been selected or developed, the level of difficulty should be evaluated. While not a perfect measure for difficulty, readability indices can provide some information regarding the grade level or difficulty level of a text based on word length, sentence length, and some other structural features. Online programs measure the readability of a text, and Microsoft Word includes this feature in the grammar/spellchecker. A more in-depth analysis can be conducted for lexical density or coherence of texts using online programs such as CLAN (http://dali.talkbank.org/clan/) or CohMetrix (cohmetrix.com) (see also Cushing, this volume). If texts are found to be too difficult, they should be revised, although modification is not a clear direct path, as reading research has shown (Koda, 2004). Some features to modify include sentence length, unfamiliar or technical words, unclear subjects, and main verbs or adding cohesion devices such as transitions or topic sentences.

Composing instruction

Integrated tasks can be confusing for test takers if they have not experienced them before, and they may resort to interpreting the task in the same way as independent tasks. A number of studies have investigated task representation in reading-to-write tasks, providing valuable insights into the nature of understanding the demands of these tasks. This research has revealed that task representation can be impacted by writers' experiences with prior writing tasks (Plakans, 2010; Ruiz-Funes, 2001; Wolfersberger, 2007). Obviously, test developers cannot control test takers' prior experience; however, carefully crafted instructions can, perhaps, minimize misinterpretation of the instructions. While this maxim is true of any assessment, one element specific to integrated task instruction demands heightened attention: use of source materials.

Mention of the source texts should appear in at least three parts of integrated task instructions. First of all, the intended purpose of the input texts should be explicit. The test takers need to know if they are responsible for the content of texts in their responses. Second, some mention regarding direct copying of phrases from the text and use of citation should be included, although this instruction is a greater issue with written than oral responses. While the wording for such instruction may be difficult, it can help minimize major problems for rating and interpreting the scores from the task. Lastly, test takers should be given the criteria on which they will be evaluated: not in detail, but to clarify what is important in their performance. If the sources are to be included in the products, then this integration should appear in the evaluation criteria. The following is an example of instructions for an integrated task (Gebril and Plakans, 2009):

> Read the question below, and then read the two passages to get more information about the topic. Write an essay on the topic giving your opinion in response to the question. Typically, an effective response will contain a minimum of 300 words. Your writing will be scored based on how well:
>
> your **ideas** are explained
> **the readings support your argument**
> **you organize your essay**
> **you** choose words
> **you** use grammar
> **you** spell **and** punctuate
>
> **IMPORTANT! READ CAREFULLY BEFORE WORKING ON THE TASK**
>
> - The two passages should help you get some ideas about the topic.
> - You may go back to the passages to check information while writing.
> - You can borrow ideas and examples from the text. However, you should mention the author's name if you do so.
> - Also, if you take exactly the same phrases or sentences mentioned in the passage, put them between two inverted commas (“”).

Developing scoring rubrics

The rubric to score integrated tasks depends on the purpose of the task, the construct as it has been defined, and the constraints of the scoring sessions. Test developers should determine if the resulting scores are a sum of separate scores on different criteria, an analytic rubric,

or one single score that represents the overall level of the performance – a holistic rubric. Research has determined that integrated tasks can be rated reliably (Gebril, 2009; Ohta *et al*., 2018; Wang *et al*., 2017); however, raters may find these tasks more challenging, particularly with content-responsible writing tasks, as they need to evaluate the quality of source integration (Gebril and Plakans, 2014). The example that follows illustrates the descriptor from a holistic rubric (Gebril and Plakans, 2009).

Level 4

A response:

- is good in clearly and correctly presenting ideas in relation to the relevant information in the reading texts. It may have weaknesses in connection to specific points made in the readings.
- uses clear organization and logical development.
- includes some minor language errors which occasionally create slight confusion, but not enough to lose the writer's meaning.

In addition to determining the form of the scale, the criteria and descriptors need development. Several approaches to this may be undertaken. First of all, a construct can help identify criteria for the scale; for example, if the construct includes reading comprehension, then the scale should have some wording regarding this skill. Another source of scale criteria can be found in the needs analysis. By scrutinizing the abilities required in a language use context, test developers may determine what criteria lead to successful integration of skills. A third method is to develop a data-driven scale. In this case, the test task is used with the target population to collect performance samples (see Galaczi and Lim, this volume). These samples are then divided into the score levels needed and analyzed for language features. The features then become criteria and descriptors for the scale.

Research on integrated assessment scores has revealed an interesting challenge for future scale development. Studies have suggested the possibility that the features distinguishing high-level performances may be dissimilar from those distinguishing lower-level performances (Gebril and Plakans, 2009; Sawaki *et al*., 2009). In order to distinguish performances at all levels, it may be important to have different criteria across score levels.

Future directions

Certainly, the current interest in integrated tasks can spark new and innovative task types to support test takers' synthesis of skills. Along with this creative development will come better understanding of the constructs these tasks are eliciting. Several areas are beacons for future work. First of all, more investigation and advancement of constructs for integrated skill assessment holds promise. A tendency exists to conceptualize integrated skills as a summation of the skills; however, the benefit of these tasks may lie in recognizing higher-order processes that are measured by amalgamated skills or in defining the synthesis of skills as a unique ability. From such explorations, new tasks will evolve, which is another area for future endeavor. In recent integrated tasks, writing and speaking have remained the dominant skills, supported by listening and/or reading. However, tasks that require writing or speaking before reading or listening may have a good deal to offer in terms of integration for comprehension and scaffolding prior knowledge (Plakans *et al.,* 2018 Trites and McGroaty, 2005). These two areas, construct refinement and task innovation, require continued research to inform, support, and extend new ideas about the evolving concept of integrated skills assessment.

Further reading

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability study. *Assessing Writing* 15: 100–17.
Using multivariate generalizability analysis, Gebril investigates reliability between independent and integrated tasks by comparing raters and task facets with the two tasks combined and with each task separately. His findings uncover similar results in reliability for the two combined and as separate tests, which provides evidence that the tasks are measuring the same construct. The results also exhibit similar results when different raters scored the two tasks and when the same raters scored the tasks. Gebril advocates using independent and integrated tasks in combination when assessing writing to provide comprehensive information on second language writing ability.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly* 10: 1–8.
This editorial introduced a special issue on integrated writing task for assessment anchored around the "promises" and "perils" he predicts as a leading scholar in the field. The promises that he sees in these tasks, which emerged in the special issue, are that the tasks (1) elicit realistic and complex literacy engagement; (2) require test takers to draw on sources; (3) are performance based, potentially avoiding a method effect and providing diagnostic or instructional use; and (4) align with construction-integration or multiliteracies views of literacy. The perils substantiated in the research of the special issue are also detailed. The construct issue emerges, which confounds writing with reading and skills required in using source materials. These tasks also are seen to require a threshold level of ability in English and may not fall on a clearly linear progression of ability levels. In addition, he finds the genres for authentic source-based writing are not well defined, especially in terms of scoring. Related to the practical issues of scoring, integrated writing can elicit texts that blur the distinction between a test taker's writing and that of the source materials. His concluding thoughts are that integrated tasks should continue to be developed and provide opportunities to assess literacy in more depth than independent writing; however, ongoing research is necessary to continue with the complexity introduced by this multi-skill, multi-text writing task.

Wang, J., Engelhard, G., Raczynski, K., Song, T. and Wolfe, E. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach, *Assessing Writing* 33: 36–47.
This study focused on source-based writing and rating using a mixed-methods design, which first quantitatively identified difficult-to-score essays, then qualitatively analyzed survey data from raters to explain the difficulty. There were four essay features mentioned consistently by raters when marking a difficult essay: (1) focus of the essay, (2) textual borrowing from source materials, (3) original development of ideas, and (4) essay organization. For these features in hard-to-score essays, raters and experts showed a lack of agreement and overall inconsistency when scoring. The study implications advocate care in preparing rater training materials to include effective prototype essays for training and selective materials to emphasize features that make rating difficult. Furthermore, rater retraining, especially with changes in source materials, is seen as important to maintain reliable scoring for integrated tasks.

References

Ascención, Y. (2005). *Validation of Reading-to-write Assessment Tasks Performed by Second Language Learners*. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing* 18: 191–208. https://doi.org/10.1016/j.jslw.2009.05.003.

Brown, J. D., Hilgers, T. and Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Written Communication* 8: 533–56. https://doi.org/10.1177/0741088391008004005.

Carroll, J. M. (1961). *Fundamental Considerations in Teaching for English Language Proficiency of Foreign Students*. Washington, DC: Routledge.

Chalhoub-Deville, M. and Deville, C. (2005). Looking back at and forward to what language testers measure. In E. Hinkel (ed.), *Handbook of Research on Second Language Teaching and Learning*. Mahwah, NJ: Erlbaum.

Cho, Y. and Choi, I. (2018). Writing from sources: Does audience matter? *Assessing Writing* 37: 25–38. https://doi.org/10.1016/j.asw.2018.03.004.

Cho, Y., Rijmen, F. and Novák, J. (2013). Investigating the effect of prompt characteristics on the comparability of TOEFL iBT integrated writing tasks. *Language Testing* 30: 513–34. https://doi.org/10.1177/0265532213478796.

Cooper, R. L. (1965). Testing. In H. B. Allen and R. N. Allen (eds.), *Teaching English as a Second Language: A Book of Readings*. New York, NY: McGraw-Hill.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly* 10: 1–8. https://doi.org/10.1080/15434303.2011.622016.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K. and James, M. (2005). Differences in written discourse in writing-only and reading-to-write prototype tasks for next generation TOEFL. *Assessing Writing* 10: 5–43.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes* 7: 140–50. https://doi.org/10.1016/j.jeap.2008.04.001.

Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. *Canadian Modern Language Journal* 58: 599–622. https://doi.org/10.3138/cmlr.58.4.599.

Farhady, H. (1979). The disjunctive fallacy between discrete point and integrative tests. *TESOL Quarterly* 13: 347–57. https://doi.org/10.2307/3585882.

———. (1983). On the plausibility of the unitary language proficiency factor. In J. W. Oller (ed.), *Issues in Language Testing Research*. Rowley, MA: Newbury House Publishers.

Fitzgrerald, J. and Shanahan, T. (2000). Reading and writing relations and their development. *Educational Psychologist* 35: 39–50. https://doi.org/10.1207/S15326985EP3501_5.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing* 26: 507–31. https://doi.org/10.1177/0265532209340188.

Gebril, A. and Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Working Papers in Second or Foreign Language Assessment* 7: 47–84.

———. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, 21: 56–73. https://doi.org/10.1016/j.asw.2014.03.002.

Grabe, W. and Cui, Z. (2016). Reading-writing relationships in first and second language academic literacy development. *Language Teaching* 49: 339–55. https://doi.org/10.1017/S0261444816000082.

Huang, H. T. and Hung, S. T. (2013). Comparing the effects of test anxiety on independent and integrated speaking performance. *TESOL Quarterly* 47*:* 444–49. https://doi.org/10.1002/tesq.69.

Huang, H. T., Huang, S. T. and Plakans, L. (2018). Topical knowledge and L2 speaking assessment: Comparing independent and integrated speaking assessments. *Language Testing* 35: 27–49. https://doi.org/10.1177/0265532216677106.

Knoch, U. and Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes. *Assessing Writing* 18: 300–8. https://doi.org/10.1016/j.asw.2013.09.003.

Koda, K. (2004). *Insights into Second Language Reading: A Cross-Linguistic Approach*. Cambridge, UK: Cambridge University Press.

Lee, Y. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing* 23: 131–66. https://doi.org/10.1191/0265532206lt325oa.

Lee, H. and Anderson, C. (2007). Validity and topic generality of a writing performance test. *Language Testing* 24: 307–30. https://doi.org/10.1177/0265532207077200.

Leki, I. and Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly* 31: 39–69. https://doi.org/10.2307/3587974.

Leki, I., Cumming, A. and Silva, T. (2008). *A Synthesis of Research on Second Language Writing in English*. New York, NY: Routledge.

Lewkowicz, J. A. (1994). Writing from Sources: Does source material help or hinder students' performance? In N. Bird *et al*. (eds.), *Language and Learning: Papers Presented at the Annual International Language in Education Conference*. Hong Kong: ERIC Document (ED 386 050).

———. (1997). The integrated testing of a second language. In C. Clapham and D. Corson (eds.), *Encyclopaedia of Language and Education*. Dordrecht, The Netherlands: Kluwer.

Li, J. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing* 22: 75–90. https://doi.org/10.1016/j.asw.2014.08.003.

Ohta, R., Plakans, L. and Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scoring: A generalizability analysis. *Assessing Writing* 38: 21–36. https://doi.org/10.1016/j.asw.2018.08.001.

Oller, J. W. (1979). *Language Tests at School: A Pragmatic Approach*. London, UK: Longman.

———. (1983). *Issues in Language Testing Research*. Rowley, MA: Newbury House.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing* 13: 111–29. https://doi.org/10.1016/j.asw.2008.07.001.

———. (2009a). The role of reading strategies in L2 writing tasks. *Journal of English for Academic Purposes* 8: 252–66. https://doi.org/10.1016/j.jeap.2009.05.001.

———. (2009b). Discourse synthesis in integrated second language assessment. *Language Testing* 26: 1–27. https://doi.org/10.1177/0265532209340192.

———. (2010). Independent vs. integrated tasks: A comparison of task representation. *TESOL Quarterly* 44: 185–94. www.jstor.org/stable/27785076.

Plakans, L. and Gebril, A. (2012). A close investigation into source use in integrated writing tasks. *Assessing Writing* 17: 18–34. https://doi.org/10.1016/j.asw.2011.09.002.

———. (2013). Using multiple sources in a listening and reading-based writing assessment task: Source text use as a predictor of score. *Second Language Writing Journal* 22*:* 217–30. https://doi.org/10.1016/j.jslw.2013.02.003.

Plakans, L., Gebril, A. and Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing* 36: 161–79. https://doi.org/10.1177/0265532216669537.

Plakans, L., Liao, J. T. and Wang, F. (2018). Integrated assessment research: Writing-into-reading. *Language Teaching* 51: 430–34. https://doi.org/10.1016/j.asw.2019.03.003.

———. (2019). "I should summarize this whole paragraph": Shared processes of reading and writing in iterative integrated tasks. *Assessing Writing* 40: 14–26.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes* 9: 109–21. https://doi.org/10.1016/0889-4906(90)90002-T.

Ruiz-Funes, M. (2001). Task representation in foreign language reading-to-write. *Foreign Language Annals* 34: 226–34. https://doi.org/10.1111/j.1944-9720.2001.tb02404.x.

Sawaki, Y., Stricker, L. J. and Oranje, A. H. (2009). Factor structure of the TOEFL internet-based test. *Language Testing* 26: 5–30. https://doi.org/10.1177/0265532208097335.

Sawaki, Y., Quinlan, T. and Lee, Y. (2013). Understanding learning strengths and weaknesses: Assessing student performance on an integrated writing task. *Language Assessment Quarterly* 10: 73–95. https://doi.org/10.1080/15434303.2011.633305.

Shanahan, T. and Lomax, R. (1986). An analysis and comparison of theoretical models of the reading-writing relationship. *Journal of Educational Psychology* 78: 116–23. https://doi.org/10.1037/0022-0663.78.2.116.

Spivey, N. (1990). Transforming texts: Constructive processes in reading and writing. *Written Communication* 7: 256–87. https://doi.org/10.1177/0741088390007002004.

Swain, M., Huang, L., Barkaoui, K., Brooks, L. and Lapkin, S. (2009). The speaking section of the TOEFL iBT (SSTiBT): Test-takers' reported strategic behaviors. *TOEFL iBT Research Report*, RR 09–30. Retrieved from www.ets.org/Media/Research/pdf/RR-09-30.pdf.

Trites, L. and McGroaty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing* 22: 174–210. https://doi.org/10.1191/0265532205lt299oa.

Wang, J., Engelhard, G., Raczynski, K., Song, T. and Wolfe, E. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing* 33: 36–47. https://doi.org/10.1016/j.asw.2017.03.003.

Weigle, S. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing* 9: 27–55. https://doi.org/10.1016/j.asw.2004.01.002.

Wolfersberger, M. A. (2007). *Second Language Writing from Sources: An Ethnographic Study of an Argument Essay Task*. Unpublished doctoral dissertation, The University of Auckland, Auckland, New Zealand.

Yang, H. C. (2014). Toward a model of strategies and summary writing performance. *Language Assessment Quarterly* 11: 403–31. https://doi.org/10.1080/15434303.2014.957381.

Yang, H. C. and Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly* 46: 80–103. https://doi.org/10.1002/tesq.6.

Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing* 14: 116–37. https://doi.org/10.1016/j.asw.2009.04.002.

22

Test-taking strategies and task design

Andrew D. Cohen

Introduction and definition of terms

The focus of this chapter builds on a perspective voiced by Davidson (2000), when he warned against *statistical determinism* when validating language assessment measures:

> Any new development in language test statistics should not divert language testers away from the far more difficult and fundamental challenges of making valid assessments. Statistical determinism is a deceptively accessible *Weltanschaung* (world view); it is far too easy to allow items to survive trialing or editing based solely on their "p" or "d" value, just as it is far too easy to judge the work by an oral interview rater based solely on inter-rater reliability correlation coefficients. Statistical evidence of educational quality should be fused with and weighed against evidence that derives from other sources.
>
> *(Davidson, 2000: 615)*

Whereas determining the construct relevance or irrelevance of variance on a language test was traditionally based on psychometrics alone, in recent years, it has become more fashionable to look beyond statistics to help understand test results. One such area of investigation has been that of test-taking strategies. In this chapter, I will take a non-psychometric approach to construct relevance in responding to the following question:

> In what ways might the study of construct-relevant and construct-irrelevant test-taking strategy use inform the design and construction of items and tasks on language assessment measures?

When designing and then constructing items and tasks for the purpose of language assessment, we are likely to make certain assumptions about them. One of them is that responding to the items and tasks correctly relies on a proper mobilization of certain cognitive processes. In line with this assumption, O'Sullivan and Weir (2011) included cognitive validity in their model for conceptualizing test validity and carrying out the validation of a

 DOI: 10.4324/9781003220756-29

measure. They saw the *cognitive validity* of a measure as dependent on the processes that respondents use in responding to language assessment items and tasks. With regard to this type of validity, they asked the following two questions:

1 Are the cognitive processes required to complete the tasks appropriately?
2 Are candidates likely to use the same cognitive processes as they would if performing the task in a "real-world" context?

If the test items and tasks are not measuring what they purport to measure, but rather something else that is "irrelevant" to the desired construct of focus, then there is a validity issue (see Chapelle and Lim, this volume). Ideally, respondents are mobilizing strategies aimed at dealing with the construct being assessed rather than avoiding it. We will first consider strategies, then look at research both from an historical perspective and in terms of current efforts and will consider practical applications as well.

Language learner strategies

The following is a working definition of *language learner strategies*:

> Thoughts and actions, consciously chosen and operationalized by language learners, to assist them in carrying out a multiplicity of tasks from the very onset of learning to the most advanced levels of target-language (TL) performance.
>
> *(Cohen, 2011a: 7)*

In order to carry out tasks on language tests, it is possible that learners will call upon their listening, speaking, reading, and writing skills, as well as the related skills of vocabulary learning, grammar, and translation.

Hence, a listening comprehension item could call for inferencing strategies while listening to a snippet of an interchange. As an example, the listener hears: "Well, if you put it that way, then I guess I do owe you an explanation, but I think you're overreacting here" and needs to indicate whether this is *most likely* part of an exchange between:

a) the boss and an employee
b) an auto mechanic and a patron
c) a husband and his wife
d) a police officer and a suspected felon.

In this case, *c* is the appropriate choice, assuming that the man's tone of voice is sarcastic and slightly condescending. Thus, in responding to language assessment measures, respondents are using both test-taking and language learner strategies.

Since the strategy literature invariably refers to metacognitive, cognitive, and affective strategies, it is important to note in this update of the 2012 chapter that recent research has found that a given strategy may assume a metacognitive, cognitive, affective, or social function from one moment to the next, depending on the nature of the interaction (Cohen and Wang, 2018). In other words, the same strategy (e.g., rehearsing a response before using it in a speaking test) could fluctuate from a metacognitive function when respondents are planning out this strategy, to a cognitive function when they are identifying the language material to include in that response, to a social function when they are being cognizant of

whom they are addressing in terms of age and gender, and to an affective function at the point at which they may encounter some frustration and take action accordingly (e.g., positive self-encouragement). The triggering of the affective function may, in turn, lead back to a metacognitive function (e.g., planning a slightly different response in the speaking task). If responding activates a sequence or cluster of strategies, then any fluctuation of functions could happen across two or more strategies, which makes strategy use even more complex (see Cohen and Wang, 2018).

Test-taking strategies

Test-management strategies

Test-management strategies are strategies for responding meaningfully to test items and tasks. These are the processes consciously selected to assist in producing a correct answer responsibly. They include logistical issues such as weighing the importance of responding to different items or tasks, keeping track of the time, and determining where to look for answers.

Test-management strategies on a reading test could deal with:

- going back and forth between a passage and a given question in order to obtain more information about just what to be looking for
- dealing with multiple-choice options systematically so as to give careful consideration to all the alternative choices and to craft a rationale for why one choice is better than the others
- time-management strategies that ensure sufficient attention to items and tasks, especially those with the highest point values on the test.

As an example, on an essay-writing task, a test-management strategy would be to outline the essay before writing it in order to ensure that it responds effectively to the writing prompt. Assume that the task is to write an essay requesting that the respondent take a stand, pro or con, on an issue such as the following from sample essay topics for the former Test of Written English published by Educational Testing Service:

> Do you agree or disagree with the following statement? Sometimes it is better not to tell the truth. Use specific reasons and details to support your answer.

A test-management strategy could be to prepare arguments in outline form in advance – e.g., taking a position in favor of white lies and listing instances where lying makes sense. Perhaps the essay would start with a justification for not telling the truth all the time as a preamble. In any case, the effective use of test-management strategies would be expected to enhance the writing of the essay.

Test-deviousness strategies

Test-deviousness strategies are used by respondents as a means of avoiding having to reveal their actual knowledge on given test tasks. The original term "test-wiseness" (Allan, 1992) was intended to convey the notion of respondents using knowledge of testing formats and other peripheral information to answer test items without going through the expected

cognitive processes (in this case, involving the selection of numerous language elements). The problem with the term *test-wiseness* is that some researchers have construed the term as referring to something positive since there is clearly a positive aspect to being "wise." But in reality, it is a misnomer since test-deviousness strategies represent construct-irrelevant strategies – assisting respondents in obtaining responses (possibly the correct ones) without engaging their second language (L2) knowledge and performance ability. In such cases, respondents are using strategies to make a cunning effort to circumvent or avoid the expected processes by relying exclusively on such things as clues in the questions and in the options. Note that we cannot slight respondents for trying to be clever. Respondents who fully understand the items may still make use of these strategies as well as respondents who do not understand the items at all. So the burden is with the test constructors to ensure that items are not susceptible to such unintended giveaways.

Test-deviousness strategies used in responding to multiple-choice (M-C) items on a reading comprehension test, for example, might include the following:

- stem-option cues – where matching is possible between the stem and an option
- grammatical cues – where only one alternative matches the stem grammatically
- similar options – where several distractors can be eliminated because they essentially say the same thing
- item giveaway – where another item already gives away the information.

(Allan, 1992)

Presumably, test constructors and language educators administering tests would want to restrict or avoid the possibility of responses to M-C items being based on the use of these construct-irrelevant strategies. With regard to a reading test, it would mean using the process of elimination (i.e., selecting an option even though it is not understood, out of a vague sense that the other options could not be correct), using clues in other items to answer an item under consideration, and selecting the option because it appears to have a word or phrase from the passage in it – possibly a key word.

The following is an example of stem-option matching. Respondents did not have to look in the text for surface matches. They were able to match directly between the stem and the correct alternative:

Question: The increased foreign awareness of Filanthropia has . . .

a) resulted in its relative poverty.
b) led to a tourist bureau investigation.
c) created the main population centers.
d) caused its extreme isolation.

Students associated "foreign" in the stem with "tourist" in option *b*, the correct answer in this case, without understanding the test item (Cohen, 2011a: 313).

A caveat at the outset

The literature related to test-taking strategies is a bit problematic in that researchers sometimes make theoretical distinctions that are no longer current. For example, terms like "memory strategies," "compensatory strategies," and "direct vs. indirect strategies" (Oxford, 1990) were found problematic, and consequently, leading strategy researchers have ceased

to use them in their studies for some decades (see Oxford, 2017). Yet these terms are still being used in some of the test-taking strategy studies described later in this chapter. Not so surprisingly, the results of studies using these terms currently have reduced weight.

Studies with test-taking strategies operationalized at a detailed level would appear most beneficial if the aim is to coach respondents in how to manage their test taking more effectively. Rather than flagging terminological problems and misconceptions in the various studies described briefly here, my preference was to provide a caveat initially and let readers determine for themselves just how effectively the various terms are operationalized.

A second issue is that strategies tend not to be functionally monolithic, but rather may fluctuate in function from one micro-moment to the next. Even though a few researchers (like Youn and Bi, 2019) acknowledge the fluctuation of strategy function, they may still treat strategies in their own studies as if they have just one immutable function, as implied by the label (e.g., metacognitive or cognitive). Actually, the Youn and Bi (2019) study added *pragmatic strategies*, which is an important addition, but likewise they assigned a single function to these specialized strategies – e.g., strategies with a single function of being sociopragmatic or pragmalinguistic. The reality might be that a given strategy (e.g., selecting an appropriate opener to start a conversation) involves both sociopragmatic and pragmalinguistic moments, if in deploying a given apology strategy, the respondent at one moment considers what is socioculturally appropriate for the given situation and at the next considers the language forms to use – in what might be a back-and-forth process. Hence, the traditional labels may not represent the reality. It is for this reason that findings from survey questionnaires and even from verbal reports relying on traditional terminology may be inadequate measures of actual strategy use.

A third issue concerns how precisely strategies are referred to in the various test-taking strategy studies. At times, strategies are referred to in such general terms that they sound more like skills than specific strategies operationalized to deal with those skill areas. Clearly, there is a continuum at play. Of real concern here is the extent to which test-taking strategy research results couched in such general terms are actually of benefit to practicing teachers. A case in point, to be described next, is a study that found that certain questionnaire items loaded on factors referred to as "processing inductively," "structuring," "analyzing," "elaborating," "reduction," "effective," "organizing thoughts," and "restructuring" (Huang, 2016). While these factors may sound impressive, the assignment of factors to individual items in this study is not necessarily logical. It would appear that strategies are complex entities that are consequently difficult to study in a top-down manner using factor analysis.

Historical origins of test-taking strategies

Early studies on test-taking strategies provided useful insights about item design. One such study by Danish researchers used introspective and retrospective verbal reports to study the responses of high school and college EFL students to three M-C test item types embedded in connected text. Students explained the alternatives that they would choose and why they thought the selected alternative was the correct one (Dollerup *et al*., 1982). Their finding that responding to each item required produced an array of strategies and that even erroneous decoding could result in a correct response provided early evidence that verbal report from respondents could help in determining test validity.

Another influential early study dealt with strategy use in responding to random and rational-deletion cloze tests (Homburg and Spaan, 1981). In a preliminary phase, respondents were given four different types of cloze (two of them involving nonsense words instead

of real words), and verbal report data were collected in order to determine their rationale for how they responded. Based on the reports of strategy use and success with different random-deletion item types, a 23-blank rational-deletion cloze was constructed with four item types presumed to require different strategies:

- recognition of parallelism across phrases
- processing within a given sentence
- the use of cues in previous sentences
- the use of cues in subsequent sentences.

It was found that success at understanding the main idea among 39 EFL students at three levels was related to success at completing blanks requiring forward reading. The message to designers of rational-deletion cloze tests would be to include in the instructions the suggestion that respondents be sure to read beyond each blank before attempting to fill the gap, in order to reduce the use of construct-irrelevant strategies.

A series of studies were conducted in the early years to determine the closeness of fit between the tester's presumptions about what was being tested and the response processes that the test takers reported (Cohen, 1984). One such study investigating test method effect in EFL reading by Israeli high school students found differences in strategy use by low- and high-proficiency students (Gordon, 1987). The researcher used four response formats:

- M-C questions in English
- M-C questions in Hebrew
- open-ended questions in English
- open-ended questions in Hebrew.

A subgroup of respondents were asked to verbalize their thoughts while they sought answers to each question. Low-proficiency students were found to process information at the *local* (sentence/word) *level*, without relating isolated bits of information to the whole text. They used individual word-centered strategies like matching alternative words to text, copying words out of the text, word-for-word translation, and formulating global impressions of text content on the basis of key words or isolated lexical items in the text or in the test questions. High-proficiency students were seen to comprehend the text at the *global level* – predicting information accurately in context and using lexical and structural knowledge to cope with linguistic difficulties. As to performance, open-ended questions in the L2 (English) were found to be the most difficult and the best discriminator between the high- and low-proficiency students, since the low-proficiency students had difficulty with them. So this study had two messages for test designers:

1 They should construct reading tasks that challenge low-proficiency students to read text at the global level.
2 If discriminating across proficiency levels is important, then they should select formats that do a better job of discrimination, such as open-ended questions in the L2.

A final early study that deserves mention was one of the few not conducted with English. It looked at strategies used by students taking reading comprehension tests in Hebrew (their first language or L1) and French (their L2) (Nevo, 1989). An M-C format was used to test comprehension of two reading passages each in Hebrew and French by 42 tenth graders.

Students answered five M-C items for each of the four texts (in the language of the text), marked alongside each item their primary and secondary strategies for answering that item (referring to a strategy checklist), responded to open-ended questions following each of the reading passages concerning the respondents' evaluation of the test items, and filled out a more general questionnaire at the end of the test. The study was groundbreaking in that it found it possible, without undue disruption, to get feedback from subjects immediately after they had responded to each item concerning the strategies used to select a response. The message here for test designers is to consider various means for piloting test measures to get a sense of their validity before finalizing them.

Critical issues and topics regarding test-taking strategies

A review of the research literature on test-taking strategies some 14 years ago identified the following topics as related to the validation of the language assessment measures (Cohen, 2006):

- conceptual frameworks for classifying test-taking strategies
- the role that the test takers' language proficiency and other characteristics play in the selection of and success in using test-taking strategies
- the role that the test takers' L1 (or dominant language) and the TL play in the selection and use of strategies
- the impact of the testing format on the selection of test-taking strategies
- the appropriateness of the research methods used to study test-taking strategies
- the effectiveness of strategy instruction for improving respondents' performance on high-stakes standardized tests.

The assumption underlying many of these studies was that the test was, in fact, a valid measure of the construct. This stance was viewed as a departure from the earlier research, which seemed more focused on whether a given approach to assessment was valid (Cohen, 2006).

Since the mandate in this chapter is to ask whether respondents are using construct-relevant strategies to produce their responses, the question emerges as to where the locus of responsibility lies. If respondents do not use construct-relevant strategies, why is this?

- Is it because items or tasks on the given instrument are of questionable validity altogether and in need of revision?
- Is it because the respondents lack knowledge of how to respond to such items and tasks and need clearer instructions or better prior test orientation?
- Is it because the students lack proficiency or have characteristics that work to their detriment (e.g., their sociocultural background, gender, age, occupational status, and so forth) and would be better served by a more appropriate set of items or tasks?

An Update on Test-Taking Strategy Research and Implications for Item Design

Let us now take a look at some of the test-taking strategy studies conducted within the last two decades in order to see what the studies tell us about how to design and construct items and tasks. The focus will be on studies dealing with (1) the validity of the measures, (2) the relationship between language proficiency and test-taking strategy use, (3) test-taking strategies by skill area, (4) the washback effect, and (5) test preparation.

Studies focusing on the validity of the measures

The use of test-deviousness strategies with listening and reading tests

One study investigated the impact of test-deviousness strategies (traditionally referred to misleadingly as "test-wiseness" strategies) – identifying and using the cues related to absurd options, similar options, and opposite options – in taking the paper-and-pencil Test of English as a Foreign Language (TOEFL) (Yang, 2000). First, 390 Chinese TOEFL candidates responded to a modified version of Rogers and Bateson's (1991) Test of Test-Wiseness (TTW) (see Yang, 2000: 58–61) and the TOEFL Practice Test B (Educational Testing Service, 1995). An item analysis of the TTW results for a subsample of 40 led to the selection of 23 respondents who were considered "test-wise" and another 17 who were deemed "test-naïve." All students were asked to provide verbal reports about the strategies that they were using while responding to a series of items that were susceptible to the use of test-deviousness strategies, from the TTW and TOEFL. It was found that 48 to 64 percent of the items across the Listening and Reading Comprehension subtests of the TOEFL Practice Test B were identified as susceptible to deviousness strategies. It was also found that the "test-wise" students had a more meaningful, thoughtful and logical and less random approach to the items than did the test-naïve students – which meant that they also had better test-management strategies. In addition, these students were more academically knowledgeable and used that knowledge to assist them in figuring out answers to questions. Finally, they extended greater effort and were more persistent in looking for cues to getting items right without having to rely on language proficiency.

This study serves as a reminder that we need to keep performing test-deviousness studies as a means of checking whether tests are giving away the answers to items. I still remember the surprising results of a student study by Avi Israel, in which the EFL respondents received just the title of an English passage and had to respond to M-C questions about it (Cohen, 1984). The more proficient students, in particular, did far too well on the items for it to have been by chance. Even some of the less proficient students almost passed the test. The items were simply too guessable.

A more recent study investigated the strategies used in eliminating distractors while answering test items (Katalayi, 2018). It focused on the extent to which examinees used construct-relevant or construct-irrelevant strategies in eliminating distractors on the Congo State English Exam M-C reading questions. Verbal report was used to collect information from six HS seniors (two high achieving, two average achieving, and two low achieving) regarding their decision-making process. The findings indicated that respondents used both construct-relevant and construct-irrelevant strategies in determining which options to deselect. A construct-irrelevant strategy was to select an option after eliminating other options without reading the text or without understanding the question fully – hence, not perceiving the actual meanings and implications of certain options. The high school students used the strategy of relying on information contained in the options and used their own powers of logic or deductive reasoning. Also, they relied on clues to the answers found in the questions.

Validating a test of integrated reading and writing tasks

It is encouraging that Educational Testing Service (ETS) – a bastion of traditional approaches to test validation – chose to fund test-taking strategy studies as part of the validation process. In fact, one such effort by ETS to construct a reading-writing subtest for the internet-based

TOEFL iBT resulted in the elimination of the subtest altogether. The study looked at test takers' understandings of and responses to what were referred to as integrated reading/writing tasks, as these interacted with their English proficiency levels, and at related issues faced by raters (Lumley and Brown, 2004a, 2004b). The tasks in question from the prototype LanguEdge Courseware (ETS, 2002) required not only comprehension of the reading text, but also synthesis, summary, and/or evaluation of its content, all considered typical requirements of writing in academic settings. The study investigated how test takers at different writing proficiency levels interpreted the demands of these integrated tasks and identified the strategies they employed in responding to them.

The study drew on verbal report data from tests taken by 30 Mandarin, 15 Cantonese, and 15 Korean students in Beijing, Hong Kong, and Melbourne, along with verbal reports by four raters in the act of rating. Score data and the student texts were also analyzed. Findings included students' responses to tasks and their descriptions of strategies used, as well as raters' reactions to the student's texts. Numerous problems with this subtest were reported. For example, raters were found to have a major problem identifying copied language versus students' own wordings. Furthermore, whereas the researchers had initially hoped to elicit detailed information on how students selected information from the reading text or transformed the language of the input text into their own words, respondents had difficulty providing accurate insights as to how they had generated their texts. The results of this study, along with other in-house research, led to the removal of the subtest from the exam. So here is a clear example of how the collecting of test-taking strategy data both from respondents and from raters of responses can lead to significant action: in this case, scrapping the test altogether.

Validating a test of reading

Another ETS validation study involving verbal report entailed the description of reading and test-taking strategies that test takers used with different item types on the reading subtest of the LanguEdge Courseware materials (ETS, 2002) developed to familiarize prospective respondents with the TOEFL iBT (Cohen and Upton, 2006). The investigation focused on strategies used to respond to more traditional *single-selection* M-C formats (i.e., *basic comprehension* and *inferencing* questions) versus *selected-response* (i.e., multiple selection and drag-and-drop) *reading-to-learn* items, which were considered innovative at the time. The latter were designed to simulate the academic task of forming a comprehensive and coherent representation of an entire text, rather than focusing on discrete points in the text. The aim of the study was to determine whether the TOEFL iBT was actually measuring what it purported to measure, as revealed through verbal reports. In a test claiming to evaluate academic reading ability, it was felt that the emphasis needed to be on designing tasks calling for test takers to actually use academic reading skills, rather than relying on test-deviousness strategies.

Verbal report data were collected from 32 students, representing four language groups (Chinese, Japanese, Korean, and Turkish) as they did tasks from the reading subtest in the LanguEdge Courseware materials. Students were randomly assigned to complete two of the six reading tasks, each consisting of a 600-to-700-word text with 12 or 13 items, and subjects' verbal report accompanying items representing 10 different item formats was evaluated to determine strategy use.

The findings indicated that tasks on the TOEFL iBT reading subtest did, in fact, call for the use of academic reading skills for passage comprehension – at least for respondents

whose language proficiency was sufficiently advanced so that they not only took the test successfully but could also report on how they did it as well. Nevertheless, it was also clear that subjects generally approached the TOEFL iBT reading subtest as a test-taking task that required that they perform reading tasks in order to complete it. Thus, responding to selected passages from the reading subtest of the LanguEdge test did not fully constitute an academic reading task for these respondents, but rather a test-taking task with academic-like aspects to it. Two reading strategies found to be common to all subtests were:

- reading a portion of the passage carefully
- repeating, paraphrasing, or translating words, phrases, or sentences (or summarizing paragraphs/the passage) to aid or improve understanding.

While the respondents were found to use an array of test-taking strategies, these were primarily test-management strategies. A caveat with respect to this finding is that the respondents were perhaps reluctant to use test-deviousness strategies because they knew that their behavior was being monitored. The six common test-management strategies were:

- going back to the question for clarification: rereading the question
- going back to the question for clarification: paraphrasing (or confirming) the question or task (except for *basic comprehension-vocabulary* and *basic comprehension-pronoun reference* items)
- reading the question and then reading the passage/portion to look for clues to the answer either before or while considering options (except in the case of *reading to learn-prose summary* and *reading to learn-schematic table* items)
- considering the options and postponing consideration of the option (except for *inferencing-insert text* items, calling for insertion of a new sentence into a displayed section of the text)
- selecting options through vocabulary, sentence, paragraph, or passage overall meaning
- discarding options based on vocabulary, sentence, paragraph, or passage overall meaning as well as discourse structure.

In terms of whether items and tasks were doing what the test designers had envisioned, the study found the reading-to-learn and inferencing items **not** to be requiring different, more academic-like approaches to reading than the basic comprehension items. Since respondents started by considering words and sentences in larger chunks of text (even whole passages), basic comprehension item types were found to elicit comparable strategies to those required by the inferencing and reading-to-learn tasks. By the time the respondents arrived at the reading-to-learn items, they already had a good sense of the passage as a whole and consequently found these items easy. When encountering the prose summary item near the end of the test, they almost immediately knew which three statements (of five) to drag into a box in order to create a summary of the text. The findings from this study would ideally lead to refinement of innovative item types so that they better approximate the construct to be tested. Again, this study would serve as a warning to test designers not to assume that items deemed "innovative" are actually eliciting the cognitive processes that they are intended to elicit.

A somewhat more recent test-taking study investigated the validity of an M-C reading test for reading assessment in the Congo, consisting of 50 M-C reading items taken from the 2011 English State Examination (Katalayi and Sivasubramaniam, 2013). The items were

administered to 496 12th-grade students. A strategies questionnaire was used to elicit information on the respondents' use of strategies during their completion of the test, wherein respondents were asked to list all strategies that they remembered using to answer each question. To assist them in this task, they were given a list of 13 strategies:

- reading the text carefully/rapidly
- reading a portion of the text carefully/rapidly
- using their knowledge of grammar or vocabulary
- drawing conclusions based on text implications
- searching for part of the text recommended
- considering options and postponing consideration re the correct option
- focusing on a familiar option
- producing one's own answer after reading the question and then looking at the options
- selecting an option causing the elimination of other options
- selecting an option by guessing.

The concern of the investigators was that assessment should evaluate the ability to cope with common life reading behaviors such as skimming or search reading, labeled as *expeditious* reading. Such reading involved the extraction of information necessary to answer specific test items. The researchers contrasted this type of reading to *careful* reading, characterized by the search for the accurate comprehension of the text, involving word identification, the understanding of syntax, and the making of inferences. Careful reading was also seen to call for comprehension of meanings both within and beyond sentences in order to construct the macro-structure of the text. The results indicated that there were more items targeting careful reading than items targeting expeditious reading.

Another study looked at whether a reading test was actually assessing the intended construct of reading comprehension applied to that measure – namely, an individual's ability to engage with, understand, interpret, and make use of written English texts to achieve day-to-day and general workplace communicative functions, as opposed to simply assessing test-taking strategies (Wu and Stone, 2016). The researchers used sample data from the Canadian English Language Proficiency Index Program-General (CELPIP-General) reading subtest. The findings were that test takers engaged more in processing the texts for comprehending meaning, less in test-management strategies, and least in test-deviousness strategies (my term). They also found that higher engagement in comprehending meaning led to higher test performance, engagement in test management showed a small negative association with test performance, and higher engagement in test-deviousness strategies led to poorer performance.

Validating a test of speaking

With regard to speaking, a study investigated the effect on test performance of strategic behaviors that test takers reported using when responding to integrated and independent speaking tasks on the speaking section of the internet-based Test of English as a Foreign Language (TOEFL iBT) (Barkaoui *et al*., 2013). Each of 30 Chinese-speaking engineering students responded to two independent and four integrated speaking tasks and provided stimulated recalls about the strategies they used when performing each task. The integrated tasks elicited a higher percentage of cognitive and affective strategies than did the independent tasks. Additionally, the integrated tasks were more alike with respect to reported strategy

use than were the independent tasks. The relationships among strategy use, task type, and task performance were found to be complex.

Another study looked at speaking strategies in completing the simulated part 3 of the International English Language Testing System (IELTS) Speaking Test (Fernandez, 2018). Twelve undergraduate and graduate students, representing seven language backgrounds, provided stimulated recall data upon the completion of the speaking task. The researcher categorized strategies according to their traditional monolithic functions as metacognitive or cognitive and added a category – *communication strategies*. Strategies in this last category were found to overlap with metacognitive strategies, obfuscating somewhat its value as a separate category. Respondents were seen to deploy clusters of strategies in the completion of speaking tasks.

The picture that emerges from test validation studies is that the field has progressed beyond the days when tests were validated simply by statistical analysis of correct and incorrect responses. At present, respondents' verbal reports are also being collected in order to better determine what it actually entails to produce correct answers to various language assessment measures.

Studies relating language proficiency to test-taking strategy use

As seen earlier through one of the early studies as well (Gordon, 1987), limited TL proficiency can prompt respondents to use construct-irrelevant strategies. This insight reflects on test validity in the sense that test constructors need to take into consideration the likely proficiency levels of the respondents and design the test accordingly, unless the test is intended for the purposes of gatekeeping – i.e., to ensure that only the highly proficient will pass.

Strategies on a high-stakes test

Some studies looking at the proficiency level and strategy use relationship have targeted high-stakes tests. One example is a study conducted to investigate the test-taking strategies that 260 randomly selected undergraduate EFL students reported using when responding to reading tasks on the Iranian National University Entrance Exam for the MA in TESL (INUEMA) (Kashkouli *et al*., 2015). Based on TOEFL reading subtest results, 80 students were placed in the high-ability group, 88 in the intermediate group, and 78 in the low-ability group. All groups completed three reading subtests of the INUEMA tasks along with a test-taking strategy questionnaire, which included 27 Likert-scale items (in Farsi), 6 asking about reported planning strategies (previewing or overviewing tasks in order to determine actions to be taken), 13 about monitoring strategies (ongoing checking of comprehension, accuracy, and/or appropriateness of response behaviors), 4 about evaluation strategies (checking on their comprehension after completing the reading items), and 4 about test-wiseness strategies (using their knowledge and experience with the taking of reading tests). The findings revealed that the intermediate group reported using test-taking strategies more than the other two groups. The results also showed that monitoring (checking comprehension, accuracy, and/or appropriateness of responses) and evaluation strategies (checking comprehension after completion of the items) were reportedly used significantly more than other the types of strategies, which meant that the respondents relied more on their academic reading skills for both specific and general comprehension of the texts and less on what this chapter is referring to as test-deviousness strategies.

A second study investigated differences in the type and frequency of strategies used by 8 high-scoring and 16 low-scoring male Iranian IELTS candidates (Ahmadian *et al*., 2016). The participants were required to read two types of academic IELTS reading comprehension texts in a mock exam. The study relied on verbal report techniques to investigate the respondents' reading strategies while completing the reading module of the IELTS test. The results showed a major difference between the two groups of candidates in the use of compensation and metacognitive strategies. The high-scoring readers reported using these strategies in different ways and more frequently than the other group. Strategies labeled in terms of their function as "cognitive" included:

- highlighting and underlining
- translation
- rereading
- changing the reading rate
- skipping
- keeping keywords in mind
- summarizing
- paraphrasing.

Strategies labeled as "metacognitive" included self-monitoring, planning, and self-evaluation.

Strategies on an objective test of writing

With regard to writing assessment, a study compared the test-taking strategies of 20 native English speakers and 20 limited-English-proficient (LEP) students on an objective test of writing skills (Gavin, 1988). The students were enrolled in the first semester of college English composition. The study looked at the respondents' awareness of item objectives, their test-taking strategies, and their test performance. Results showed that the test questions were generally more difficult for the LEP students than for the native speakers and that as a means of coping, the LEP students used test-deviousness strategies that were construct irrelevant. When unaware of the test item objectives, the LEP students focused more on rules dealing with grammar, mechanics, or style than did the natives, who were focused more on meaning. They were also more likely to choose answers at random or to perform lower-level analysis.

The study brings up the issue of intuition and how difficult it can be for someone with limited TL proficiency to perform the kind of intuiting that will ensure correct test answers. Perhaps prior orientation about the test could include guidance as to how to use powers of intuition constructively in determining the correct response. Some years ago, I conducted a study comparing results when instructions as to how to read a Hebrew L1 and an English L2 text to summarize them (in the L1, Hebrew) and how to write the summary were provided and when they were not provided. The results showed that whereas more explicit instructions did not improve the summaries of the L1 text, they did make a significant difference when summarizing the English L2 passages (Cohen, 1993).

Strategies on a test of listening comprehension

A study involving the assessment of listening comprehension looked at the influence of language proficiency on test-taking strategy use among 75 Taiwanese college students in their

third and final year of a mandatory three-year EFL course (Chang, 2009). The participants completed a listening test comprising four M-C questions, three short answer questions, and three true-or-false questions about four stories, each three minutes long. The students completed a test-taking questionnaire before and after responding to the test. In addition, after each subtest, they reported the strategies used based on a strategy list and also added their own strategies not on the checklist. The most marked difference between the high-language-proficiency (HLP) and low-language-proficiency (LLP) students was how they used a strategy.

For example, when the listening input was repeated, the LLP students reported listening just for the main ideas at the first hearing and paid attention to the details only at the second hearing. In contrast, the HLP students reported trying to understand as much as they could the first time and used the second hearing for confirmation only. Another difference was that while both HLP and LLP students used the strategy "trying to hear every word," their reported purpose was different. Whereas the HLP students tried to hear every word for fear of missing any information that might affect their scores, the LLP students tried to match the words that they had learned with the words that they heard. According to Chang (2009), this implied that students may use the same strategy (in this case, trying to hear every word) for different purposes. Again, the message to the test designer may be to make the instructions (and prior test-taking orientation) clear as to the best way to deal with such listening tasks.

A more recent study focused on how 96 advanced, upper-intermediate, and intermediate Iranian EFL university students compared in their use of metacognitive and cognitive strategies in the completion of a listening comprehension test (Vahdany *et al*., 2016). It was found that the high-proficiency more than the intermediate listeners used self-management strategies, such as attending selectively and guessing the meaning of unfamiliar idioms by using the context and logic to infer the correct answers. In addition, such test-taking strategy use was found to have a positive correlation with listening test performance.

Strategies on reading and on reading-writing tasks

One assessment of ability across skills explored the test-taking strategies of 52 HLP, MLP, and LLP seventh-grade EFL students while completing five reading tasks and two integrated reading-writing tasks, locally constructed and assessed using the Common European Framework scale (Nikolov, 2006). The five reading tasks called for matching 10 questions with the correct 10 (out of 12) response choices. The reading-writing tasks involved identifying material in the passages and copying some of it:

> Task 1: Read a 60-word text about a pen pal and fill in a questionnaire about the person.
> Task 2: Complete a birthday party invitation with 8 or 10 well-chosen words.

The study was conducted to provide insights into how learners go about solving tasks and what they think about and rely on while doing them. The results showed that the strategies that the children applied were not inherently effective or ineffective and that the successful use of the strategies depended on how appropriately they applied them to the actual item in the task (e.g., "guessing," "relying on cognates," "applying a rule," or "skipping the rubrics for an item"). The students provided verbal report as they did the tasks in two 45-minute sessions (without prompting from the 26 researcher assistants). The researcher noted that strategies differed in type and magnitude, some involving major

cognitive processes and others referred to as "tiny tricks." More will be said about this study in the Research Methods section.

Strategies on a gap-filling test

With regard to gap-filling, six HLP and six LLP Japanese EFL university students completed a 16-item gap-filling test while providing verbal report (Yamashita, 2003). The text comprised 336 words, and the items were all deemed to require text-level understanding for closure. Results showed that both HLP and LLP readers reported most frequently using text-level information. Although there were several cases in which the items were answered correctly with local grammatical clues and extra-textual background knowledge, overall, the gap-filling test generated text-level processing and differentiated well between the HLP and LLP readers. Hence, the researcher concluded that the findings supported the claim that a gap-filling test can be used as a test to measure higher-order processing ability.

In taking a closer look at the data, it was found that the HLP readers reported using text-level information more frequently than the LLP readers. Qualitative analysis of the protocols revealed further differences between the groups. Whereas the LLP readers had sensitivity about the type of information necessary to answer the items, they were less successful than the HLP readers in their attempts and resorted to other sources such as clause-level information and guessing. A key difference between the groups was in the way that the clause-level information was used. The HLP readers were able to give different weight to different types of information according to their importance in understanding the text. For them, clause-level information was only used as a source for confirming their answers. The LLP readers, on the other hand, were less able to use text-level information and put heavier emphasis on local grammatical information.

A more recent study explored Korean EFL students' test-taking strategy use on gap-filling inference items from the College Scholastic Aptitude Test (CSAT) by means of retrospective verbal report (Nam, 2015). Eight college freshmen provided Korean L1 verbal report regarding their processes in answering three gap-filling M-C inference items with five alternatives and only one gap to fill in. The students with higher proficiency levels tended to use more variety in their test-taking strategies – e.g., predicting text meaning, confirming guesses, and looking for sentences that conveyed the main idea. LLP students often relied on familiar words and distractors as avoidance strategies and reported using test-deviousness strategies such as matching words.

Strategies on a test of pragmatic ability in speaking

In recent years, there has been heightened interest in test-taking strategies on measures of pragmatic ability. One such study looked at the reported strategy use of LLP and HLP test takers with their rated performance on L2 pragmatics as assessed on speaking tasks (Youn and Bi, 2019). Thirty university-level ESL learners completed four pragmatic speaking tasks that differed in the formality of pragmatic actions: (1) a monologic speaking task calling for the provision of peer feedback on classmates' in-class writing, (2) a role-play request to a professor for a recommendation letter during an office hour visit, (3) a role-play refusal by a professor to a student request to change an upcoming class presentation schedule because there is an upcoming final exam in another class, and (4) a role-play with a classmate to negotiate an agreeable meeting time and place to discuss an upcoming class project. Two raters scored their performance using analytical rating criteria. The test

takers' retrospective verbal reports were transcribed and analyzed to develop a valid coding scheme that consisted of cognitive, metacognitive, and pragmatic strategies. The test takers' reported strategy use was compared to the raters' scores for them. Compared with the LLP test takers, the HLP respondents utilized diverse strategies more frequently, ranging from varied pragmatic strategies to strategies specifically related to managing task demands. Across all respondents, metacognitive strategies unique to task-based performance (e.g., assessing the task-related situation) were more commonly reported than the cognitive (e.g., comprehending task instructions, linking prior knowledge or experience) and pragmatic strategies (ranging from the use of general pragmatic awareness to situation-related sociopragmatic and pragmalinguistic strategies).

Studies focusing on test-taking strategies by skill area

Listening comprehension

With regard to the skill of listening, a recent study investigated the impact of test-taking strategies on listening comprehension performance by Omani 11th-grade students (Bulushi *et al*., 2019). The sample included 20 female students from a government school in an experimental group and 27 in a control group. An eight-week treatment involved instruction in the use of metacognitive, cognitive, and affective strategies before, during, and after taking a listening comprehension measure. Even though the students in the control group were found to benefit from practice in listening comprehension, the students in the experimental group derived an extra benefit from the strategy instruction that they received. They perceived the use of test-taking strategies not only as enhancing their listening, but also as enhancing performance in other language skill areas as well, such as reading.

Speaking

With regard to the skill of speaking, a recent study investigated the relationship of test-taking strategies with performance on the speaking component of a large-scale standardized English proficiency measure in Taiwan, the General English Proficiency Test-Intermediate (GEPTI-S) (Huang, 2016). A total of 244 Taiwanese EFL learners took the GEPTI-S and completed a strategy inventory which featured strategies specific to the three task types on the English proficiency measure: namely, reading the given texts aloud, answering the proposed questions, and describing pictures. This study included strategies that were specific enough in nature so that there was little ambiguity as to how to operationalize them:

- focusing on the keywords
- rehearsing responses before making them
- paying attention to blended sounds and intonation
- using familiar sentence patterns in responses
- simplifying answers
- slowing down when responding
- rehearsing or skipping unknown words in read-aloud tasks.

Exploratory factor analysis and two-step structural equation modeling performed on the data revealed that the GEPTI-S induced construct-relevant strategic behaviors, which offered evidence to support claims of test validity.

Reading

With regard to the skill of reading, a study investigated the role that metacognitive and cognitive strategies played for 98 Saudi Arabian EFL learners in responding to a reading test with four task formats by having the respondents also complete a strategy questionnaire (Assiri, 2014). Respondents reported that *planning strategies* helped them manage their use of comprehension, memory, and retrieval abilities in response to constructed-response questions. They reportedly used *monitoring strategies* to aid their performance on tasks that required judgment, such as true/false questions, and *evaluation strategies* to boost performance on selected-format tasks, such as M-C questions.

In another study, researchers examined whether the previewing of M-C cloze alternatives would affect item response accuracy and response time among Chinese learners and English native speakers and whether the relation varied as a function of English proficiency level (Guo *et al*., 2016). The researchers examined the eye-movement patterns of participants who completed an online M-C sentence completion task, which meant that the text was more difficult to preview than the original due to the deletions. The results showed that, relative to no previewing, previewing was associated with a significantly lower probability of answering an item correctly. Importantly, these relations varied across English proficiency levels such that HLP respondents performed better without previewing, but there was no difference for lower-intermediate learners of English. These findings suggested that previewing a cloze text with deletions interfered with the comprehension process, particularly for HLP respondents.

Studies focusing on the washback effect

At least one study has looked at the washback effect of taking high-stakes tests on report of test-taking strategy use. The study focused on the responses of 284 non-English majors in China to a test-taking strategy questionnaire. It compared the responses of those who had taken the College English Test (CET) with those of students who had not (Wei, 2014). What distinguished this study from others is that instead of doing item-by-item or task-by-task analysis of reported strategies while a respondent was taking a measure or just afterward, the study relied on respondents' self-report of their strategies when taking tests. The findings showed that reported strategy use was more oriented toward test-taking strategies (test-management and test-deviousness strategies) than toward language learning and use strategies (cognitive and metacognitive). With regard to possible washback effects, the group who had taken the CET reported somewhat more use of cognitive strategies than those who had not and only slightly more use of test-management and test-deviousness strategies.

Studies focusing on test preparation

To examine the effects of preparing for a high-stakes listening test, a study focused on the 16-item listening subtest from the International English Language Testing System (IELTS) (Winke and Lim, 2017). One experimental group took four practice tests and received explicit strategy instruction with regard to test taking, another experimental group took the same four practice tests but did not receive strategy instruction, and a control group did not take practice tests. Listening test preparation was not found to have any effect on performance for the given listening subtest. The main effect of test preparation seemed to be that

of familiarizing the respondents with the test format and the item types, especially items that were relatively complex and unfamiliar to the test takers.

Research methods for investigating test-taking strategies

In this section, let us first look at the different types of verbal report as applied to test-taking strategy research. Then let us briefly look at the fact that strategies are rarely used in isolation, but rather in consort with other strategies.

The use of verbal report

Verbal report became a primary research tool for this endeavor in the early 1980s (see Cohen, 1984, 2013). Verbal reports include data reflecting one or more of the following approaches.

1 *Self-report* – learners' descriptions of what they do, characterized by generalized statements: in this case, about their test-taking strategies – for example, "On multiple-choice items, I tend to scan the reading passage for possible surface matches between information in the text and that same information appearing in one of the alternative choices."
2 *Self-observation* – the inspection of specific, contextualized language behavior, either introspectively – that is, within 20 seconds of the mental event – or retrospectively – for instance, "What I just did was to skim through the reading passage for possible surface matches between information in the text and that same information appearing in one of the alternative choices."
3 *Self-revelation* – "think-aloud," stream-of-consciousness disclosure of thought processes while the information is being attended to – for example, "Hmm . . . I wonder if the information in one of these alternative choices also appears in the text."

Whereas the test-taking strategy studies described in this chapter all relied in one way or another on verbal report, most of them involved self-report by means of survey questionnaires. Many fewer involved think-aloud, introspective, or retrospective data.

While verbal report continues to play a key role in test-taking strategy research, there have been changes in procedures for conducting such verbal reports aimed at improving the reliability and validity of the results. One has been to provide a model for respondents as to the kinds of responses that are considered appropriate (see, for example, Cohen and Upton, 2006) in order that they produce enough relevant data to make the study viable. In addition, researchers now – unlike in the past – may intrude and ask probing questions during data collection in order to make sure, for instance, that the respondents indicate not just their rationale for selecting, say, *b* as the correct alternative in M-C items, but also their reasons for rejecting *a*, *c*, and *d*. Respondents have also been asked to listen to a tape recording or read a transcript of their verbal report session in order to complement those data with any further insights they may have (Nyhus, 1994). The bottom line is that verbal report does not constitute a single approach but rather is a vehicle for enlisting a number of different approaches.

The challenge of describing the clustering of strategies

An important finding from the Nikolov (2006) study described earlier was that some strategies tended to occur in combination with others. For example, "reading the text in English"

and "translating" were often combined with "phonetic reading of unfamiliar vocabulary items." Also, respondents varied as to how they used metacognitive strategies to approach the different tasks. For instance, they tended to check the instructions for several tasks but neglected to do so for the others, or they started by reading the example in the first task and then proceeded to respond to the rest of the items without paying attention to the rubrics or the examples in other tasks. Given the individual variation, Nikolov found that efforts to categorize the data were highly problematic as many strategies overlapped, and others could be subdivided into other strategies.

In corroboration of Nikolov's finding, language learner strategy experts have agreed that strategies are, in fact, often used in combination (Cohen, 2007: 35–36), despite the fact that they are often treated in the literature as if they occurred in an isolated fashion.

Recommendations for practice

With regard to test-management strategies, the message for test designers would be to make sure respondents receive clear instructions as to how to make use of such strategies:

1 Read the instructions for the whole test and for each question carefully
2 Make use of key words in the questions and underline them
3 Identify difficult questions that require more time to process, answer the easier questions first, and delay answering the difficult questions
4 Allot time per question according to its difficulty and follow these time allotments
5 Before handing in the test, check over the responses.

(Amer, 1993)

In this section, let us look at just a few illustrative test-taking strategies that emerge from this work with respect to a few of the prominent formats for assessing language (adapted from Cohen, 2011b).

Strategies in taking cloze tests

Respondents could be alerted both in prior test-orientation sessions and perhaps even in the instructions to strategies such as the following:

- Looking at the words and phrases immediately preceding and following blanks, as well as at the extended context, for clues as to how to fill them in.
- Rephrasing sentences to make sure that the intended meaning is clear.
- Recalling knowledge of the passage and any other prior knowledge that could be useful for filling in the blanks.

Strategies in dealing with M-C Items

The following are some test-management strategies for reading and listening tests that could be included in test-orientation sessions and perhaps in the instructions for the tests as well.

Test-management strategies for reading

- Go back to the question for clarification – reread it carefully or paraphrase it in order to verify or confirm understanding of the question or task, since the necessary clue may be in the wording of the question.
- Read the question(s) first and then read the passage/portion in order to expedite the search for the material in the text needed to answer the question.
- Select or eliminate options based on vocabulary knowledge or on the meaning of the sentence, paragraph, overall passage, or discourse structure.
- Perform a mental translation of parts of the text to see if the material makes sense in the L1, and then summarize the passage as a check for comprehension.

Test-management strategies for listening items

- Verify if the options match elements of the listening text or the question in terms of key words, specific details, inferences about details, and level of specificity.
- Check back to part or all of a prior question as a guide to selecting a response to the item at hand.
- Determine the level of detail required in answering a given question so as to reject an option that is either too general or too specific.
- Identify relevant background knowledge and then utilize it in an appropriate way.
- If unsure of the response, make an educated guess, drawing on a combination of strategies such as those listed here.

Recommendations for test-taking strategy instruction

The study on test-taking strategies in listening reported on earlier (Chang, 2009) included recommendations for how to integrate strategy instruction for taking such tests into regular adult EFL classroom instruction. While the following steps are intended primarily for listening assessment, they are meant to be illustrative of how to raise student awareness of how they can use test-taking strategies to their benefit:

1 The teacher provides students with a checklist of strategy use for a given task.
2 Before listening, students look at the strategy checklist for a minute and check off the strategies that they have ready for the task.
3 After listening, students spend a few minutes thinking about the strategies that they actually used for the task.
4 Students are provided a script so that they can confirm what they hear with what they read.
5 Students are encouraged to listen to the texts as many times as they wish without looking at the script.
6 More or less successful students are assigned to groups after a task so that they can compare and discuss the strategies used.
7 Teachers pick one or two students (without revealing these students' names) who perform the best and share their strategy use with the whole class.

Further directions in research

Ideally, item and task designers will be increasingly cognizant of the need to prompt response strategies that call on the desired language knowledge and skills. Further research describing test-taking strategies used to respond to items and tasks in various TLs and using a timely sampling of elicitation and response formats will undoubtedly support this effort. Aside from TL and testing format, researchers need to be mindful of:

- the respondents' TL proficiency level, gender, sociocultural group, and other characteristics;
- the most appropriate methods for collecting the data; and
- the means for meaningfully linking strategy use with performance on the language measures.

With regard to the final bullet point, it would be helpful to monitor the extent to which adapting test-taking strategies to fit the demands of the assessment items and tasks actually contributes to their performance. Another concern would be the actual contribution of research on test-taking strategies to making language tests more valid.

Finally, research should continue to investigate the impact on test performance of guides for respondents as to how to take high-stakes standardized tests. In a study of 43 students at a coaching school in Taiwan, for example, high scorers were found to focus on their understanding of the passages, use the strategies taught by the coaching school only as an auxiliary to comprehension, and tended to personalize strategy use (Tian, 2000). The low scorers, on the other hand, followed the test-taking strategy instruction more mechanically, focused on word-level strategy use, and used test-deviousness strategies as a way of circumventing comprehension. These findings should serve as a warning to test designers that strategy instruction materials may not necessarily help those who need them the most and perhaps most benefit those who least need assistance.

By way of closing, what might we predict the field of test-taking strategies will look like in in the future? What debates will remain? What issues will be unresolved? Predictably, there will always be new testing formats heralded in the literature, and so there will be a continuing need to conduct test-taking strategy research as a means for validating them. Dynamic assessment (DA) was a relatively new means of assessment when the earlier version of this chapter was published in 2012. DA is based on the Vygotskyian perspective that assessment occurs not in isolation from instruction but as an inseparable feature of it. This approach to assessment challenges the widespread acceptance of independent performance as the privileged indicator of individuals' abilities and calls for assessors to abandon their role as observers of learner behavior in favor of a commitment to joint problem solving aimed at supporting learner development, whereby teachers collaborate with learners on increasingly complex tasks (Poehner, 2007, 2008; see also Antón and García, this volume). It would appear beneficial to conduct test-taking strategy research looking at respondents' test-taking and language learning strategy use on given instances of DA appearing at different points along a continuum from language assessment to language instruction.

More recent forms of assessment include those involving, say, pragmatics (see Cohen, 2020) and especially those that involve the use of technology, such as in game-based assessment (Russell and Vásquez, 2013; Sawaki, this volume).

Further reading

Bowles, M. A. (2010). *The Think-Aloud Controversy in Second Language Research*. Abingdon, UK: Routledge.
The book deals with the validity and use of respondents' verbal reports during the performance of language tasks. After presenting theoretical background and empirical research on the validity of think alouds, the author gives an overview of how think alouds have been used in L2 language research, as well as a meta-analysis of findings from studies involving think alouds on verbal tasks. The volume also offers guidance regarding the practical issues of data collection and analysis.

Brown, J. D. (ed.). (2013). *New Ways of Classroom Assessment. Revised*. ERIC ED549559 (396 pp.). Alexandria, VA: Teachers of English to Speakers of Other Languages.
While not focused on test-taking strategies per se, the volume constitutes a compendium of everyday classroom assessment activities that provide a way of observing or scoring students' performances and giving feedback that is meant to inform students and teachers about the effectiveness of the teaching and learning taking place. Each activity comes with suggestions as to how to give feedback in the form of a score or other information (e.g., notes in the margin, written prose reactions, oral critiques, teacher conferences). Many of the entries utilize other possible feedback perspectives aside from that of the teacher: namely, self-assessment, peer assessment, and outsider assessment – often used in conjunction with teacher assessment. One entry on "Preparing students for tests" by Alastair Allan (pp. 205–09) expressly deals with what I would term test-deviousness strategies. Although not necessarily calling upon the teacher to be a collaborator in the assessment process as does dynamic assessment (Poehner, 2007, 2008) the assessment activities in this volume are more aligned with DA than are traditional language assessment activities. For table of contents, go to www.tesol.org/read-and-publish/bookstore/toc/NW_classroomassessmentrevised_TOC (accessed April 14, 2019).

Green, A. J. F. (1998). *Using Verbal Protocols in Language Testing Research: A Handbook*. Cambridge, UK: Cambridge University Press.
Referring to verbal report as *verbal protocol analysis* (VPA), the author notes that it is a complex methodology, warranting orientation of users to maximize its benefits. The book provides background regarding VPA in research on language assessment (the design of data collection methods, the development of coding schemes for analyzing the data, and the actual analysis of the data).

Gass, S. M. and Mackey, A. (2000). *Simulated Recall Methodology in Second Language Research*. Mahwah, NJ: Lawrence Erlbaum.
The book focuses on retrospective verbal report data, referred to as *stimulated recall*, and gives recommendations for how to collect and analyze such data. The authors also consider issues of reliability and validity and uses for stimulated recall – for example, in comprehending and producing oral language; understanding the dynamics of L2 classroom interaction; looking at processes in L2 reading; and investigating L2 syntactic development, vocabulary acquisition, and pragmatics.

Kelly, M. (2019). *10 Reading Comprehension Strategies All Students Need*. Retrieved April 14, 2019, from www.thoughtco.com/reading-comprehension-strategies-7952.
This is just one recent offering from ThoughtCo, an internet-based company that provides many articles on language assessment, accessible through a search engine.

References

Ahmadian, M., Poulaki, S. and Farahani, E. (2016). Reading strategies used by high scoring and low scoring IELTS candidates: A think-aloud study. *Theory and Practice in Language Studies* 6: 408–16. http://dx.doi.org/10.17507/tpls.0602.25.

Allan, A. (1992). Development and validation of a scale to measure test-wiseness in EFL/ESL reading test-takers. *Language Testing* 9: 101–22. https://doi.org/10.1177/026553229200900201.

Amer, A. A. (1993). Teaching EFL students to use a test-taking strategy. *Language Testing* 10: 71–77. https://doi.org/10.1177/026553229301000104.

Assiri, M. (2014). Metacognitive and cognitive strategies use and performance on a reading test with multiple-format tasks. *Arab World English Journal* 5: 187–202.

Barkaoui, K., Brooks, L., Swain, M. and Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics* 34: 304–24. https://doi.org/10.1093/applin/ams046.

Bulushi, A., Al Seyabi, F. and Al-Busaidi, S. (2019). The role of test taking strategies in improving Omani students' listening comprehension. *International Journal of Education* 10: 57–77. https://doi.org/10.5296/ije.v10i4.13951.

Chang, A. C.-S. (2009). EFL listeners' task-based strategies and their relationship with listening performance. *TESL-EJ* 13. Retrieved April 9, 2019, from www.tesl-ej.org/wordpress/issues/volume13/ej50/ej50a1/.

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing* 1: 70–81. https://doi.org/10.1177/026553228400100106.

———. (1993). The role of instructions in testing summarizing ability. In D. Douglas and C. Chapelle (eds.), *A New Decade of Language Testing: Collaboration and Cooperation*. Alexandria, VA: TESOL, 132–60.

———. (2006). The coming of age of research on test-taking strategies: Collaboration and cooperation. *Language Assessment Quarterly* 3: 307–31. https://doi.org/10.1080/15434300701333129.

———. (2007). Coming to terms with language learner strategies: Surveying the experts. In A. D. Cohen and E. Macaro (eds.), *Language Learner Strategies: 30 Years of Research and Practice*. Oxford, UK: Oxford University Press, 29–45.

———. (2011a). *Strategies in Learning and Using a Second Language: Research and Practice*. Harlow, UK: Longman/ Pearson Education.

———. (2011b). Test-taking strategies. In C. Coombe, P. Davidson, B. O'Sullivan and S. Stoynoff (eds.), *The Cambridge Guide to Assessment*. Cambridge, UK: Cambridge University Press, 96–104.

———. (2013). Verbal report. In C. A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. Oxford, UK: Wiley-Blackwell. https://sites.google.com/a/umn.edu/andrewdcohen/publications/research-methodology (Retrieved from "Research Methodology" list, April 14, 2019).

———. (2020). Considerations in assessing pragmatic appropriateness in spoken language. *Language Teaching* 53: 183–202. https://doi.org/10.1017/S0261444819000156.

Cohen, A. D. and Upton, T. A. (2006). *Strategies in Responding to the New TOEFL Reading Tasks* (Monograph No. 33). Princeton, NJ: ETS. https://sites.google.com/a/umn.edu/andrewdcohen/publications/assessment (Retrieved from "Assessment" list, April 14, 2019).

Cohen, A. D. and Wang, I. K.-H. (2018). Fluctuation in the functions of language learner strategies. *System* 74: 169–82. https://doi.org/10.1016/j.system.2018.03.011.

Davidson, F. (2000). The language tester's statistical toolbox. *System* 28: 605–17. https://doi.org/10.1016/S0346-251X(00)00041-5.

Dollerup, C., Glahn, E. and Rosenberg Hansen, C. (1982). Reading strategies and test-solving techniques in an EFL-reading comprehension test – a preliminary report. *Journal of Applied Language Study* 1: 93–99.

Educational Testing Service (ETS). (1995). *TOEFL Practice Test B in TOEFL Practice Tests*, vol. 1. Princeton, NJ: ETS, 55–106.

———. (2002). *LanguEdge Courseware Score Interpretation Guide*. Princeton, NJ: Educational Testing Service.

Fernandez, C. J. (2018). Behind a spoken performance: Test takers' strategic reactions in a simulated part 3 of the IELTS speaking test. *Language Testing in Asia* 8. Retrieved April 9, 2019, from https://doi.org/10.1186/s40468-018-0073-4.

Gavin, C. A. (1988). *The Strategies of Native and Limited English Proficient Test-Takers as Revealed by Think Aloud Protocols*. Unpublished EdD thesis, Rutgers University, New Brunswick, NJ.

Gordon, C. (1987). *The Effect of Testing Method on Achievement in Reading Comprehension Tests in English as a Foreign Language*. Unpublished master of arts thesis, Tel-Aviv University, Ramat-Aviv, Israel.

Guo, Q., Kim, Y.-S. G., Yang, L. and Liu, L. (2016). Does previewing answer choice options improve performance on reading tests? *Reading and Writing* 29: 745–60. https://doi.org/10.1007/s11145-016-9626-z.

Homburg, T. J. and Spaan, M. C. (1981). ESL reading proficiency assessment: Testing strategies. In M. Hines and W. Rutherford (eds.), *On TESOL'81*. Washington, DC: TESOL, 25–33.

Huang, H.-T. D. (2016). Exploring strategy use in L2 speaking assessment. *System* 63: 13–27. https://doi.org/10.1016/j.system.2016.08.009.

Kashkouli, Z., Barati, H. and Nejad Ansari, D. (2015). An investigation into the test-taking strategies employed for a high-stake test: Implications for test validation. *International Journal of Research Studies in Language Learning* 4: 61–72. https://doi.org/10.5861/ijrsll.2014.852.

Katalayi, G. B. (2018). Elimination of distracters: A construct-irrelevant strategy? An investigation of examinees' response decision processes in an EFL multiple-choice reading test. *Theory and Practice in Language Studies* 8: 749–58. http://dx.doi.org/10.17507/tpls.0807.05.

Katalayi, G. B. and Sivasubramaniam, S. (2013). Careful reading versus expeditious reading: Investigating the construct validity of a multiple-choice reading test. *Theory and Practice in Language Studies* 3: 877–84. https://doi.org/:10.4304/tpls.3.6.877-884.

Lumley, T. and Brown, A. (2004a). Test-taker and rater perspectives on integrated reading and writing tasks in the Next Generation TOEFL. *Language Testing Update* 35: 75–79.

———. (2004b). *Test Taker Response to Integrated Reading/Writing Tasks in TOEFL: Evidence from Writers, Texts, and Raters*. Final Report to ETS. Language Testing Research Centre, The University of Melbourne, Melbourne, Australia.

Nam, Y.-K. (2015). Korean EFL Students' Strategy Use in Gap-Filling Inference Items. *English Teaching* 70: 81–107. https://doi.org/10.15858/engtea.70.4.201512.81.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing* 6: 199–215. https://doi.org/10.1177/026553228900600206.

Nikolov, M. (2006). Test-taking strategies of 12-and 13-year-old Hungarian learners of EFL: Why whales have migraines. *Language Learning* 56: 1–51. https://doi.org/10.1111/j.0023-8333.2006.00341.x.

Nyhus, S. E. (1994). *Attitudes of Non-Native Speakers of English Toward the Use of Verbal Report to Elicit Their Reading Comprehension Strategies*. Plan B Masters Paper, Department of ESL, University of Minnesota, Minneapolis.

O'Sullivan, B. and Weir, C. J. (2010). Language testing = validation. In B. O'Sullivan (ed.), *Language Testing: Theories and Practices*. Basingstoke, UK: Palgrave Macmillan.

———. (2011). Test development and validation. In B. O'Sullivan (ed.), *Language Testing: Theories and Practice*. Basingstoke, UK: Palgrave Macmillan, 13–32.

Oxford, R. (2017). *Teaching and Researching Language Learning Strategies: Self-Regulation in Context*, 2nd edn. Abingdon, UK: Routledge.

Poehner, M. E. (2007). Beyond the test: L2 dynamic assessment and the transcendence of mediated learning. *Modern Language Journal* 91: 323–40. https://doi.org/10.1111/j.1540-4781.2007.00583.x.

———. (2008). *Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting L2 Development*. New York, NY: Springer.

Rogers, W. T. and Bateson, D. J. (1991). The influence of test-wiseness on the performance of high school seniors on school leaving examinations. *Applied Measurement in Education* 4: 159–83. https://doi.org/10.1207/s15324818ame0402_5.

Russell, V. and Vásquez, C. (2018). Assessing the effectiveness of a web-based tutorial for interlanguage pragmatic development prior to studying abroad. *IALLT Journal of Language Learning Technologies* 48: 69–96. Retrieved April 14, 2019, from https://journals.ku.edu/IALLT/article/view/8579.

Tian, S. (2000). *TOEFL Reading Comprehension: Strategies Used by Taiwanese Students with Coaching-school Training*. Unpublished PhD dissertation, Teachers College, Columbia University, New York.

Vahdany, F., Akbari, E., Shahrestani, F. and Askari, A. (2016). The relationship between cognitive and metacognitive strategy use and EFL listing test performance. *Theory and Practice in Language Studies* 6: 385–91. http://dx.doi.org/10.17507/tpls.0602.22.

Wei, X. (2014). The intensity and direction of CET washback on Chinese college students' test-taking strategy use. *Theory and Practice in Language Studies* 4: 1171–77. https://doi.org/:10.4304/tpls.4.6.1171-1177.

Wu, A. D. and Stone, J. E. (2016). Validation through understanding test-taking strategies: An illustration with the CELPIP-General reading pilot test using structural equation modeling. *Journal of Psychoeducational Assessment* 34: 362–79. https://doi.org/10.1177/0734282915608575.

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing* 20: 267–93. https://doi.org/10.1191/0265532203lt257oa.

Yang, P. (2000). *Effects of Test-Wiseness Upon Performance on the Test of English as a Foreign Language*. Unpublished PhD dissertation, University of Alberta, Edmonton, CN.

Youn, S. J. and Bi, N. Z. (2019). Investigating test-takers' strategy use in task-based L2 pragmatic speaking assessment. *Intercultural Pragmatics* 16: 185–215. https://doi.org/10.1515/ip-2019-0009.

Section 7
Prototyping and field tests


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

23

Prototyping new item types

Susan Nissan and Elizabeth Park

Introduction

We begin this chapter by defining prototyping and differentiating it from other terms that refer to evaluating test materials before they are used operationally. Next, we consider variables that inform the appropriate sample size and the kinds of information that prototyping can provide: for example, task design (such as task directions), task format, and task timing. The chapter covers many of these aspects of task design by describing (1) prototyping tasks for new or expanded constructs and (2) prototyping alternative tasks for known constructs. Although prototyping is carried out by many language testing programs, the examples in this chapter are from tests designed at Educational Testing Service (ETS) only. The authors, both on the test development staff of ETS, have worked on numerous prototyping projects and draw on that experience to provide the examples in this chapter. We conclude the chapter with some thoughts about the use of tools that new technologies offer.

In a language testing context, prototyping refers to trying out new test materials in order to evaluate their appropriateness for use prior to making final decisions about the operational test design. Prototyping may provide information about the design of a new test or about revisions to an ongoing test. Before prototyping begins, a test construct must be defined, and the abilities to be measured as part of the construct must be carefully considered and documented. By prototyping a variety of task types, we can investigate the extent to which theories and expectations about what the tasks measure are accurate (e.g., by comparing examinees' performance with established measures of the construct) and the extent to which construct-irrelevant variance occurs because of task format (e.g., by interviewing examinees about their prototype experience).

Numerous factors contribute to decisions about the scope and type of prototyping that is needed. If the task types are novel and little is understood about how examinees might perform, multiple iterations of prototyping events may be necessary, and several of the prototyping examples in this chapter refer to subsequent iterations (see Measuring a new construct: pragmatic understanding, Measuring an expanded construct: higher-level reading abilities, and Developing a new rubric). When introducing a well-known task type into a revised test, the reasons for the revision (e.g., to improve reliability and/or discrimination or

DOI: 10.4324/9781003220756-31

to align the test with current understanding of the construct) will inform the types of prototyping questions to be addressed.

In addition to *prototyping*, other terms are used to describe various ways of trying out new task types before they are used operationally. Some of these terms are used interchangeably with each other and with the term *prototyping*, while others indicate real differences in the purpose or stage of pre-operational administration. Based on our experience, test design ideally consists of three stages: prototyping, piloting, and field testing (each of the three may require iterative investigations). We offer these terms and their definitions to clarify what *prototyping* does and does not refer to in this chapter.

Usability testing is a type of prototyping that focuses on interface features and directions in computer-based tests to determine how easy it is for examinees to understand and use the interfaces. A small group of examinees (e.g., five to ten) can provide useful information about these issues. Participants are observed as they interact with an interface, and a questionnaire can be used to obtain their feedback.

The pilot phase, referred to as *item trialling* or *pilot testing*, provides feedback to test designers about the appropriateness and efficacy of new task types for gauging test takers' knowledge, skills, and abilities (see also Read, this volume). The focus of this phase is to evaluate new task types, task directions, and scoring rubrics (i.e., scoring guidelines). The terms *pre-testing* and *item tryouts* refer to trying out specific exemplars of task types that already exist in the operational test specifications. The pilot phase requires a representative sample of the test population to provide information about item characteristics such as difficulty and discrimination values. The particular items in the pilot are analyzed for how well they perform as individual exemplars of the task types in use. Items with poor performance at this stage are revised or replaced. This purpose is different from one of the major purposes of prototyping, which is to determine how well a proposed task type functions before it is considered for operational use.

During the "field test" phase, one or more versions of an entire test form is administered to determine its feasibility as a full operational test form. Field testing generally focuses on information about the functioning of the test as a whole, including such issues as test timing, reliability, and creating an appropriate score scale. In order to obtain good item statistics and test timing information, field testing typically requires a large number of candidates (e.g., 200 to 1,500) who represent the target test population. These later stages of test design are addressed in detail in Kremmel, Eberharter, and Holzknecht's chapter (this volume).

Sampling for prototyping

Throughout this chapter, sample sizes are identified as small, medium, and large, ranging from 5–20, 20–100, and more than 100, respectively. The size of the sample depends on the type and number of questions the prototype study is designed to address. A large sample for multiple-choice items can reveal if the tasks are doable for the population, if items are of appropriate difficulty, and if test directions are clear. For high-stakes tests[1], pilot and field test studies with an even larger representative sample typically follow prototype studies. Studies with this larger sample are required before the test becomes operational to ensure the most equitable, reliable, and valid measurement possible.

Because the test populations for many high-stakes English language tests come from a variety of countries, the sample must be representative of a wide range of national and linguistic backgrounds. The better the prototyping sample reflects the characteristics of the test population, the better it will predict how the items will perform if the task types become

operational, so each of the most important characteristics of the test population should be considered.

In some cases, it may be helpful to include native speakers (NSs), for example, if the intent is to measure a new aspect of a construct. During the design phase of the TOEFL iBT test, NSs were used in several prototype studies, including one that investigated pragmatic understanding (PU), an aspect of the listening construct that had not previously been systematically measured. Test designers included NSs in the prototype sample to determine whether NSs would be able to answer these types of questions. It would not be appropriate to expect non-native speakers (NNSs) to perform tasks that were problematic for NSs.

However, a large sample is not always necessary or appropriate. In usability studies where examinees are observed interacting with computer interfaces, the sample may be small. For example, Fulcher (2003) describes a model for interface design with three iterative groups of five participants each. When trying out an item type to determine whether it provides appropriate evidence for a previously unmeasured construct, it may be worthwhile to try it out initially on a small number of participants, even as few as ten. If participants understand the directions and appropriate evidence is obtained, additional studies can be undertaken with larger, more representative samples. This iterative approach is common during prototyping as typically it is not possible to address every question with one prototype study.

Usually the prototype sample for productive skills is smaller than for receptive skills since human raters analyze the performance tasks, and human rating is costly and time consuming. Determining an adequate sample size will depend on the nature of the construct, the number of score points within the prototype rubric, and the number of criteria within the rubric. When a goal of prototyping is to develop a new scoring rubric, a medium-size sample may be necessary as it is important that the prototype sample reflects the full range of abilities of the operational test population to confirm that all the score points of the rubric are appropriate. Additional data collection (e.g., pilot testing, field testing) is likely necessary with larger sample sizes to validate and refine the rubric.

Viability of measuring new and expanded constructs

In this section, we provide examples of prototyping efforts that help define a construct prior to operationalization. While an aspect of a construct may theoretically be relevant to assess, if tasks cannot be designed to capture appropriate evidence, then claims cannot be made about test takers' proficiency in that aspect of the construct. The three examples of prototype studies in this section were designed to confirm that a new aspect of a construct can be measured. These prototype studies addressed issues related to item design and wording of directions. In addition to analyzing prototype task performance data, evidence of the new construct was obtained from questionnaires and interviews administered during the studies. Study designers also collected evidence of the feasibility of generating sufficient numbers of items for maintaining an ongoing testing program.

Measuring a new construct: pragmatic understanding

The TOEFL 2000 Listening Framework (Bejar *et al*., 2000) recommended that the listening skills measured in TOEFL iBT include a measure of pragmatic understanding (PU). One early prototype study investigated the feasibility of systematically measuring PU in the new test (Jamieson and Denner, 2000). PU was defined as understanding speaker purpose

(the function of an utterance), speaker attitude (or stance), and understanding whether an utterance was a fact or an opinion. One purpose of the study was to determine whether item types could be designed that measure this aspect of comprehension, and if so, whether examinees could answer them. Another purpose was to compare performance on PU items to performance on basic understanding (BU) items that measured main ideas and details and to compare the types of strategies participants used to respond to the BU and the PU items.

As described in Chapelle *et al*. (2008), six sets of items were developed using a variety of stimulus types (e.g., lectures, classroom discussions). Each stimulus was followed by 4–6 items: 2–3 BU items and 2–3 PU items. Although all the items were multiple choice, variations of the multiple-choice format were trialed. For example, one item type included replaying a portion of the stimulus in the stem of the item to help examinees focus on the pragmatic features of a particular part of the stimulus.

An instrument was developed to collect usability information about the extent to which participants could answer the questions (e.g., whether examinees understood the directions) and to collect the thought processes that participants used to respond to the items. Retrospective interviews were employed to obtain this information (see Bowles, 2010).

The study materials were administered to 20 participants – 15 NNSs, and five NSs – all of whom were enrolled in a mid-size university in the United States. The NNSs were placed into three ability groups according to their TOEFL paper-based test (PBT)[2] scores. This study was administered to a small number of participants because it was the first attempt to develop PU tasks systematically, and the probability was high that changes to the tasks would be made based on information from the test taker interviews.

Participants listened to the audio stimulus and responded to items for all six sets. The participants were then shown the items a second time, and an interviewer asked two questions about each item. The first question, "Did you understand what you needed to do to respond to the item?" was intended to confirm that the directions were clear and that participants knew how to respond to the task. The second question, "Why did you choose that answer?" was intended to identify which strategies the participants used to answer the questions. The strategies were later analyzed to determine whether the strategies were construct relevant or not. These interviews were taped and transcribed.

Directions and option format

The item directions were clear to the participants for the most part. A majority of the items had a multiple-choice format, with some variations that did not cause any confusion. However, one format had "audio options," where the options were not written on the screen but only presented orally. This format was attractive to some test designers as no reading ability was required to respond, and it was thought that it might measure listening ability more purely than other item formats. The first instance of this item format created some confusion for participants, but the second instance was not problematic for them. However, this item format was not pursued for both practical and construct-related reasons: listening to audio options took longer than reading written options, and comparable evidence was obtainable with the written option format.

Comparison to basic comprehension questions

Analyses were conducted to compare the mean scores on the BU and the PU items to see whether these item types differed in difficulty and to compare the types and frequency of strategies used to respond to these different item types.

Results indicated that both BU and PU items were difficult for the low-level ability group (based on TOEFL PBT scores), of medium difficulty for the mid-level group, and easiest for the high-level and native speaker groups. The total proportion correct was about the same for the two item types: 0.83 for all BU items and 0.82 for all PU items. However, the pragmatics items separated the mid-level ability group more effectively than did the BU items. There was a 23 percent difference in proportion correct between the low-level group performance and the mid-level performance on the pragmatics items and a 20 percent difference between the middle and high levels. With the BU items, there was a 30 percent difference between the low and middle levels and only a 6 percent difference between the middle and high groups, suggesting that perhaps something different was being measured and that these items could provide useful measurement information, particularly about examinees at the middle and higher levels of proficiency.

Strategy use

The retrospective interviews provided insight into the types and frequency of strategies used to respond to the items. For BU items, the most common strategies were to recall specific content, recall lexical information, recall general content, make inferences, and use a process of elimination. For the PU items, the most common strategies were to make inferences, recall lexical information, recall general content, recall the narrative, and interpret the intonation. Although there is some overlap of strategies used, the differences support the hypothesis that the pragmatics items provide evidence about different aspects of the listening construct than do BU items.

Establishing an operational definition of PU was achieved by reanalyzing what the 12 PU prototype items were measuring and the responses to the retrospective questions about these items. The outcome was a definition of PU that focuses on the understanding of speaker purpose (function of utterance), stance (speaker attitude), and recognizing whether a specific utterance is presented as a fact or an opinion. It was recommended that this definition of PU be confirmed and investigated further in a subsequent prototype study.

In sum, participants were able to respond to the pragmatics items, and most of the items were answered correctly by the majority of the participants. Moreover, they seemed to contribute to distinguishing between ability groups. The differences between BU and PU were demonstrated by the different strategies used to respond to the items that measured those abilities.

Measuring an expanded construct: higher-level reading abilities

Another major innovation of the TOEFL iBT was to investigate the feasibility of obtaining evidence for the purpose identified in the TOEFL 2000 Reading Framework (Enright *et al*., 2000) as "reading to learn." This purpose was further investigated and defined in Enright and Schedl (1999) as an academic purpose for reading that requires the reader to construct an organized mental framework of important information in texts. Reading-to-learn tasks require readers to show evidence of being able both to comprehend individual points and to use a framework such as a table or a summary to reconstruct the major ideas and important supporting information from the text.

With this goal in mind, the design team proposed three prototype tasks: (1) a schematic table completion, (2) a summary, and (3) an outline completion. To receive credit on the schematic table task (Figure 23.1), examinees needed both to select the important ideas

Directions: Complete the table below by classifying each of the answer choices according to whether it is characteristic of

1) **the applied arts, or**
2) **the fine arts**

Not all of the phrases will be used.

Applied Arts	• • •
Fine Arts	• •

Answer choices:

Object's purpose is primarily aesthetic.	Basic form of objects varies little across cultures.
Objects serve a functional purpose.	Artists work in concert with their materials.
Incidental details of objects do not vary.	Object's place of origin is difficult to determine.
Artists work to overcome the limitations of their materials.	

Figure 23.1 Schematic table

about the reading stimulus from a list of possible choices and to place them into the correct categories in a table. The summary task (Figure 23.2) required examinees to select from six answer choices three major ideas from the passage and place them in order in a summary format. The intended order of the second task type was the order that a good reader would use in summarizing and remembering the information (i.e., the order of the mental framework for remembering the information). For the outline completion task, answer choices were to be selected and placed into a partially completed outline in a specific order.

To investigate the effect of memory load on the construct of reading to learn, the design team proposed two versions of the reading-to-learn tasks: a variant that displayed the reading passage alongside the items and a variant that did not. In an academic context, reading to learn involves remembering information, so some members of the design team thought that participants should not be allowed to review the passage when responding to the reading-to-learn item. Other team members were concerned that it was unfair to withhold the text and that examinees would be unable to remember the information in a test situation.

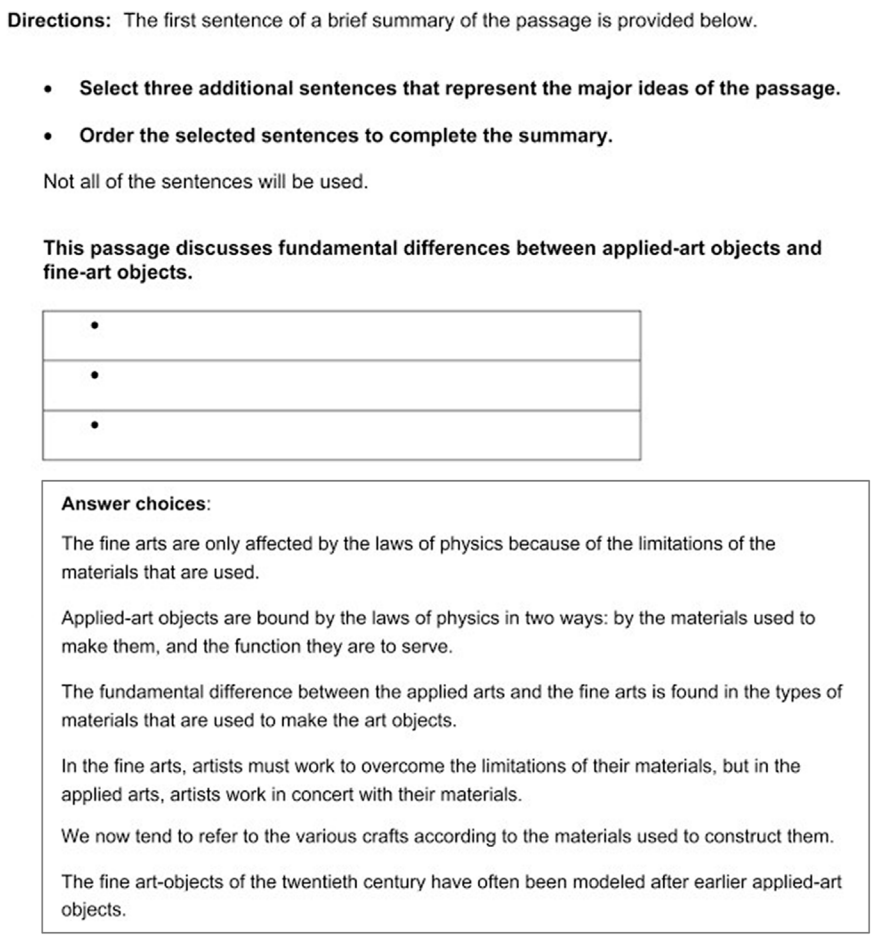


Figure 23.2 Summary task

These reading-to-learn tasks were prototyped with an initial group of 475 candidates who were selected to match the characteristics of the international TOEFL test population in terms of their academic status (graduate/undergraduate), gender, first language, and English ability, based on their TOEFL paper-based test scores. This sample was relatively large so that sufficient data about the multiple variables being investigated could be generated. The three task types were presented to comparable groups of candidates, and equal numbers of candidates of similar abilities were given the same test questions with texts either available or unavailable for review.

The goals of the initial round of prototyping were (1) to investigate the feasibility and sustainability of creating ongoing exemplars of these types of texts and tasks; (2) to investigate the viability of the task types and to see how well they discriminated among examinees, with and without text; and (3) to investigate the issue of how to score these more complex tasks, as it was thought that the reading-to-learn task types should be weighted more heavily than traditional four-option multiple-choice items because they would take more time and measure more global comprehension of the text.

Sustainability/replicability

A practical consideration for testing programs that require large numbers of forms is whether the test can be maintained over time. One question related to the prototype task types was whether an item pool could be created with passages of appropriate length and complexity to create the kinds of tasks described in the reading test framework (Enright *et al*., 2000). The design team recruited test development staff to create a small pool of passages and tasks that could be used in prototyping and to evaluate the feasibility of finding sufficient passages to sustain these tasks in the operational test.

Test developers used a passage evaluation form to track the lexical, grammatical, rhetorical, and topical components of the passages they considered for use and were able to create a pool of passages that was representative of the types of passages described in the reading framework (Enright *et al*., 2000). They concluded that sustaining such a pool of passages would be possible. However, they found it difficult to find passages in which the *order* of presentation of major ideas was critical for understanding. For example, in the summary task presented in Figure 23.2, the order of the three major points of the text is not critical. If examinees select the three major points (#1, #2, #4), one would assume they had understood the major text information. For texts describing a process, the order matters, but a pool of texts describing processes would not be representative of academic texts in general. Prototype performance of examinees also supported the view that this requirement, noted in the item directions, was not well understood and not likely to discriminate well.

Decisions made based on the reading-to-learn prototype study

Results of the initial round of prototyping indicated that the two groups under the different test conditions – passage available versus passage not available – did not differ significantly. Some candidates performed better with the texts available while others performed better without the texts. Therefore, the design team recommended access to the passage since this might make examinees more comfortable. The outline task was seen as focusing too much on details that could be better assessed by basic comprehension questions, and it did not continue to the next stages of prototyping and piloting.

Performance was analyzed for consistency of scores with ability levels, and different ways of scoring the items were considered. It was determined that these reading-to-learn task types should be weighted more heavily than the traditional four-option multiple-choice items because they take more time to complete, measure more global comprehension of the text, and call upon higher-level skills. Since prototyping demonstrated that few participants selected all the correct answers, the items would be too difficult if examinees needed to answer all parts correctly to receive credit. Thus, a simple partial credit model was used to score the initial prototypes, where the number of points awarded for a response task depended on the number of correct answers to be selected and the number of total answer choices. The details of awarding partial credit in the operational test were determined in subsequent studies and analyses.

Measuring an expanded construct: integrated skills

Prior to the launch of the TOEFL iBT, a comprehensive needs-and-task analysis was conducted to inform the design of the assessment (Rosenfeld *et al*., 2001). Analysis of academic

writing tasks in the target language use context supported the need to assess writing based on knowledge and experience. The analysis also indicated that a measure of academic writing and speaking should include both subject matter content and integrated language skills. Until the early 2000s, many direct writing assessments required examinees to write an essay that expressed and supported an opinion, but those tasks did not necessarily measure integrated skills or require accurate subject matter in the response. Similarly, speaking assessments also focused on supporting an opinion without reference to content information. Thus, the design teams conducted significant and iterative prototyping efforts prior to piloting and field testing the expanded construct.

Viability of measuring multiple skills with one or more integrated tasks

For an early round of prototyping, the design teams developed tasks consisting of a reading or listening stimulus with both selected response (SR) tasks and constructed response (CR) speaking or writing tasks. The goal was to explore the possibility of obtaining information about reading or listening ability in addition to writing or speaking ability. CR tasks presented a prompt that required "content responsibility" in the response. Content responsibility refers to evaluating responses based on the accuracy and completeness of the response, in relation to the stimulus material. This type of task broadens the target language use domain, specifically, to write or speak about academic content that is read or heard.

The design teams produced four variations of integrated skills SR+CR task type: reading-writing (RW), reading-listening (RL), reading-speaking (RS), and listening-speaking (LS). Reading stimuli included either a short (~350 words) or a longer (~650 words) passage, and listening stimuli included an excerpt from a university lecture or a campus-situated conversation. In designing these prototypes, the teams faced the following challenges: (1) creating a viable measure of integrated skills, (2) avoiding direct overlap between SR and CR items to minimize the potential for double jeopardy (i.e., penalty), and (3) ensuring items were appropriately difficult.

Regarding the first challenge, prototypes in the study elicited some evidence about listening and speaking abilities and about reading and writing abilities. However, the RS and RW prototypes did not yield relevant evidence about the construct. Since the reading passage was available during the speaking and writing response time, some examinees located a relevant portion of the passage and simply copied or read aloud from the text.

To address the second challenge, the design team prototyped tasks with SR items to assess one portion of the stimulus and CR items to assess another portion. The goal was to minimize the impact that responding to SR tasks might have on CR tasks. A review of responses from the prototype study indicated that shorter reading passages and conversations did not provide sufficient evidence of sustained writing ability. On a positive note, writing about the longer lectures provided evidence for the targeted claims about writing ability. However, the tasks about listening ability tended to focus on minor points rather than key information.

The design team found the third challenge difficult to resolve. On the one hand, stimulus texts for CR tasks should be accessible so examinees would be able to demonstrate their productive skill ability. However, to measure reading and listening ability, the stimuli required a level of complexity appropriate for the construct and testing population. These conflicting requirements for stimulus features were noted by designers.

Given these challenges, the design team did not recommend including this iteration of constructed response tasks in the high-stakes TOEFL iBT. However, these SR+CR tasks can

provide useful practice of language abilities, and thus a version of the prototyped tasks was included in a publication called *LanguEdge* (Powers *et al*., 2003).

Prototyping to inform the design of the TOEFL iBT writing section

The next phase of prototyping compared the RW and LW tasks with longer texts to a new prototype task that incorporated writing about information that was both read and listened to. The reading-listening-writing (RLW) task included a reading passage that presented an issue or an explanation, followed by a lecture that typically called into question the various points in the reading passage, and then a writing prompt that asked examinees to explain the points called into question. This task design attempted to capture evidence of academic writing that reflects analysis of multiple sources and synthesis of complex information.

This task could be replicated in multiple forms, and numerous reading passages and accompanying lectures could be created. Moreover, it did not seem to be possible to "coach" examinees for this task, as examinees needed to use information from both the oral and written texts in their written response. The topics discussed were academic, and even if there was some general familiarity, prior knowledge did not seem to assist with correctly completing these tasks. By using a prompt like "How did the information in the lecture call into question the information in the reading passage?" the advantage of copying from the reading text was minimized. The more proficient writing responses that were elicited included the kind of writing that was targeted in the construct as defined in the TOEFL 2000 Writing Framework (Cumming *et al*., 2000).

Evaluating alternative tasks that measure a defined construct

The previous section provided examples of extensive prototyping for new constructs and novel task types. However, prototyping for tried-and-true tasks is also recommended, particularly if those tasks are being adapted for a different delivery platform (e.g., computer, smartphone) or are being repurposed for different testing populations. Additionally, if some aspects of a test are being revised but the interpretive claims based on the test results are to remain the same, then prototyping can help verify that those revisions do not impact the validity of the measurement.

Evaluating enhancements to an existing construct

After the TOEFL iBT had been operational for a decade, recommendations were made for enhancing the writing section (Cumming *et al*., 2013). The recommendations included specifications to identify the intended purpose, audience, and genre of responses to the writing tasks and to replace the independent task with a content-responsible task. Four new or revised writing tasks were prototyped between 2014 and 2015: a revised reading-listening-writing (RLW) task, an online forum task, an email task, and a classification task. After several rounds of prototyping, the email task and the classification task were abandoned as they yielded insufficient evidence of the targeted writing construct. However, the revised RLW task and the online forum task showed promise. The revised RLW task specifies the audience, purpose, and genre of the task more clearly than the current operational RLW task since it is framed as a report to the class; it also focuses more on writing skills and less on cognitive skills not directly related to the writing construct (Cho and Choi, 2018). The online forum task also defines the intended audience, purpose, and genre of the task; it is

modeled on student contributions to online discussion forums. Like the current independent writing task, it asks examinees to convey their opinions, but it is also source responsive, requiring examinees to integrate information from existing posts. Scoring rubrics for the two viable tasks were drafted and revised.

These item types were not operationalized in the TOEFL iBT because they would require a change to the score scale. Revising score scales of high-stakes international assessments is disruptive to score users, examinees, and other stakeholders, and such revisions should be made judiciously. However, exploratory development continues, and when the opportunity arises, this item type may be utilized in the TOEFL family of assessments.

Alternative task contexts

Detailed item writing guidelines help test developers ensure that tasks presented in different test forms measure the same construct – i.e., are parallel to each other. However, overly constrained guidelines can have a narrowing effect on item generation and variation. Tests that require a large pool of unique items need to balance this tension carefully. One example of this balancing act comes from the Test of English for International Communication (TOEIC), specifically the TOEIC speaking test. In 2013, a study was conducted to revise the TOEIC speaking item writing guidelines to increase the contexts of a given communication task: proposing a solution to a problem. The purpose of the study was to verify that an increased variety of contexts (e.g., number of interlocutors, speaker roles, setting) in the audio stimulus would not affect measurement of the speaking construct.

The design team began by revisiting the original task design. The original task was to propose a solution based on a problematic situation presented in a voicemail message. As this type of communication task occurs in a wide variety of contexts in the real world, the design team proposed increasing the number of contexts for the item stimulus material. Thus, the goal was to investigate which variables relating to context would or would not affect the construct.

The team developed prototypes to explore (1) the number of interlocutors heard in the stimuli (one, two, or three), (2) the setting of the stimuli (face-to-face conversation, meeting, teleconference, voicemail message, telephone conversation, radio talk show, etc.), and (3) the timing of the response (synchronous, asynchronous, semi-asynchronous). The purpose of the study was not to verify the viability of a new part of a construct but rather to verify that the revisions to the guidelines would not affect construct measurement. Therefore, the items were prototyped with subject matter experts specializing in second language acquisition and English language assessment rather than students. The content experts found that telephone calls and meetings were reflective of the domain of the construct (spoken English in the workplace and daily life) and meaningful to score users as these two task types were most similar to the original context. Items for a pilot study were developed based on the recommendations from the subject matter experts and were administered to 992 English learners (ELs) in Korea and Taiwan. The results indicated that the increased variety in contexts would not affect measurement of the construct.

Keyboarding versus drag-and-drop responses

Even if a construct or task type is familiar to test users and examinees, prototyping can nevertheless provide insight, particularly when deciding between different technologies for capturing responses. As part of the launch of the TOEIC Bridge writing test in 2019, the

design team considered tasks to assess examinees' ability to produce grammatically correct sentences. One promising task type was based on a traditional classroom task that presents a sentence with the word order scrambled. The students must then re-sequence the words to form a grammatically correct sentence. Although the test was intended for beginning and low-intermediate learners (i.e., CEFR levels A1 to B1), the medium of the assessment, including the test instructions, would be English. As the assessment would be administered via computer, the challenge facing the design team was how to adapt the classroom task for computer-based delivery while ensuring the accessibility of the item directions (see also Sawaki, this volume).

The design team identified two plausible methods for capturing examinee responses: typing via standard keyboard and using a computer mouse to drag and drop words on the computer screen. As keyboarding would be more closely aligned with the writing construct, the team created prototypes that required examinees to read the words/phrases on screen and type the sentence in the correct sequence. Prototypes were administered to a small group of beginning-to-intermediate ELs.

Upon review of responses and participant interviews, the design team reached two conclusions. First, successful responses provided evidence that examinees could generate grammatically correct sentences. Second, the evidence that the unsuccessful responses provided was not always clear. Some unsuccessful responses received partial credit due to construct-relevant issues: namely, the examinees' limited writing proficiency. However, other responses failed as a result of construct-irrelevant factors. For example, some unsuccessful responses consisted of well-formed sentences that warranted low scores because they included words or phrases that were not part of the stimulus material.

To minimize the impact of construct-irrelevant variance, the design team decided to pursue the drag-and-drop task format. New prototype exemplars were developed, with each word or phrase appearing in a box that could be moved via computer mouse to targets on screen. While successful responses still provided evidence to support the construct, unsuccessful responses for the drag-and-drop prototype clearly indicated that examinees did not have the English language proficiency to generate grammatically accurate sentences. The drag-and-drop items moved forward to pilot and field-testing phases.

New testing populations

Even if a construct and task are well established, prototyping remains a good practice when a task is being considered for a new test population. One example comes from the launch of the TOEIC Bridge speaking test in 2019. The design team sought to verify that an item type in use since 2006 on the TOEIC speaking test[3] could be adapted for a less-proficient testing population (e.g., CEFR A1 to B1). Moreover, the test instructions would be provided in English and would need to be accessible to ELs with lower proficiency levels.

The design team looked to balance the linguistic load of the item directions with adequate scaffolding to support spoken responses. To this end, two types of prototypes were developed. The first set of prototypes utilized the same structure as the current TOEIC speaking test item type: a photograph was presented with open-ended directions (e.g., "In this part of the test, you will describe the picture on your screen in as much detail as you can"). The second set of prototypes displayed the same photographs as the first set, but the directions were modified to provide scaffolding for the task (e.g., "Look at the picture on your screen. Describe where the people are and what they are doing. Provide as much detail as you can").

Both sets of prototypes were administered to a small group of low-proficiency ELs who were then interviewed.

The design team came to favor the second set of prototypes. Responses at the highest band of the rubric (score band 3) for both prototypes provided evidence that examinees could use spoken English to describe people, objects, places, and activities. However, the second set of prototypes elicited a greater variety of unsuccessful responses that allowed for more nuanced differentiation among participants. For the scaffolded prototype, unsuccessful responses were categorized by a limited choice of vocabulary and use of structures at score band 2, by only using isolated words or phrases at score band 1, or having no English in the response at the lowest score band (score band 0). The scaffolded task directions moved forward to piloting and field testing.

Prototyping to inform development of scoring rubrics

To evaluate responses to new constructed response tasks, draft scoring rubrics are prototyped to investigate whether raters can use them to successfully differentiate examinees according to ability level. The draft rubrics are then refined by clarifying the distinctions between different levels or to change the number of levels, by decreasing the number if the responses cannot be sorted into the number of levels defined by the prototype scoring rubric, or by increasing the number if they fail to distinguish among examinees of clearly different abilities.

Scoring rubrics often undergo additional revisions after the prototype phase of design. When additional tasks are created, when responses from additional examinees are obtained, and when additional raters attempt to use a new rubric, refinements in band descriptions can be made, up to the first reporting of operational scores.

Developing a new rubric

Responses to prototype tasks can be utilized to develop an appropriate scoring rubric for the construct to be measured. This approach was employed to develop the scoring rubric for the independent and integrated tasks in the speaking section of the TOEFL iBT. For the speaking section of the TOEFL iBT, a total of five workshops were conducted. The initial four sessions were held for different small groups of experts: ESL instructors, oral proficiency raters, university administrators, and applied linguists. During the initial workshops, experts identified characteristics of proficient academic speaking, which expanded on initial definitions in the speaking framework (Butler *et al*., 2000). The groups (1) listened to responses to prototype tasks, (2) placed the responses according to proficiency level, and (3) identified the salient features of the responses at the different levels. The discussion of these features and the levels they represented were summarized in a preliminary scoring rubric.

The fifth workshop included one representative from each of the groups of experts who listened to responses, described their salient features, and attempted to use the draft rubric. At this session, the preliminary rubric was refined into a four-point scoring scale for the integrated tasks and a five-point scale for the independent tasks. Subsequent prototype efforts and pilot studies explored whether these scoring rubrics were appropriate (Chapelle *et al*., 2008), and the test launched with four-point scales. (See Harsch and Martin [2012] for a thorough description of developing a writing rubric.)

A similar approach was utilized to develop the scoring rubrics for the TOEFL iBT integrated writing tasks. Note that this approach does not automatically prescribe the number of

levels that will be on the operational scoring rubric. Rather, the number of levels is a direct result of the number of distinct bands or groupings of salient features of the responses that can be described by the panels of experts. Ultimately, this approach led to speaking rubrics with four levels and writing rubrics with five levels.

Adapting a rubric

Institutions that administer the TOEFL ITP[4] assessment series articulated the need for a speaking test. To explore the possibility of adding a speaking measure to the TOEFL ITP, a prototype study was conducted in 2019 to determine the viability of using existing task types and their scoring rubrics with a different population. Two task types from the TOEFL iBT and one task type from the TOEFL Junior[5] speaking test were selected. The tasks were administered to 50 examinees representative of the TOEFL ITP population to determine (1) the viability of these tasks for the TOEFL ITP population and (2) the appropriateness of the score bands defined on the TOEFL iBT and TOEFL Junior rubrics for the TOEFL ITP population.

The results indicated that the highest band on the TOEFL iBT speaking rubric was not relevant for the TOEFL ITP population and that additional distinctions were needed at the lower bands. Therefore, the TOEFL iBT speaking rubric was modified to a five-band rubric with more detailed levels identified for lower-ability examinees. Similarly, the highest band of the TOEFL Junior speaking rubric did not describe the highest ability level of the TOEFL ITP sample. As a result, the two lowest bands of the TOEFL Junior speaking rubric were combined into one band, and a higher band description was added. These draft rubrics will subsequently be used with a larger pilot study to confirm their viability for the TOEFL ITP population and the targeted tasks.

Summary and future developments

We began with a definition of prototyping and emphasized that prototyping should begin only when a clear and agreed-upon definition of the construct exists. The size of the prototyping sample is determined by the number and nature of the questions being addressed as well as by practical issues such as cost, time, and resources.

The examples of prototype studies illustrate how prototyping can inform decisions about the viability of measuring certain aspects of a construct and about the task types that provide the best evidence for a specific construct. A process for developing and trying out a scoring rubric based on prototype responses was described. The included studies investigated only some of the features that can be better understood through prototyping.

What new directions might language testing prototyping take in the future? The focus in this chapter has been on prototyping for the design of large-scale proficiency assessments. An important aspect for future prototype studies is to consider specific learning contexts. Initial, small-scale prototype efforts to inform teaching and learning might be carried out in the classroom in cooperation with, or led by, classroom teachers. The use of self-assessment, peer-assessment, and group work provide different kinds of opportunities for formative evaluation and diagnostic feedback. In the classroom context, new, innovative communicative tasks can provide teachers with information about students' instructional needs as well as assess how well students are able to use the information they have learned in class in communicative situations.

Prototyping related to the plethora of new technologies could address three areas: the design of new assessment tools (e.g., to support automated scoring engines and spoken dialog systems – see Xi, this volume), enhanced analyses of prototype results (e.g., using eye-tracking software), and prototyping task types that expand construct coverage afforded by technology innovations (e.g., video interviews). As the proportion of second language learners continues to increase throughout the world, a variety of technologies, learning situations, and tasks to meet the learning needs of this expanding and diverse population will continue to be researched and prototyped in the future.

Further reading

Chapelle, C., Enright, M. and Jamieson, J. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY: Routledge.
This volume provides a detailed description of the research and development efforts to revise the TOEFL from 1990 to 2005. By describing this evolution, key principles in educational measurement are explained. The volume integrates the results of empirical studies and validity arguments that support the TOEFL iBT. Early chapters of the volume present the rationales for the revisions and a description of the process used to define the construct. Middle chapters provide detailed accounts of numerous research studies and prototyping efforts that informed the design of the test. The volume concludes with a validity argument for the test.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment, An Advanced Resource Book*. New York, NY: Routledge.
In the first section of this valuable resource book, the authors review issues of validity, test constructs, models, and frameworks. They consider the relationship between these abstract models, the development of a test framework, and actual test specifications. They explain how items and tasks are written and prototyped. They also consider ethics and standards for testing, test validation, and use. The second section of the book includes excerpts from highly influential papers by experts in these same areas. The third section focuses on group activities related to the core concepts of the book, such as analyzing items and tasks, creating arguments for test validation, and writing an assessment statement for a test.

Wolf, M. K. and Butler, Y. G. (eds.). (2017). *English Language Proficiency Assessments for Young Learners*. New York, NY: Routledge.
This volume draws on examples of English language proficiency assessments for school-age children from the US and around the globe to highlight test development and validation processes. For those interested in examples of prototyping efforts, Chapters 3, 5, 7, and 10 may be of particulate note.

Notes

1 Examples of high-stakes tests for English language learners include such tests as TOEFL iBT and IELTS, tests of admission for foreign students to English language colleges and universities; TOEIC, a test of international workplace English; English tests required for doctors and nurses who trained in non-English-speaking countries and work in English-speaking countries, etc.
2 The TOEFL paper-based test, a multiple-choice test, measured abilities in reading, listening, and structure.
3 The TOEIC speaking test, an assessment of workplace English, measures beginning to advanced level (i.e., CEFR A2 to C1) learners.
4 The TOEFL ITP tests are administered to measure English proficiency at an institution, primarily for placement into an English language program and/or to monitor student progress.
5 The TOEFL Junior tests are intended for students aged 11 and older, primarily for placement and to monitor progress.

References

Bejar, I., Douglas, D., Jamieson, J., Nissan, S. and Turner, J. (2000). *TOEFL® 2000 Listening Framework: A Working Paper* (TOEFL® Monograph No. MS-19). Princeton, NJ: Educational Testing Service.

Bowles, M. A. (2010). *The Think-Aloud Controversy in Second Language Research*. London, UK: Routledge.

Butler, F., Eignor, D., Jones, S., McNamara, T. and Suomi, B. (2000). *TOEFL® 2000 Speaking Framework: A Working Paper* (TOEFL® Monograph No. MS-20). Princeton, NJ: Educational Testing Service.

Chapelle, C., Enright, M. and Jamieson, J. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Cho, Y. and Choi, I. (2018). Writing from sources: Does audience matter? *Assessing Writing* 37: 25–38. https://doi.org/10.1016/j.asw.2018.03.004.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly* 10: 1–8. https://doi.org/10.1080/15434303.2011.622016.

Cumming, A., Kantor, R., Powers, D., Santos, T. and Taylor, C. (2000). *TOEFL® 2000 Writing Framework: A Working Paper* (TOEFL® Monograph No. MS-18). Princeton, NJ: Educational Testing Service.

Enright, M., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P. and Schedl, M. (2000). *TOEFL® 2000 Reading Framework: A Working Paper*. TOEFL Monograph Series MS – 17. RM-00–4. Princeton, NJ: Educational Testing Service.

Enright, M. and Schedl, M. (1999). *Reading for a Reason: Using Reader Purpose to Guide Test Design* (Internal Report). Princeton, NJ: Educational Testing Service.

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing* 20: 384–408. https://doi.org/10.1191/0265532203lt265oa.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment, An Advanced Resource Book*. New York, NY: Routledge.

Harsch, C. and Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing* 17: 228–50. https://doi.org/10.1016/j.asw.2012.06.003.

Jamieson, J. and Denner, A. (2000). *Report on Prototype Project #7: Creation of Tasks and Observations for Listening Claim 2 – Listening for Pragmatic Understanding in a Variety of Text Types* (Internal report). Princeton, NJ: Educational Testing Service.

Park, E. and Bredlau, E. (2014). *Expanding the Question Formats of the TOEIC Speaking Test*. (ETS White Paper). Princeton, NJ: Educational Testing Service.

Powers, D. E., Roever, C., Huff, K. L. and Trapani, C. S. (2003). *Validating LanguEdge™ Courseware Scores Against Faculty Ratings and Student Self-assessments*. (ETS Research Report Series): i–25.

Rosenfeld, M., Leung, S. and Oltman, P. K. (2001). *The Reading, Writing, Speaking, and Listening Tasks Important for Academic Success at the Undergraduate and Graduate Levels*. TOEFL Monograph Series MS – 21. RM-01–03. Princeton, NJ: Educational Testing Service.

24

Pre-operational testing

Benjamin Kremmel, Kathrin Eberharter, and Franz Holzknecht

Introduction/definitions

Pre-operational testing is the systematic process of gathering evidence to support claims about the validity of a test's score use and interpretation before the operational use, through trialling and researching test tasks or items. It usually involves the stages of prototyping, piloting, and field testing (Fulcher, 2010). *Prototyping* (see Nissan and Park, this volume) can be understood as the creation and mini-trial of individual pre-production models of tasks based on specifications to investigate whether any assumptions might need to be reconsidered or refined. This can involve reviews and evaluations from subject experts, so-called *alpha testing* (Fulcher and Davidson, 2007), as well as administering the prototype task or item to a handful of candidates, which should be sampled from or similar to the actual target test population (i.e., *beta testing*). *Piloting* "refers to the process of trialling items with a larger group of people than would normally be used in prototyping" (Fulcher, 2010: 179; see also Read, this volume). While at the *beta prototyping* stage, mainly qualitative data will be collected and analyzed from a relatively small sample, piloting should also incorporate a quantitative component. The sample size will naturally depend on the test purpose, but it should be large enough (N = 30+) to allow for statistically meaningful conclusions. Piloting is often used to tentatively check the reliability, difficulty, and discrimination of items through item analysis, which can also be achieved using sub-components of the test, rather than administering the full test (Fulcher, 2010). Kenyon and MacGregor (2012) refer to this phase as exploratory in nature, as changes to the test and its features (e.g., scoring criteria, administration conditions) may still be viable. The widespread practice of additional pre-testing on native speakers during this phase can be considered contentious (Alderson *et al*., 1995), not least because of the ambiguity and controversy surrounding the term *native speaker* in today's world.

Field testing, on the other hand, is often considered "confirmatory, a final check that everything is working as intended and the test is ready to become operational within the context for which it is being developed" (Kenyon and MacGregor, 2012: 295). This stage is often only feasible for large-scale testing contexts, in that field testing involves the trialling of complete test versions on a large sample of the test population. This

DOI: 10.4324/9781003220756-32

allows for more robust item analyses, and therefore generalizations to the larger population, but also enables investigations into systematic test bias through differential item functioning (DIF) analyses, or dimensionality issues pertaining to the internal test structure (Fulcher, 2010). It is also the final opportunity to identify technical issues to do with test delivery or invigilation in case no prior operational acceptance or readiness testing has been conducted. It is worth highlighting, however, that due to the iterative nature of test design in general, these three stages will inevitably be iterative processes as well that are best conceptualized as dynamic or cyclical rather than linear (Fulcher and Davidson, 2007).

After these three stages of pre-operational testing (or iterations thereof), test developers decide whether they have gathered sufficient evidence to warrant the launch of the test for the intended purpose. This does not imply that validation research on the test or further development of the test will or should end at this point. Rather, it means that the test has undergone best-practice quality control before becoming operational and that it is now ready to be used for yielding useful information and inferences for score users.

Pre-operational testing is therefore key to all test development and thus features in all major guidelines for good assessment practice (e.g., EALTA, ILTA, ALTE) and handbooks on language testing and assessment. It is also a pivotal element of essentially all current validation frameworks, and we will illustrate the central role of pre-operational testing in two argument-based validation approaches.

Historical perspectives

While the view that pre-operational testing is critical for test development appears consensual and with no dissenting perspectives identifiable over the decades, there is a lack of scholarly literature on pre-operational testing as a research area itself, so it is difficult to point to the origins of the practice. References to "experimentally" administering a test or test parts to "a representative sample of students for whom the test is intended" (Lado, 1961: 329) can already be found in the earliest academic publications on language testing, which have established the field, but it is probably fair to assume that the process of trying out instruments predates these considerably as a standard practice in psychological measurement.

Forms of pre-operational testing have thus also featured strongly in guidelines for assessment practice from the beginning. The Guidelines for Practice of the International Language Testing Association (ILTA) have listed pre-testing of items and also the trialling of scoring guides as important "responsibilities of test designers and test writers" from their first inception (ILTA, 2007). The European Association for Language Testing and Assessment's (EALTA) Guidelines for Good Practice make test developers consider the questions "Are the tests piloted?" "What is the normal size of the pilot sample, and how does it compare with the test population?" "What information is collected during piloting (teachers' opinions, students' opinions, results . . .)?" "How is pilot data analysed?" and "How are changes to the test agreed upon after the analyses of the evidence collected in the pilot?" (EALTA, 2006). *The Manual for Language Test Development and Examining* by the Association of Language Testers in Europe (ALTE) also maintains that "some form of try-out is needed as test takers can still respond to items in unexpected ways" (ALTE, 2011). Likewise, the Standards for Educational and Psychological Measurement call for piloting and documentation of this phase of trying out newly developed items and tasks (AERA *et al*., 2014). While not a content area in itself, descriptions of some form of pre-testing have regularly featured

in many journal articles, even before these guideline documents, when new instruments or experiments were designed for research studies.

Depending on the nature and purpose of instruments, however, these activities and the research methods and data involved have been known to vary widely, as there is no one standard pre-operational testing procedure (Kenyon and MacGregor, 2012). This has resulted in terminological inconsistencies in the limited literature available. While Fulcher (2010) discusses prototyping, piloting and field testing as the three stages of pre-operational testing as outlined and adapted for this chapter, Fulcher and Davidson (2007) only distinguish between prototyping and field testing. Chapelle (2008) exclusively uses the term *prototyping* to refer to all pre-operational testing activities. Bachman and Palmer, initially subsuming all pre-operational testing procedures under "pre-testing" (1996), later refer to the same processes as the "trialling stage" (Bachman and Palmer, 2010), of which piloting may be understood as an early stage and field testing to occur at a later stage. Wise and Plake (2016), on the other hand, seem to use the terms *pilot* and *field testing* almost interchangeably. Wendler and Walker (2006) further use the term *item tryout*, which appears to correspond to *field testing* in Fulcher's taxonomy.

Fulcher's (2010) tripartite model of *prototyping*, *piloting*, and *field testing*, however, seems the most useful as it describes different forms and phases of pre-operational testing in detail and provides necessary distinctions between the three. It also offers clear labels that will be encountered in most journal articles, handbooks, or introductory textbooks on language assessment and technical reports and manuals of tests. Lastly, his definition renders it clear that pre-operational testing must go beyond examining psychometric or statistical properties of items and tasks to provide the evidence necessary to support claims about items and tasks functioning as intended. We therefore agree with Kenyon and MacGregor that quantitative item analysis alone is insufficient "to cover all claims about the inferences and decisions that may be made on the basis of a test performance" (2012: 296), but that pre-operational testing needs to involve a comprehensive investigation using mixed-methods in order to establish the basis for any valid and useful claims about a test's operational readiness and thus its score interpretation.

To illustrate the significance of pre-operational testing for valid inferences based on test performance, we will now discuss the integral role of pre-operational testing in two validation frameworks, with a particular focus on argument-based frameworks and how this stage might affect various aspects of the validation argument.

Critical issues and topics

As Alderson *et al*. (1995) have argued, "However well-designed an examination may be, and however carefully it has been edited, it is not possible to know how it will work until it has been tried out" (p. 73). Pre-operational testing is thus an essential phase in test development. Sireci and Zenisky (2016) see pre-operational testing as one of the main ways of reducing threats to validity, including surveying examinees about the test experience, anxiety, engagement, and other accompanying data collection to aid in gauging the appropriacy and functioning of the test. This section will thus outline why this phase is key to support claims about the validity of a test's score interpretations by exemplifying the significant role of pre-operational testing in two prominent argument-based validation models. What these two models share is Toulmin's (1958) argument structure that links claims to data by supporting warrants or counter-rebuttals with backing evidence. In both models, pre-operational testing is fundamental to providing these links in the validation arguments.

The pivotal role of pre-operational testing in Chapelle's (2008) validity argument is highlighted by the prominence of the term *prototyping* in the chapter titles of the edited volume that illustrates the framework on the new Test of English as a Foreign Language (TOEFL) exam. Pre-operational testing in this argument-based framework is critical for the *evaluation* inference, the *explanation* inference, the *generalization* inference, and the *extrapolation* inference (see Chapelle and Lee, this volume). In the *evaluation* inference, it contributes data and backing to claims about the appropriateness and interpretability of scoring rubrics and procedures and administration conditions (Enright *et al*., 2008a, 2008b). In the *explanation* inference, it addresses and supports questions of task difficulty, task characteristics, the fit of observed to predicted performance features, the internal structure of the test, and the construct relevance of task-related strategies employed by test takers (Enright *et al*., 2008a, 2008b). In the *generalization* inference, pre-operational testing may investigate facets of measurement that contribute to score variability or the effect of different factors in the assessment (e.g., number of tasks and ratings, rating designs, composite score reporting) on reliability (Enright *et al*., 2008b). For the *extrapolation* inference, Enright *et al*. (2008b) suggest that pre-testing (or *second phase prototyping* in their terminology) could examine the relation of candidates' performances on prototype tasks to those on other measures (e.g., self-assessment or teacher assessment).

In a similar vein, pre-operational testing can also be argued to be an integral part in validation through an *assessment use argument* (AUA) (Bachman and Palmer, 2010). Kenyon and MacGregor (2012) illustrated this by means of a case study already in the first edition of this handbook. They maintained that pre-operational testing can be, or rather should be, embedded in the AUA to support links between performance, assessment records, score interpretations, decisions, and consequences. According to them, the main benefit of situating pre-operational testing within an AUA is that it requires the test developer to explicitly state the claims and thus identify which evidence needs to be gathered and presented to buttress these.

For claiming *consistency* of assessment records, for instance, pre-operational testing will require checking the reliability of tasks, items, and scoring criteria already in the piloting phase as materials can still be refined at this stage. Only a test that yields sufficient reliability estimates in field testing would be considered ready for operational use and thus form a solid foundation or backing for the claim of consistent assessment records.

When linking test records to interpretations, pre-operational testing can be critical to contribute to the *meaningfulness* argument as the quantitative and qualitative data gathered before the operational launch would inform warrants about what examinees and raters do with and think about items and tasks. This would usually involve questions such as the comprehensibility of instructions and items, how engaging tasks and materials are, the appropriateness of the test administration including timing, and what amount and what kind of responses are generated by items and tasks. Depending on the purpose and context of a test, pre-operational testing will also address questions of *impartiality* (e.g., through bias analyses), *generalizability* (e.g., through needs analyses, expert or test taker judgments on the authenticity of materials), *relevance*, and *sufficiency* and therefore contribute backing or rebuttals to the interpretation level of the AUA. While Kenyon and MacGregor (2012) state that piloting (and thus implicitly the gathering of more qualitative data) may not be necessary for new replacement items for an existing test, it could be argued that such a traditional approach with a heavy emphasis on psychometric properties and correlational analyses may not be sufficient to establish a robust base for the validation argument of these items. We do agree, however, that the remit of pre-operational testing and the questions it seeks to answer

are contingent on the novelty of the assessment, the stakes involved, the resources available, and prior research or experience that forms the basis of the test.

Pre-operational testing may also play a role in addressing some of the warrants needed for supporting claims on the link between interpretations and decisions, as well as between decisions and consequences. While the traditional understanding of pre-testing may often be limited to investigating the statistical functioning of items, the notion of the AUA goes beyond this to subsume all research prior to operational use that is necessary to support claims at any level, including at the decision and consequences levels. At the decision level, it would focus "on examining alignment of decisions to be made to values and legal requirements of the stakeholder context, as well as the equitability to all stakeholders of decisions made" (Kenyon and MacGregor, 2012: 298). At the consequences level of the AUA framework, this would mean that developers would also need to investigate the beneficence of consequences before operational use. While Kenyon and MacGregor (2012) acknowledge that this may at the very least "call for a wider variety of qualitative methodological approaches" (p. 298), this expansion of the concept of pre-operational testing could also be seen as overly ambitious, not least because the impact and consequences of a test are likely to only be measurable and researchable once a test is in full operation. Thinking through the rationales, decisions, and intended and unintended potential consequences of a test is certainly necessary in the process of test construction but probably does not fall directly within considerations of pre-operational testing as defined at the beginning of this chapter.

With *field testing* specifically, a further critical decision needs to be taken, which pertains to the design or form of the field test. Wendler and Walker (2006) outline three main options: (1) embedding a section of new items in an operational administration of the test, (2) embedding individual items within an operational test form, or (3) field testing new items in a stand-alone administration. Each of these approaches has advantages and drawbacks. (1) and (2) are usually implemented so that the responses on the new items are not taken into account when determining and reporting test scores. They both guarantee large samples of items and of representative test takers and thus robust data. It may allow test developers to have more confidence in the field-tested items being taken seriously and answered with effort under live test conditions, unless the new items are easily identifiable as field test items or sections. There is, however, an ethical dimension to this kind of "item-seeding" that seems noteworthy. These new items may extend the testing time for some or all test takers, which may impact performance (e.g., fatigue). If administration time is not lengthened, it is disadvantaging candidates by unfairly increasing the number of items to be answered in the time given. If newly seeded items are used to replace items for the reasons outlined, there is a risk of jeopardizing the reliability of the assessment. Moreover, as these new items are of uncertain quality at the time of administration, it cannot be ruled out that flawed items or tasks might "throw" candidates and negatively affect their performance.

Option (3) does not suffer from the problematic aspects discussed for options (1) and (2) but often requires more resources (e.g., recruiting a new test population), and given that test takers know that it is a special administration rather than the real test, they may lack the motivation to answer with their best efforts. This, in turn, may create a challenge for the interpretability of the results gleaned from the pre-testing. Item difficulty statistics generated from such a low-stakes field test, for instance, may change considerably when it comes to the operational administration of items in a real high-stakes setting. A systematic increase in facility values might therefore need to be factored in by test developers in this scenario due to the lack of representativeness (e.g., in terms of motivation) of the field-testing sample. At its worst, test developers might face systematic efforts to play the

system by field test candidatures underperforming on purpose with the intention of affecting/manipulating the difficulty level of the operational test. It is also this option that is often associated with the most test security concerns, as new items are clearly identifiable as such and thus may be leaked by candidates (see section on practical recommendations in this chapter).

Current contributions and research

Not much seems to have changed since the last publication of this volume, where Kenyon and MacGregor note a "lack of literature directly addressing pre-operational testing" (2012: 301). Indeed, publications specific for the field of language testing that address issues of pre-operational testing are few and far between.

To this day, Chapelle et al. (2008) remains a significant and exhaustive account of how a validity argument and construct considerations can be concentrated to drive all stages of revising a major language test, the TOEFL. The chapters in the first section of the book provide an overview of the foundation work that went into developing the blueprint of the revised test. The second part of the volume includes three chapters that provide in-depth descriptions of the procedures involved in prototyping new assessment tasks (Enright et al., 2008a), assessment measures across the productive and receptive skills (Enright et al., 2008b), and, finally, the test as a whole (Huff et al., 2008). In one of the last chapters of the book, Wang et al. (2008) report on the outcomes of the final pre-operational field test. Each chapter clearly builds on the evidence-based approach laid out by the initial sections of the volume as they include a set of guiding questions linking each phase and the evidence that is collected back to the intended use and interpretation of the test scores. Thus, the chapters illustrate ways in which test developers can address the key issues and collect the necessary evidence at the stage of pre-operational testing in order to refute the rebuttals to the validity claims of a newly developed or operational test.

Since the publication of Chapelle *et al*. (2008) and the previous version of this chapter, which called for more publication of pre-operational testing practices (Kenyon and MacGregor, 2012), others have published accounts and descriptions of pre-operational testing. In a collection of papers first presented at the Georgetown University Round Table 2016 (J. M. Davis *et al*., 2018), Lin and MacGregor (2018) illustrate with a number of case studies how particular questions pertinent to pre-operational testing can be addressed effectively. Their account encompasses two pre-operational testing studies, a small-scale study into the role of delivery mode (handwritten versus keyboard), and a larger study of score comparability across computer-adaptive, semi-adaptive, and print versions of a test. The authors also discuss the kinds of statistical analyses that can be carried out once a test is operational and needs to be monitored.

In 2015, Cambridge Assessment dedicated issue 59 of the quarterly published *Research Notes* to shedding more light on the test development processes established for Cambridge English language tests. In this issue, Corrigan and Crump (2015) illustrate the item development of two reading items, including various pre-operational versions and accompanying statistics from pre-operational tests as well as item performance in the operational test. Showing how item difficulty may vary with small changes to an item and supplementing this with data from pre-operational and operational test runs is rarely done. However, we feel that publications of this kind are useful as illustrative examples of the benefits of pre-operational testing for those test developers currently working on establishing best-practice procedures at their institutions.

The development and piloting of new testing procedures is also the focus of the three-part IELTS partnership research series (Berry *et al*., 2018; Nakatsuhara *et al*., 2016, 2017) as well as the chapter by L. Davis *et al*. (2018). The three-part series reports on three related phases aimed at exploring the validity and feasibility of assessing speaking via online conferencing technologies. The first phase investigated test taker performance and perception as well as examiner behaviour through a small-scale experiment with 32 test takers and four examiners (Nakatsuhara *et al*., 2016). Phase 2 reports on a follow-up to the first experiment and seeks to answer similar questions of test comparability across delivery modes, but on a larger scale (99 test takers, ten examiners) and employing more refined statistical analyses (Nakatsuhara *et al*., 2017). Finally, in their third report, the authors present research focusing on questions of the usability and scoring validity of a speaking examination delivered via a developed online platform (Berry *et al*., 2018). It is particularly noteworthy for this large-scale research effort that each phase built on the results of the previous and employed a mixed-methods approach. Similar in its outset to the IELTS partnership research series, L. Davis *et al*. (2018) also report on a pilot study investigating the technical feasibility and participant perception of computer-mediated speaking assessments. While exploratory and concise in nature, their study clearly demonstrates some of the basic steps that can be taken if test developers seek to establish new delivery modes and testing procedures.

Knoch and Chapelle's (2017) article on validating rating processes within an argument-based framework also outlines the potential contribution of pre-operational testing efforts to validation arguments. For each of the six potential inferences in the validity argument, the authors compiled a list of warrants and assumptions that test developers might want to address, and they provide suggestions as to how evidence can be gathered to support each of them, making it a useful additional resource when planning pre-operational testing for the productive skills.

While there are recent publications that make pre-operational procedures visible and discuss related issues, we would argue that there is still more scope for publications in this area. The apparent lack of literature addressing pre-operational testing systematically, despite the fact that a lot of these activities are likely undertaken by many test developers, may well be due to a range of reasons. First and foremost, it might not occur to practitioners entrenched in their daily work of developing and running a language test that the methods they employ and difficulties they face are worth sharing with a broader audience. Each test development unit is likely to have had to overcome various practical issues connected with pre-operational testing, requiring innovative approaches. From the researchers' as well as practitioners' perspective, even those approaches that proved less successful might be worth knowing about. Publishing this kind of research makes the design decisions, and the rationales for those decisions, transparent and open to independent scrutiny, which is a standard that we would and should expect of any research and development process.

Another reason little is published on pre-operational testing might be the fact that despite it being a crucial step in test development, not all test developers can be assumed to pre-test their tasks. When the structure of an examination does not allow for item seeding or when a test is undergoing a major review, field tests, which can be demanding in terms of resources, need to take place. However, the risks of exposing new tasks to a larger population and the fact that psychometric item properties in pre-operational tests might vary considerably from those from live administrations might also deter test developers from investing in pre-operational testing efforts.

Finally, the apparent lack of literature might to some extent also be caused by quite practical constraints. The diverse and inconsistent terminology surrounding pre-operational

procedures in and of itself makes locating relevant contributions difficult. Also, the fact that research has taken place in the context of pre-operational testing is rarely put into the foreground and is often treated as a methodological side note. However, as argued here, this should not be tolerated by the field as these are challenges that can be overcome.

Main research methods

Depending on the specific assessment context, pre-operational testing seeks to answer a multitude of different research questions relating to areas including, but not limited to, test content, test methods, test length, test taker perceptions, and test difficulty, and hence it necessitates the use of a variety of research methods. Although the choice of research methods is determined to some extent by the type of test (e.g., task level versus item level, standalone versus seeded/embedded, pen and paper versus computer-based), ideally pre-operational testing research should comprise both quantitative and qualitative methods in order to mitigate common limitations associated with either approach. While quantitative methods allow for large sample sizes and often result in generalizable findings, research questions can only be investigated superficially without substantiation on a more individual level. On the other hand, although qualitative methods offer detailed and personal insights of different stakeholders, findings are often not generalizable due to small sample sizes. By combining the two approaches in mixed-methods designs during the *prototyping*, *pilot-testing*, and *field-testing* stages, language testers can have more confidence in the results and thus develop more valid assessment instruments.

It would be beyond the scope of this chapter to outline all the different quantitative and qualitative research methods in detail (for a comprehensive overview of methods in second language research, see Gass and Mackay, 2005, 2007); however, the use of questionnaires in combination with stimulated recalls for pre-operational testing will be described briefly to illustrate how they could be combined to gain evidence for a validity argument. The major advantage of questionnaires is that different kinds of data can be collected from a large sample of respondents. Pre-operational testing questionnaires can be useful both in the *pilot-testing* stage – for example, by asking stakeholders attitudinal questions about the suitability of a new test format or the comprehensibility of instructions – as well as the *field-testing* stage by including factual items on biodata and language background to ensure that the field-testing sample is similar to the target population of the operational test. Questionnaires can also be used to collect behavioural data: for instance, by asking test takers about their use of test-taking strategies to ensure that newly developed items do not elicit test-wise behaviour. However, as questionnaires are measuring these aspects only indirectly, qualitative methods such as stimulated recalls could be used simultaneously on a smaller sub-sample to get direct insights into stakeholders' attitudes and behaviours. This way, test developers can triangulate their findings and achieve more robust results. Although the use of questionnaires in pre-operational testing may seem straightforward, practitioners need to ensure that questionnaires are developed thoroughly and that the data is analysed appropriately. Good resources are Brown (2001) and Dörnyei and Taguchi (2009), who describe in detail the various steps in questionnaire development and analysis. For detailed information on the stimulated recall method, see Bowles (2010) and Gass and Mackey (2000).

One crucial step in pre-operational testing is to investigate the psychometric properties of newly designed items, tasks, task formats, and rating scales. This generally happens in the *field-testing* stage, as large samples are usually necessary to ensure sufficient statistical power. The methods used in this stage differ between the receptive skills and the productive

skills. For the receptive skills, psychometrics methods are used to establish whether test items are reliable, of acceptable difficulty, of equal difficulty for different groups of test takers (e.g., females and males, test takers from different backgrounds, etc.), and discriminating well between weaker and stronger test takers. Only if new items pass these quality-control criteria can they be used for operational testing. The two most commonly used methods to establish statistical performance of test items for the receptive skills are classical test theory (CTT) and item response theory (IRT, or modern test theory). Although both methods have advantages and disadvantages (see also the chapters by Brown and Ockey in this volume), the underlying statistical assumptions of IRT are more robust than those of CTT (Magno, 2009). While CTT is mostly based on correlational analysis, IRT is informed by probability theory and takes into account both item difficulty and person ability to calculate person and item fit statistics (Kaplan and Saccuzo, 1997). Therefore, IRT "makes it possible to estimate sample-free item difficulty and item-free person ability" (Green, 2013: xiii). The results of CTT analyses, on the other hand, are strongly dependent on the sample (Hambleton, 2000; Magno, 2009) but can still be very useful if the field-testing sample is truly representative of the target population for the operational test. In order to perform meaningful IRT analysis, larger sample sizes are generally needed (200 or more participants are usually suggested in the literature), which can be challenging to achieve for smaller-scale testing programs. Whichever statistical method is used in pre-operational testing for the receptive skills, it is crucial that only those items with acceptable psychometric performance are used for operational testing.

For the productive skills, in order to determine the difficulty of newly developed tasks as well as their reliability and discrimination, it is necessary to investigate task performances and ratings of performances in conjunction with rating scales and rating procedures. This is again best accomplished by employing both quantitative and qualitative research methods. For example, in order to determine whether a newly developed writing test and rating scale can be used for operational testing, a research plan could involve many-facet Rasch measurement to study rater reliability and scale functioning, analyses based on generalizability theory to determine how many ratings and tasks are necessary to achieve a satisfactory level of generalizability for the intended test purpose, and interviews or introspective studies with scale users to get first-hand insights into the scale's user friendliness (see also Knoch and Chapelle, 2017). At the same time, the task performances could be looked at through different kinds of text analyses to investigate, for instance, in what ways handwritten responses differ from machine-typed responses (Lin and MacGregor, 2018) or to study whether writing tasks elicit types of performances in line with test developers' expectations.

As these examples show, there are many methods and combinations of methods that practitioners can use in pre-operational testing. An established validation framework can help identify the most important questions and evidence needed prior to designing an effective pre-operational research agenda.

Recommendations for practice

As we have highlighted in this chapter, pre-operational testing is complex and can vary greatly for different assessment contexts. The form of pre-operational testing necessary for a particular exam depends on factors such as the type of test, the intended test purpose, and the target population. Perhaps the most limiting factor of pre-operational testing, however, is practical constraints. Piloting and field testing are expensive and take time, which makes comprehensive pre-operational testing difficult to achieve for many smaller-scale

testing programmes. However, we would argue that putting enough resources into thorough pre-operational testing is crucial for making claims about the validity of test score interpretations. Even if there are not enough resources available for large-scale *field testing*, *prototyping* newly developed tasks with subject-matter experts and subsequent *piloting* on a small sample (30+) are generally achievable for most testing boards or even classroom contexts and already go a long way in avoiding some of the pitfalls associated with test development. That said, for high stakes tests, *field testing* on a large enough sample representative of the target population is essential to ensure that all items in the operational test have sound psychometric footing.

In this section, we present some practical advice on pre-operational testing. Not all the questions raised here will be applicable to every assessment context, and the list does not claim to be exhaustive. The recommendations are rather intended as general practical advice to help test developers think through some of the important aspects of pre-operational testing.

Planning

Due its complexity, drawing up a clear plan for all stages of pre-operational testing (*prototyping*, *pilot testing*, and *field testing*) is key. The plan will differ depending on various factors such as test purpose, high stakes/low stakes, etc.; however, some important questions at this stage are:

1 Are the aims of the pre-testing effort clearly defined by research questions?
2 Will the intended research methods generate useful data to answer all the research questions?
3 Are the resources sufficient for the intended form of pre-operational testing (budget, timing, personnel, rooms)?
4 Is the intended sample for pre-operational testing truly representative of the target population for the operational test? If not, are the attributes in which the populations differ likely to impact the outcomes?
5 Is the intended sample size large enough to allow for meaningful statistical analyses? (As a general rule, we would recommend a minimum of 30 participants for pilot testing, 100 for field testing if CTT is used and 200 if IRT is used.)
6 Will enough items and tasks be field tested to ensure sufficient numbers for the operational test? (A certain proportion of items and tasks will inevitably fail to pass the quality criteria.)

Administration

The procedures of pre-operational test administration are in many ways similar to those of operational testing. However, it is important to consider and try out all administrative procedures at this stage to ensure smooth test administration once the test is operational.

General aspects

1 Are there clear and standardized guidelines for test administration?
2 Are test administrators trained?
3 Are test administrators required to sign non-disclosure agreements (especially for high-stakes tests)?

4 If questionnaires are used for data collection, have they been thoroughly developed and piloted on a representative sample?
5 Is there a clear procedure for reporting back issues to do with test administration?

Administration of pen-and-paper tests

Although computer-based testing has become more and more common, there are still a large number of pen-and-paper tests, including high-stakes tests such as school leaving exams. Certain practical aspects of pen-and-paper testing are very different from computer-based assessment, and testing boards should therefore consider the following questions:

1 Do test takers use pens which are clearly visible on the paper and do not erase easily?
2 Are all materials posted securely (especially for high-stakes tests)?
3 Have all materials been sent back or are test booklets – or even pages thereof – missing?
4 Is there sufficient room to store all materials?
5 Is the technical equipment at the test administration centres up to date so it does not impact test performance (e.g., headphones, speakers, microphones, video players, etc.)?

Administration of computer-based tests

As outlined earlier, computer-based testing differs in some important aspects from pen-and-paper-based testing (see Sawaki, this volume). Coombe highlights that "in computer-based assessment, field testing is also conducted as a stress test" (2018: 21), in which questions such as the following should be investigated:

1 Is the testing platform functioning as intended?
2 Are there enough computers available to accommodate all test takers?
3 Are the computers up to date so they do not impact test performance (e.g., hardware, software, headphones, speakers, microphones, video players, etc.)?
4 Are recorded performances captured at high enough quality to ensure that raters can grade them reliably (sound or video quality)?
5 Are there any problems with internet connectivity at the test administration centres (for online tests)?
6 Is the server capacity sufficient?

Data retrieval, analysis, and storage

In most cases, pre-operational testing involves the collection of large amounts of data. Even for smaller-scale testing projects, the data collected can be quite varied and may include video-recorded interviews with subject-matter experts, responses to a short online questionnaire, and the test responses themselves. Managing all this data efficiently is important and can be achieved by considering questions such as the following:

1 For pen-and-paper tests: Can the data be converted to a digital format in an efficient way (i.e., can answer sheets be scanned, or do responses need to be entered manually)?
2 For computer-based tests: Does the platform used for test administration allow for efficient and reliable data retrieval on item and task levels?
3 Are the methods of data analysis appropriate?

4 Can the different data sources be cross-compared and triangulated?
5 Is the data stored securely (especially for high-stakes tests)?

Documentation and dissemination

As we have highlighted in previous sections of this chapter, published articles focusing on pre-operational testing are still sparse. This is unfortunate as reporting on procedures concerned with pre-operational testing is a valuable opportunity to build trust with stakeholders and contribute knowledge and experience to the wider expert community. We therefore recommend that test developers also consider the following questions during pre-operational testing:

1 Have all steps of pre-operational testing been documented thoroughly?
2 Are there plans to share experiences and expertise gained through pre-operational testing with stakeholders and other test developers?
3 Are there plans to disseminate some of the key findings in the form of publicly available technical reports or peer-reviewed journal articles?

Future directions

In this chapter, we have outlined the purposes and challenges of pre-operational testing. In line with the authors of its previous edition published in 2012, we reiterate that pre-operational testing plays a pivotal role in professional test development and is an integral part of building a validity argument for an assessment instrument.

While a need for more research and publication of pre-operational testing seems to have begun to be recognized by recent publications (see earlier in this chapter), we argue that more effort still needs to go into presenting and publishing insights and results from pre-operational testing. With hardly any publications clearly dedicated to this area, the field is still far away from conducting "debates in the academic literature weighing the pros and cons of alternate approaches and discussing the most efficient and effective methods," as Kenyon and MacGregor had hoped in their outlook almost ten years ago (2012: 305). In order to be able to conceptualize a theoretically sound and effective pre-operational testing endeavour, publications that focus in particular on the various challenges of delivering pre-operational testing within the time and resource constraints of actual test development are still needed. In a way, pre-operational testing thus remains an overlooked and underappreciated aspect of test development to this day. In the long run, however, establishing a firm research agenda and dedicating work to this area might constitute a beneficial contribution to the field as a whole as it has the potential to help build professional practices across various language-testing contexts. For this to be the case, such research would have to keep the needs and questions of those colleagues who produce language tests with very limited resources in mind and address them in a meaningful way. Making pre-operational testing more visible and working toward scalable solutions to problems many practitioners share is also likely to bring pre-operational testing to contexts where it is currently not established – be it for security concerns or a perceived lack of predictive power of pre-operational testing data.

We feel that the increasing digitalization of language assessment has the potential to drive the field's interest in pre-operational testing, from small-scale piloting to large-scale field testing. Due to rapid improvements in internet connectivity in many parts of the world, language tests are likely to move toward online delivery soon if they are not delivered online

already. Stakeholders' perceptions of online examinations, as well as their ability to interact with digital tests, may vary considerably and have an effect on the use and interpretation of test scores. As many practitioners face the challenges of digital innovation, this development could serve as a great opportunity to address the current research and publication gap and could be a starting point to address and innovate questions of pre-operational testing in many contexts.

Further reading

Bachman, L. and Palmer, A. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.
This work describes the assessment use argument at length and helps contextualize the importance of pre-operational testing for various elements of a validation argument at large.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.
This edited volume details the pre-operational testing for the revision of a large-scale, high-stakes test and discusses the implications for various inferences in the validation argument.

Fulcher, G. (2010). *Practical Language Testing*. New York, NY: Routledge.
The author offers an entire chapter on the phases of pre-operational testing and provides plenty of examples and practical advice in expanding Fulcher and Davidson's (2007) initial description of pre-operational testing.

Lin, C. and MacGregor, D. (2018). Using a validation framework as a guide for planning analyses and collecting information in preoperational and operational testing. In J. Davis, J. Norris, M. Malone, T. McKay and Y. Son (eds.), *Useful Assessment and Evaluation in Language Education*. Washington, DC: Georgetown University Press, 201–16.
This paper is based on the previous version of this handbook chapter and exemplifies how pre-operational testing issues can be situated in a validation argument.

References

AERA [American Educational Research Association], APA [American Psychological Association], NCME [National Council on Measurement in Education], Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Alderson, J. C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

ALTE [Association of Language Testers in Europe]. (2011). *Manual for Language Test Development and Examining for Use with the CEFR*. Strasbourg: Language Policy Division. Retrieved from www.coe.int/t/dg4/linguistic/ManualtLangageTest-Alte2011_EN.pdf.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford, UK: Oxford University Press.

———. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

Berry, V., Nakatsuhara, F., Inoue, C. and Galaczi, E. (2018). Exploring the use of video-conferencing technology to deliver the IELTS Speaking Test: Phase 3 technical trial. *IELTS Partnership Research Papers*, 2018/1. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Retrieved from www.ielts.org/teaching-and-research/research-reports.

Bowles, M. (2010). *The Think-Aloud Controversy in Second Language Research*. New York, NY: Routledge.

Brown, J. D. (2001). *Using Surveys in Language Programs*. Cambridge, UK: Cambridge University Press.

Chapelle, C., Enright, M. and Jamieson, J. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. New York, NY: Routledge. https://doi.org/10.4324/9780203937891.

Coombe, C. (2018). *An A to Z of Second Language Assessment: How Language Teachers Understand Assessment Concepts*. London, UK: British Council.

Corrigan, M. and Crump, P. (2015). Item analysis. *Cambridge English: Research Notes* 59: 4–9.

Davis, J. M., Norris, J. M., Malone, M. E., McKay, T. H. and Son, Y. (2018). *Useful Assessment and Evaluation in Language Education*. Washington, DC: Georgetown University Press.

Davis, L., Timpe-Laughlin, V., Gu, L. and Ockey, G. (2018). Face-to-face speaking assessment in the digital age: Interactive speaking tasks online. In J. Davis, J. Norris, M. Malone, T. McKay and Y. Son (eds.), *Useful Assessment and Evaluation in Language Education*. Washington, DC: Georgetown University Press, 115–30.

Dörnyei, Z. and Taguchi, T. (2009). *Questionnaires in Second Language Research: Construction, Administration, and Processing* (2nd ed.). New York, NY: Routledge. https://doi.org/10.4324/9780203864739

EALTA [European Association for Language Testing and Assessment]. (2006). *Guidelines for Good Practice in Language Testing and Assessment*. Retrieved from www.ealta.eu.org/documents/archive/guidelines/English.pdf.

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R. N., Mollaun, P., Nissan, S., Powers, D. E. and Schedl, M. (2008a). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright and J. M. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language ™*. New York, NY: Routledge, 97–144.

Enright, M. K., Bridgeman, B., Eignor, D., Lee, Y.-W. and Powers, D. E. (2008b). Prototyping measures of listening, reading, speaking, and writing. In C. A. Chapelle, M. K. Enright and J. M. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language ™*. New York, NY: Routledge, 145–86.

Fulcher, G. (2010). *Practical Language Testing*. New York, NY: Routledge.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. New York, NY: Routledge.

Gass, S. and Mackey, A. (2000). *Stimulated Recall Methodology in Second Language Research*. Mahwah, NJ: Lawrence Erlbaum Associates.

———. (2005). *Second Language Research: Methodology and Design*. Mahwah, NJ: Lawrence Erlbaum Associates.

———. (2007). *Data Elicitation for Second and Foreign Language Research*. New York: Routledge.

Green, R. (2013). *Statistical Analyses for Language Testers*. New York, NY: Palgrave Macmillan.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care* 38: 60–65.

Huff, K., Powers, D. E., Kantor, R. N., Mollaun, P., Nissan, S. and Schedl, M. (2008). Prototyping a new test. In C. A. Chapelle, M. K. Enright and J. M. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language ™*. New York, NY: Routledge, 187–226.

ILTA [International Language Testing Association]. (2007). *The ILTA Guidelines for Practice*. Retrieved from www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf.

Kaplan, R. M. and Saccuzo, D. P. (1997). *Psychological Testing: Principles, Applications and Issues*. Pacific Grove, CA: Brooks Cole Publishing.

Kenyon, D. M. and MacGregor, D. (2012). Pre-operational testing. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. Abingdon, UK: Routledge.

Knoch, U. and Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing* 4: 1–23. http://doi.org/10.1177/0265532217710049.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. London, UK: Longman.

Lin, C. and MacGregor, D. (2018). Using a validation framework as a guide for planning analyses and collecting information in preoperational and operational testing. In J. Davis, J. Norris, M. Malone, T. McKay and Y. Son (eds.), *Useful Assessment and Evaluation in Language Education*. Washington, DC: Georgetown University Press, 201–16.

Magno, C. (2009). Demonstrating the difference between Classical Test Theory and Item Response Theory using derived test data. *The International Journal of Educational and Psychological Measurement* 1: 1–11. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1426043.

Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E. (2016). Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery. A preliminary comparison of test-taker and examiner behaviour. *IELTS Partnership Research Papers*, *1*. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Retrieved from www.ielts.org/teaching-and- research/research-reports.

———. (2017). Exploring performance across two delivery modes for the IELTS Speaking Test: Face-to-face and video-conferencing delivery (Phase 2). *IELTS Partnership Research Papers*, *3*. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Retrieved from www.ielts.org/teaching-and-research/research-reports.

Sireci, S. G. and Zenisky, A. L. (2016). Computerized innovative item formats. In S. Lane, M. R. Raymond, and T. M. Haladyna (eds.), *Handbook of Test Development*. New York, NY: Routledge, 315–34.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Wang, L., Eignor, D. and Enright, M. K. (2008). A final analysis. In C. A. Chapelle, M. K. Enright and J. M. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language* ™. New York, NY: Routledge, 259–318.

Wendler, C. and Walker, M. (2006). Practical issues in designing and maintaining multiple test forms for large-scale programs. In S. Downing and T. Haladyna (eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, L. L. and Plake, B. S. (2016). Test design and development following the Standards for educational and psychological testing. In S. Lane, M. R. Raymond and T. M. Haladyna (eds.), *Handbook of Test Development*. New York, NY: Routledge, 19–39.

25

Piloting vocabulary tests

John Read

Introduction

This chapter differs from the other two in Section 7 of the book in two respects: first, it focuses on assessing a particular area of language knowledge – vocabulary – and second, it deals with the somewhat ill-defined topic of pilot testing. By contrast, the other chapters (Nissan and Park, this volume; Kremmel *et al*., this volume) range more widely over types of items and tasks, and they deal with steps in the test development process that are better defined in language assessment.

Let us take up the latter point first. Various authors (see, e.g., Bachman and Palmer, 1996: 248; Moussavi, 1999: 284) have pointed out that there is a confusing array of terms which are partly if not fully synonymous – *pilot*, *pre-test*, *trial* and *tryout* – referring to a particular event or activity, along with the corresponding set of verbal forms – *piloting*, *pre-testing*, etc. – that refer to a more general process. In their dictionary, Davies *et al*. (1999) have separate main entries for *pre-test* and *trial(ling)*, but they acknowledge that the terms are not clearly distinguished in the literature. Thus, unfortunately, as in so many other areas of applied linguistics, we have to live with a lack of standardization in the terminology that is used.

Fulcher (2010), Kenyon and McGregor (2012) and Reed (2014: 877) all make a helpful distinction between pilot testing and field testing on two bases. One is that pilot testing is typically conducted on a smaller scale than field testing, and the other is that pilot testing occurs at an earlier stage of the test development process, often with the purpose of trying out whether innovative formats or tasks will work effectively. By contrast, field testing occurs after the overall structure and content of the test have been determined and is conducted on a larger scale. This distinction has also been adopted by Kremmel, Eberharter, and Holzknecht (Chapter 24, this volume), and so I will use the terms *pilot test* and *pilot testing* according to this way of defining them. It is an appropriate way to describe the work on developing an innovative vocabulary test format to be discussed later in this chapter.

It is also necessary to define the scope of vocabulary testing. Vocabulary is obviously an indispensable component of language use, to the extent that almost any kind of language

 DOI: 10.4324/9781003220756-33

test is in a sense measuring the test takers' vocabulary ability, even if this ability is not an explicit focus of the scoring or rating. I distinguish between "discrete" and "embedded" assessments of vocabulary knowledge and use (Read, 2000: 8–10). The embedded type involves a focus on vocabulary as part of the assessment of a broader construct, as when a reading comprehension test includes some items that measure whether the test takers can understand particular words or phrases that occur in the reading passages or when the raters in a speaking test evaluate the range and accuracy of vocabulary use as one of the criteria for rating each candidate's performance. However, normally, vocabulary tests are a discrete form of assessment, in that they are intended to focus specifically on some aspect of vocabulary knowledge or use, based on a pre-selected set of words. This chapter will be limited to the discussion of discrete vocabulary testing in this sense.

Historical perspective

Vocabulary tests in the multiple-choice format were the classic type of test item in foreign language tests during the middle decades of the twentieth century when, at least in the US and its international sphere of influence, the purely psychometric approach to test design and analysis was at its height (Spolsky, 1995). For example, I have traced elsewhere (Read, 2000: 138–47) the history of vocabulary assessment in the Test of English as a Foreign Language (TOEFL) from 1964 until the late 1990s. In the original version of the test, multiple-choice vocabulary items presenting the target words in isolation or very short sentence contexts were highly reliable and efficient measures which correlated well with the reading section of the test. It was only under external pressure from applied linguists and language teachers that the TOEFL program progressively moved to more contextualised formats for assessing vocabulary knowledge, and eventually, it was embedded in the assessment of reading comprehension. Numerous other large-scale language proficiency and achievement tests also took advantage of the good psychometric qualities of multiple-choice and other types of objectively scorable vocabulary test items.

This means that a number of the early textbooks on language testing (e.g., Lado, 1961; Heaton, 1988) have substantial sections on vocabulary tests, including a range of sample item types and advice on both the writing and piloting of items. Such coverage of vocabulary testing is less of a standard feature in more recent texts, with some exceptions being Hughes (2003: Chapter 13) and Brown and Abeywickrama (2010: Chapter 11). In her book on classroom assessment of younger learners, Jang (2014) emphasises the importance of assessing vocabulary knowledge in the development of academic language proficiency.

More generally, piloting is a somewhat hidden element of test design and development, in the sense that it is mentioned only in passing in most published papers on vocabulary tests. For instance, in a short article for teachers on the Vocabulary Size Test, Nation and Beglar (2007) state that "[a] lot of care was put into making the distractors [of the multiple-choice items] so that they were genuine choices and they were carefully checked in pilot testing" (p. 12). However, there is no specific information on how many people participated in the pilot and who they were, what the procedure was, what kind of feedback was obtained, or what changes were made to the draft items.

Piloting conventional test items

The basic principle of most conventional vocabulary test items is to require the learners to show that they can make a link between a word in the target language and its meaning, as

expressed by a synonym or definition in L2, a translation equivalent in L1, or a picture. For this purpose, the most common format is probably the familiar multiple-choice type, with its numerous variants:

I was astonished to hear what happened.

a. eager b. quite sad c. very pleased d. very surprised

Sonia has always been a friendly person, but recently she has been _____ kind and generous.

a. mostly b. particularly c. primarily d. fully

The other widely used format is word-definition matching, as in:

1	environment		
2	principle	_____	close detailed study of something
3	response	_____	money received regularly
4	factor	_____	surrounding area or conditions
5	analysis		
6	income		

These item types require the test takers to select responses from the options given. In the piloting of such items, the kind of questions we need to consider are:

- Is there in fact only one correct response for each item?
- Is the correct response clearly expressed and not easily guessed by being longer or noticeably more specific than the other options?
- Is each of the non-correct options functioning effectively by attracting at least some of the less-proficient test takers?
- Is the item at an appropriate level of difficulty?

Apart from selected-response items, there are others which require the test takers to supply a response, the most common of which is the gap-filling (or completion) task. In gap-filling items, the simplest case is that the target word is deleted from a single sentence.

Land that can produce good crops is _____.
They went out to _____ the people on the sinking boat.

A gap-filling sentence needs to provide the clues needed to figure out what the missing word is. Thus, in the second sample item, the test takers need to know what *sinking* means in order to have a reasonable chance of supplying the appropriate word. Sometimes the sentences are more or less definitions, as in the first example above, rather than the more natural sentential context for the word represented by the second item. Gap-filling items may also be created from whole paragraphs or longer texts by selectively deleting content words at regular intervals. These are often called “cloze” tests, although strictly speaking, the standard cloze procedure involves deleting both content and function words at a fixed interval of, say, every seventh word.

The main issues with gap-filling items are:

- Is it possible for the test takers to supply an acceptable response based on the contextual information provided?
- Given that very often more than one word can fill a blank, which responses actually supplied by the participants in a trial are correct? For the second example given previously, the intended answer may be *rescue*, but words like *save*, *help* and *assist* can presumably also be accepted. But what about *welcome* or *see*? A pilot test can elicit a range of responses to be evaluated so that decisions can be made as to which words will be included in the answer key. If the intention is to elicit just one target word, the sentence will need to be revised to try to rule out other possible answers; alternatively, the first letter of the target word may be supplied.
- Do the words supplied by the test takers need to be correct in terms of grammar and spelling? For instance, can some credit be given for *rescuing* or *reskue* in the earlier example? One can argue either way on this issue, but the point is that it should be resolved at the piloting stage so that those who score the operational test can be given clear guidelines.

Piloting a new format: word associates

Vocabulary test formats that are used in operational tests for assessment purposes, such as those discussed in the previous section, tend to be variations on a theme. The more innovative types of test are likely to be developed and used by researchers interested in investigating how vocabulary knowledge expands (or is lost), how words are stored and accessed in the brain, and other research questions of that kind. If they show some initial promise, these research instruments are often explored by a number of researchers in succession as measures in their own right. This can be seen as a process of "serial piloting," in which a format is tried out in several variants and in a variety of educational contexts, with the result that, over time, the potential of the format is thoroughly explored.

Devising the format: the initial studies

I can illustrate this process by reference to a type of vocabulary test that I devised in 1990: the word associates format. The test is based on the concept of word association, which has long had a significant role in psychological and psycholinguistic research. The basic research technique is a simple elicitation procedure: participants are presented with a series of words one by one and asked to supply the first word they think of in response to each stimulus word. For native speakers, the responses to particular words are remarkably stable, and there are published norms for word associations in English (Palermo and Jenkins, 1964; Deese, 1965). On the other hand, research by Meara (2009) and his students in the 1980s showed that second language learners produced responses that were much less consistent or interpretable, so the standard technique seemed an unpromising basis for either gaining an understanding of the learners' lexical networks (which was Meara's concern) or assessing the depth of their vocabulary knowledge (my own interest).

Working with Meara in London at the time, I took an alternative approach to my research objective by creating a test format that required learners to *select* words associated with particular stimulus words rather than supplying them. After drafting an initial set of items,

discussing them with a group of graduate students, and trying them out with a class of ESL students in London, I settled on the following item format:

edit			
arithmetic	film	pole	publishing
revise	risk	surface	text

Four of the words were associates of *edit* (*film*, *publishing*, *revise*, *text*), whereas the other four were distractors. There were three types of relationship between the target word and the associates:

- paradigmatic: the two words were either synonyms or similar in meaning, and belonged to the same word class: e.g., *edit* and *revise* (both verbs)
- syntagmatic: the two words could occur one after the other, as a collocation: e.g., *edit* and *text* (verb + noun)
- analytic: the associate represents one aspect of the meaning of the target word and is likely to form part of its dictionary definition: e.g., *edit* and *publishing*.

One of the first things I found out from the initial pilot was the test takers needed to know what kinds of association they were looking for; thus, for subsequent piloting, I prepared a practice sheet containing instructions and sample items which illustrated the three types of relationships. I went on to write two forms of a 50-item word associates test based on samples of items from the University Word List (Xue and Nation, 1984).

After my return to New Zealand, I took a number of steps to pilot the test further. I asked ten EAP teachers at the English Language Institute (ELI) of my university to work through the items individually and record what they thought were the correct answers. I made two interesting observations: some of the teachers reported that the test took longer to complete than they expected, and often, they could find two or three associates for an item quite readily but had more difficulty in finding the fourth one. The latter observation strengthened my belief that test takers should be told in the instructions that there were four associates to be found in each item. Of the 100 items, 27 were revised because more than one of the teachers chose a distractor or could not find one of the associates. Five of the teachers checked the revised items and no further problems were found.

Both forms of the revised test were administered to over 100 students in the intensive EAP programme at the ELI on separate days. The full details of the analysis can be found in Read (1993), but briefly, the two forms proved to be reliable (KR-20: .92), and they correlated reasonably well with a longer test of the UWL words that used a matching format (Form A: r = .76; Form B: r = .81). A Rasch partial credit analysis showed that 12 students were "misfitting" on at least one form of the test, in the sense that their response patterns did not fit the norm established by the pilot participants as a whole. An inspection of the test papers indicated that most of these test takers scored well overall but did not respond to certain items, presumably because they did not know the target word. This suggested that willingness to guess might be a significant factor in the learners' response behaviour.

Another step in the piloting was to ask eight learners individually to work through a selection of the items and provide a verbal report on why they chose each of their responses. This proved to be quite revealing. Those who were less proficient tended not to attempt any responses to an item if they did not recognise the target word. However, some more

proficient learners were not so inhibited, as in this explanation by one of them for his responses to the "denominator" item:

denominator			
common	develop	divide	eloquent
fraction	mathematics	species	western

Researcher (R): Do you know this word?
Learner (L): I don't know this word . . . but I'm trying to look at all the words that . . . what are the similar in meaning because I know that four words can go together . . . [Pause] I'm not sure that my idea will be right . . .
R: You've circled 'fraction' . . . 'divide' . . . 'species' . . .
L: Or 'develop' . . . So do you want my reason?
R: Yeah.
L: Because I don't know this word before and I tried to looking for some words that have similar in meaning. I think it's a guess – 'fraction,' 'divide,' 'species' – all things are cut in smaller – yeah – and then it develop bigger amount . . . but I don't know . . .

(Read, 1993: 366)

The learner applied a strategy that allowed to him to identify two of the four associates correctly while emphasizing that he was just guessing. Another learner was able to choose three of the four associates of *diffuse* in a similar manner. Thus, the verbal report evidence called into question the inference I wanted to make at the item level: that the selection of associates was a measure of how well the target word was known.

In order to address some of these concerns, I undertook a second study based on a redesigned item format (Read, 1998). Since the original study had shown that words entered into somewhat different semantic relations according to their parts of speech, the new format was intended specifically to assess knowledge of adjectives. The associates and distractors were presented in two clusters, with possible paradigmatic and analytic associates on the left side and possible syntagmatic associates on the right, as in this example:

conscious

awake healthy knowing laughing	face decision effort student

In this case, there were two associates on the left (*awake* and *knowing*) and two on the right (*decision* and *effort*), but in other items, the pattern was either one and three or three and one, in order to reduce the prospects of successful guessing. The variant patterns also took account of the fact that the different uses of some adjectives could be better represented by three collocates (e.g., *broad – river*, *broad – shoulders*, *broad – smile*), whereas three distinct meanings of *secure* were appropriately expressed through three paradigmatic associates: *confident*, *fixed*, and *safe*. A draft set of 46 items was tried out with a class of 16 Japanese learners and with four language teachers. Based on the problems revealed by this feedback, six items were discarded and another four were revised. This illustrates a useful guideline for piloting: that a draft test should contain more items than are needed for the

operational version so that poorly performing items can be removed without making it necessary to pilot replacement items.

Two trials were conducted with students in an intensive EAP programme, both involving concurrent measures of the learners' knowledge of the target adjectives. In the first trial, 84 learners took both the word associates test and another test that required them to match words and definitions. The correlation between the two tests was quite high, at .86. In the second, smaller trial, 15 learners not only took the two written tests but also participated in an interview based on the Vocabulary Knowledge Scale (Wesche and Paribakht, 1996) to probe their knowledge of the target adjectives in greater depth. The results showed that the interview scores correlated more highly with the matching test (.92) than the word associates test (.76). Although these findings were based on a small sample, they suggested – along with some other evidence – that the revised format was still affected by the test takers' willingness to guess what the associates of the target word might be.

Two North American studies

Since these two initial studies, a number of other researchers have carried out further piloting of the word associates format, adapting it for different learner populations and investigating variations in the design features.

Qian (1999) adopted an edited version of the Read (1998) test as his depth of vocabulary knowledge (DVK) measure in a study of the relative contributions of breadth and depth of vocabulary knowledge to reading comprehension performance. His participants were 74 Chinese and Korean students taking an intensive ESL course at a university in Canada. He conducted his own piloting of the test, which resulted in the replacement of eight of the original items because of concerns that some of the associates were not unambiguously correct.

Subsequently, the same DVK format was used in a larger-scale project (Qian and Schedl, 2004) to investigate the possibility of including the word associates format as an efficient measure of vocabulary knowledge in the Test of English as a Foreign Language (TOEFL). The original target words were replaced by 40 adjectives that had been tested in the multiple-choice format in an earlier (pre-1995) version of TOEFL. In this case, there were 207 ESL students from a much wider range of language backgrounds, again taking an intensive university course in Canada. A time limit of 35 minutes was set, reflecting the time taken by the test takers in Qian's 1999 study, and almost all the test takers were able to complete the items within that period. After the test had been administered, 30 randomly selected participants were interviewed about their experience of taking it. All but two of them reported that they were able to understand the nature of the test task once they had read the instructions, and an inspection of all the test papers showed that only 10 test takers had responded inappropriately by, for instance, selecting more than four responses for numerous items. Furthermore, all except one of the interviewees stated that they were reluctant to guess what the associates were for items with a target word that was unfamiliar to them. The researchers concluded that the DVK test was comparable in quality to the existing TOEFL multiple-choice vocabulary test and was worthy of further evaluation. One concern they raised was that the TOEFL item writers had found it difficult to select nouns as syntagmatic associates that were indisputably the correct responses, which would have been a problem in producing multiple forms of the test. However, in the end, the DVK did not fit the design of the new internet-based TOEFL and was not put to operational use.

Extending the format: French learners in the Netherlands

A number of researchers in the Netherlands have adapted the word associates format for their purposes and have explored several variations in the basic design. In particular, Tine Greidanus and her associates (Greidanus and Nienhuis, 2001; Greidanus *et al.,* 2004, 2005) saw the format as a potentially useful means of assessing the vocabulary knowledge of Dutch students who were advanced learners of French. Greidanus and Nienhuis (2001) considered it more appropriate for such learners to write word associates items in which the distractors as well as the associates were semantically related to the target word. They selected a sample of 50 words from the most frequent 5,000 words in French and wrote pairs of items like this for *rive* (bank):

rive	*artificiel* (artificial)	***rive***	*bord* (bank)
	bord (bank)		*fleuve* (river)
	côté (side)		*gauche* (left)
	fleuve (river)		*paquet* (parcel)
	gauche (left)		*prudent* (careful)
	vague (wave)		*tombe* (tomb)

First, note that there are just three associates and three distractors for each item, so each associate represents one of the three semantic relationships identified in my original study (Read, 1993): paradigmatic, syntagmatic, and analytic. The three associates in both of these items are *bord*, *fleuve*, and *gauche*; the difference is that the distractors in the item on the left are related in meaning to the target word to some degree, whereas the distractors in the item on the right are quite unrelated.

Preliminary versions of the two 50-item tests (each with half related and half unrelated items) were tried out with four university teachers of French, two of whom were native speakers. Items were modified when these judges did not select the intended associates. Clearly, the use of semantically related distractors increases the likelihood of debate over whether the associates are really more acceptable responses than the distractors. The revised tests were taken by 54 undergraduate students at two universities in Amsterdam. Collectively the items with semantically unrelated distractors were significantly easier than those with related distractors, as might be expected. The researchers also calculated separate scores for the syntagmatic and paradigmatic associates and found the means for the paradigmatic associates were always significantly higher than those for the syntagmatic ones. This finding, which was confirmed by Greidanus *et al*. (2005), contrasted with Qian and Schedl's (2004) study, in which the means for the two types of associate were virtually the same.

In a follow-up study, Greidanus *et al*. (2004) introduced a new variable: a differing number of associates in the items. They initially prepared test items for 65 French words, 22 with two associates, 34 with three associates and 9 with four associates. In order to ensure that they considered each of the 6 words in an item, the test takers were required to explicitly accept or reject each word as an associate, and their total score included both the number of associates and the number of distractors correctly identified. Another innovation was to include students who were native speakers of French in the trialling. The test was piloted twice, and after each pilot, the researchers applied item analysis statistics to the responses of the French-speaking students as the basis for revising or discarding items containing associates that were not selected frequently enough or distractors that proved too attractive.

In both trials, the mean scores showed a consistent pattern of variation according to the frequency level of the words (from first 1,000 to fifth 1,000), and the proficiency level of the students (e.g., third- to second- to first-year Dutch university students).

The third study (Greidanus *et al.,* 2005) built on the earlier ones in various ways. The items assessed less common vocabulary, in the fifth 1,000 to tenth 1,000 range of French words, and the participants included not only French and Dutch students but also British students studying French at the University of Edinburgh. Once again, the distractors were semantically related to the target words. The authors piloted a preliminary version of the test with five university teachers of French and three university students and found that distractors in 18 of the items were too attractive as possible answers and needed to be replaced with less appealing alternatives. In fact, two versions of the test were produced: one with a fixed number of associates for each item and the other with a variable number of associates. The results showed that there was essentially no difference in performance on the two versions, which puzzled the authors a little because they had hypothesized that the fixed-number version would be easier since it could make it easier to identify associates by a process of elimination. This result suggests that these test takers took each item on its merits and mostly did not adopt the kind of strategic approach to responding that the researchers expected.

Adapting the format for Dutch children

Two other Dutch researchers, Schoonen and Verhallen (2008), applied the word associates format in a rather different educational context. Their interest was in assessing the deep word knowledge of primary school children in the Netherlands. In an earlier study (Verhallen and Schoonen, 1993), they used an elaborate interview procedure to demonstrate that children with Dutch as a first language had not only larger vocabularies but also richer semantic concepts associated with individual words than children for whom Dutch was a second language. However, the interview procedure was time consuming, and they wanted to find a more practical alternative. They adapted Read's (1993) word associates format, first by reducing the number of associates and distractors from eight to six and then presenting them visually as a "word web," as in Figure 25.1:

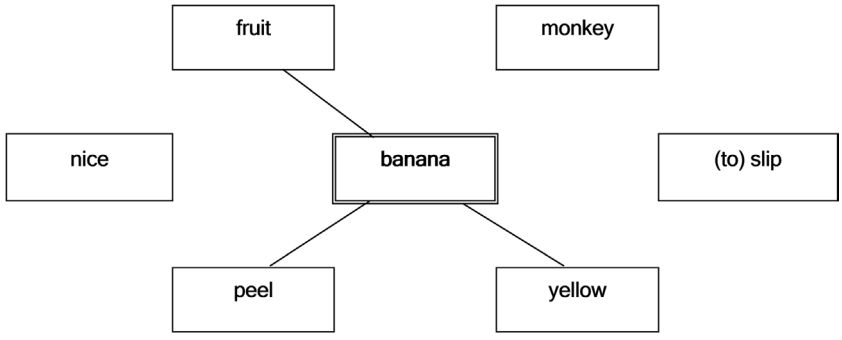


Figure 25.1 Sample item from the WAT (translated from Dutch)

Source: Schoonen and Verhallen, 2008: 219

The target word *banana* is at the centre, and the test takers' task is to draw lines from there to the three associates, as shown in the figure. Note that, like Greidanus and her colleagues, Schoonen and Verhallen decided that the distractors should also be semantically related to the target word. They argued that depth of knowledge should be conceived as the ability to define words in the decontextualized manner that is important for school learning. Thus, the associates represent the core meaning of the word, whereas the distractors are related to it in a more context-specific way. It is true that bananas are "nice" to eat, for humans and for "monkeys," and circus clowns tend to "slip" on banana skins, but the associates *fruit* (paradigmatic), *peel* (syntagmatic), and *yellow* (analytic) are more basic elements of the meaning. The children had to mark all three associates correctly in order to be given one point for the item; there was no partial-credit scoring.

The researchers wrote an experimental test comprising 27 items and conducted a pilot study with 133 children in Grades 3 and 5. In order to introduce the test to the children, the test administrator worked through two practice items with each child on a blackboard to discuss the kind of relationships they should look for and to prompt the child to draw the lines. In addition to analysing the responses, the researchers recorded children "thinking aloud" as they worked through the test items to gain a better understanding of the strategies the children were following in selecting their responses. Some items were revised, and more were added to make a total of 50, which were divided into two test forms of 25 items each because of a concern that a 50-item test might be too long for nine-year-old children, the youngest participants in the study. The full study involved 822 nine- and eleven-year-old students in 19 primary schools in the Netherlands, who took the test in class groups in school after receiving standardized instructions from a test assistant.

More research on the design features

Subsequently, Schmitt *et al*. (2011) went back to the Read (1998) format, using adjectives as target words, and they conducted two pilot studies which represent the most comprehensive investigation of design features of the word associates format to date. In Study 1, they wrote their own 40-item test of adjectives from the Academic Word List. After taking the whole test, the participants, who were 18 EAP students in Japan, were interviewed to elicit an independent measure of their knowledge of ten of the target words. The results provided support for Read's (1993) original concerns about the validity of item scores as measures of knowledge of the target words. On the one hand, test takers with scores of 0 or 1 were shown to have little if any knowledge of the word in the interview, and those who scored 4 generally had good knowledge. However, item scores of 2 or 3 could indicate anything from no understanding to good knowledge of the target word. Overall, in nearly half the cases, the WAT appeared to be overestimating the learners' knowledge of the target words. In addition, the researchers recorded think-aloud protocols as the test takers worked through ten further items. From their analysis of the strategies that the learners adopted, they concluded that a score of 3 or 4 was a good indicator that the test takers had at least partial knowledge of the target word meaning; otherwise, they were more likely to be making guesses based on knowledge of word parts, loose semantic associations, or connections among the associates and distractors.

The second study by Schmitt *et al*. (2011) used another set of academic English adjectives as target words. The participants, 28 international students at the University of Nottingham, took 20 word associates items as a pen-and paper test and worked through another

14 items in an interview, which broadly confirmed the findings in Study 1. The authors suggest that the format "appears to be better at indicating what learners do know rather than what they do not know" (2011: 11). Test takers who had no knowledge of target word (according to what they revealed in the interview) tended to make inferences based first on what they knew about the meanings of the associates and distractors and then on the orthographic form of the words. Other variables in this study were as follows:

- Both six-option and eight-option items, which seemed to perform at a similar level.
- Three different scoring methods. The "all-or-nothing" method used by Schoonen and Verhallen (2008) worked better for six-option items, whereas awarding one point for each associate correctly identified seemed preferable for eight-option items. The more complex method used by Greidanus *et al*. (2004) did not appear to be worth the extra effort involved in applying it.
- Three types of relationship between the target word and the distractors: unrelated, related in meaning, and related in form (e.g., *special* and *obvious* as distractors for *spurious*). The form-based distractors did not perform satisfactorily at all. On the other hand, the semantically related distractors worked better with six-option items, and the unrelated distractors gave better results with the eight-option items.
- The distribution of the associates between the paradigmatic and syntagmatic types: 1–3, 2–2, or 3–1. The researchers found the 2–2 pattern produced the highest correlation between the word associates and interview results. They suggested that the 1–3 pattern was more susceptible to successful guessing because three nouns that could all collocate with a particular target adjective tended to be similar to each other in meaning.

A recent study by Zhang *et al*. (2017) is one of the first to apply the word associates format to the assessment of a non-European language: Chinese. The authors modelled their research on Qian's (1999) study by comparing an associates test they developed with an independent measure of vocabulary knowledge as predictors of the reading comprehension ability of foreign learners of Chinese. In addition, they investigated two design features of the format. First, they extended the finding of Greidanus *et al*. (2005) by showing that the learners performed better on the test when they knew there was a fixed number of associates for each test item; they tended to miss associates when informed that the number varied from one item to the next. Secondly, the authors confirmed the conclusion of Schmitt *et al*. (2011) that the all-or-nothing scoring method worked best for test items consisting of three associates and three distractors, as compared to the other two methods.

In this study, there were a couple of issues that were specific to the Chinese language. One was a decision to use only two-character (compound) verbs and adjectives as the target vocabulary items since these are much more common than one-character words in modern Chinese. Although the authors were careful to select vocabulary that would be known to the participants, they also acknowledged a potential problem that could occur if a test taker did not recognise one of the characters in a compound word, despite being familiar with the word in its spoken form. This would block the learner's access to the word meaning in a way that does not arise in the testing of vocabulary in languages with alphabetic writing systems.

Conclusion

Thus, the word associates test is a very interesting case of an apparently simple test format which turns out to raise multiple design issues and has generated a considerable amount of

research involving at least four target languages (English, French, Dutch, and Chinese) and a variety of learners in several different settings. In some respects, the various studies can be regarded as investigations of the validity of this kind of test rather than pilot studies in the conventional sense, but they are generally small-scale projects that address the same sort of questions as – and adopt similar methods to – those that should be used in piloting new types of vocabulary tests. One interesting point about all this work on the word associates format is that, to the best of my knowledge, none of the studies has yet led to the adoption of the format as a vocabulary measure in a large-scale operational test. Nevertheless, collectively, the research provides very useful guidelines for any test designer considering the use of the format for an operational assessment purpose, although, of course, it does not obviate the need to pilot any such test in the educational context for which it is intended.

Let us review the word associates literature as a way of summarising the key elements involved in piloting vocabulary tests.

1 Test developers are faced with multiple decisions in determining the particular form that their test will take: the criteria for selecting the target words, the presentation format of the items, how many response options there will be, whether the distractors will be related in meaning or form to the target word, whether there will be the same number of associates for each item, and how the responses will be scored.
2 Once a draft set of items has been written, it is the normal practice to ask a number of native speakers or other proficient users of the target language – particularly language teachers – to work through the items, to see if they can readily identify the intended responses and spot any problems with the way the items have been written.
3 The next stage in piloting is to administer a revised version of the test to learners who are similar in educational background, target language proficiency, and other respects to the intended population of test takers. No single pilot study can address all the potential issues, and so more than one is likely to be needed if the test is to be used for formal assessment purposes.
4 Since the test takers will almost certainly have had no previous experience with the word associates format, it is essential to brief them thoroughly through written and oral instructions so that they understand what kind of relationships they are looking for, and the adequacy of the briefing needs to be monitored as part of the pilot. There should also be some practice items for them to work through.
5 In addition to administering the pen-and-paper test on a group basis, it is desirable to ask some test takers to provide individual verbal reports as they work through the items, in order to gain insights into their response behaviour.
6 A trial with a reasonable number of learners (at least 30, but the more the better) will generate the data required to estimate the reliability of the test and to apply item analysis statistics to the individual items.
7 Another helpful means of evaluating the pilot test results is to have some independent measure of the test takers' knowledge of the target words, through giving a separate written test in a different format or by interviewing the learners individually.

Steps 4 and 5 are not as necessary with more standard test items, such as the multiple-choice and gap-filling types, but the other steps can be applied to any kind of vocabulary testing. The higher the stakes involved in the assessment for the test takers the more important it is to ensure that the test material is adequately piloted before it is used for decision-making purposes. However, even in the lower-stakes environment of the language classroom,

teachers can engage in a form of serial piloting by treating the first administration of a test that they give to their students as a pilot, which provides the basis for revising and improving the test progressively as it is used with subsequent groups of students.

Future developments

The basic procedures just outlined for piloting vocabulary tests are unlikely to change a great deal in the foreseeable future, but the tests themselves and the contexts in which they are used will be influenced by a number of developing trends. The first is the now-routine use of computers to administer language tests. The item types that are typically chosen for vocabulary tests lend themselves well to computer delivery, which allows for rapid scoring and analysis of the test takers' responses. In addition, the computer can be programmed to record data which is useful at the piloting stage, such as test taker keystrokes, changes made to responses, and the time taken to respond to particular items or the test as a whole.

The precise measurement of reaction time may play a role in another emerging area: diagnostic language assessment. Diagnosis is receiving fresh attention as a purpose for language assessment (see, e.g., Alderson *et al*., 2014; Jang and Sinclair, this volume), and, although various approaches are being explored, it is likely that measures of vocabulary and grammar knowledge will play a significant part in identifying areas of weakness needing attention by learners and their teachers. Traditionally, test takers have been assessed on their ability to recognise and supply the correct language forms without time pressure, but from a diagnostic perspective, learners' speed of access to their lexical knowledge is also important for fluency in all four skills. Cremer and Schoonen (2013) used a computer-based adaptation of the Schoonen and Verhallen (2008) WAT to assess the accessibility of semantic word knowledge for Dutch school children in their reading. And a series of recent studies by Harrington (2018) and his associates have found that a measure of response speed can usefully add to the predictive power of a simple computer-based vocabulary test in assessing students' readiness for academic study through the medium of English. However, this research also shows that careful piloting of such tests with particular student populations is needed to ensure that they will work effectively for their intended purpose.

A third trend in vocabulary studies is the current interest in collocations and other kinds of multi-word lexical items. Gyllstad (2020) provides a useful overview of work to design formats for this purpose. Of course, the standard word associates format includes a measure of collocational knowledge, and other researchers have mostly adopted conventional formats, such as multiple-choice and gap-filling, to assess whether learners can match the forms and meanings of common adjective-noun and verb-noun collocates. For larger lexical units, which often have non-transparent meanings and/or discourse functions, it is necessary to provide a linguistic context for the lexical item as part of the test format. This may require piloting to determine whether expert language users agree that each target multi-word item fits the context appropriately.

Further reading

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

This volume in the authoritative Cambridge Language Assessment series was the first book-length treatment of second language vocabulary testing. It reviews relevant theory and research in language testing and second language acquisition, discusses the validity of four influential vocabulary tests,

and offers practical advice on the issues involved in designing and scoring vocabulary measures for a variety of purposes.

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Basingstoke: Palgrave Macmillan.
This book offers a comprehensive guide to research on vocabulary learning. It discusses both the substantive research questions to be investigated and the methodological issues that arise in designing good studies. There is a substantial section on measuring vocabulary knowledge which includes a critical review of many published vocabulary tests and guidance on how to develop new tests for various purposes. The author also outlines ten vocabulary research projects that new researchers could usefully undertake.

Webb, S. (ed.). (2020). *The Routledge Handbook of Vocabulary Studies*, London: Routledge.
This is the most up-to-date and wide-ranging survey of research and professional practice on second language vocabulary. It includes a section of six chapters on measuring different aspects of vocabulary knowledge, along with another chapter by the present author which reviews key issues in designing vocabulary tests.

Zhang, D. and Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: Issues, findings and suggestions. *Asia-Pacific Journal of Second and Foreign Language Education* 2: 1–30.
Zhang and Koda give a more complete account of all the research on the word associates format than has been possible in this chapter. It includes a systematic discussion of the design issues that have been investigated, as well as the procedures for administering the test, the different scoring methods and relevant characteristics of the test takers. The authors also suggest future directions for research on the format.

References

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L. and Ullakonoja, R. (2014). *The Diagnosis of Reading in a Second or Foreign Language*. New York, NY: Routledge.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, H. D. and Abeywickrama, P. (2010). *Language Assessment: Principles and Classroom Practices*, 2nd edn. New York, NY: Pearson Longman.

Cremer, M. and Schoonen, R. (2013). The role of accessibility of semantic word knowledge in monolingual and bilingual fifth-grade reading. *Applied Psycholinguistics* 34: 1195–217. https://doi.org/10.1017/S0142716412000203.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *A Dictionary of Language Testing*. Cambridge: Cambridge University Press.

Deese, J. (1965). *The Structure of Associations in Language and Thought*. Baltimore, MD: Johns Hopkins Press.

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Arnold.

Greidanus, T., Beks, B. and Wakely, R. (2005). Testing the development of French word knowledge by advanced Dutch and English-speaking learners and native speakers. *The Modern Language Journal* 89: 221–33. https://doi.org/10.3138/cmlr.62.4.509.

Greidanus, T., Bogaards, P., Van der Linden, E., Nienhuis, L. and Wolf, T. (2004). The construction and validation of a deep word knowledge test for advanced learners of French. In P. Bogaards and B. Laufer (eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing*. Amsterdam: John Benjamins.

Greidanus, T. and Nienhuis, L. (2001). Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. *The Modern Language Journal* 85: 567–77. https://doi.org/10.1111/0026-7902.00126.

Gyllstad, H. (2020). Measuring knowledge of multiword items. In S. Webb (ed.), *The Routledge Handbook of Vocabulary Studies*. London: Routledge.

Harrington, M. (2018). *Lexical Facility: Size, Recognition Speed and Consistency as Dimensions of Second Language Vocabulary Knowledge*. London: Palgrave Macmillan.
Heaton, J. B. (1988). *Writing English Language Tests*, 2nd edn. London: Longman.
Hughes, A. (2003). *Testing for Language Teachers*, 2nd edn. Cambridge: Cambridge University Press.
Jang, E. E. (2014). *Focus on Assessment*. Oxford: Oxford University Press.
Kenyon, D. M. and MacGregor, D. (2012). Pre-operational testing. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. London: Routledge.
Lado, R. (1961). *Language Testing*. London: Longman.
Meara, P. (2009). *Connected Words: Word Associations and Second Language Vocabulary Acquisition*. Amsterdam: John Benjamins.
Moussavi, S. A. (1999). *A Dictionary of Language Testing*, 2nd edn. Tehran: Rahnama Publications.
Nation, I. S. P. and Beglar, D. (2007). A vocabulary size test. *The Language Teacher* 31: 9–13.
Palermo, D. and Jenkins, J. (1964). *Word Association Norms*. Minneapolis, MN: University of Minnesota Press.
Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review* 56: 282–308. https://doi.org/10.3138/cmlr.56.2.282.
Qian, D. D. and Schedl, M. (2004). Evaluating an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21: 28–52. https://doi.org/10.1191/0265532204lt273oa.
Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10: 355–71. https://doi.org/10.1177/026553229301000308.
———. (1998). Validating a test to measure depth of vocabulary knowledge. In A. Kunnan (ed.), *Validation in Language Assessment*. Mahwah, NJ: Lawrence Erlbaum.
———. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.
Reed, D. J. (2014). Field testing of test items and tasks. In A. J. Kunnan (ed.), *The Companion to Language Assessment*. Malden, MA: Wiley Blackwell.
Schmitt, N., Ng, J. W. C. and Garras, J. (2011). The word associates format: Validation evidence. *Language Testing* 28: 105–26. https://doi.org/10.1177/0265532210373605.
Schoonen, R. and Verhallen, M. (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing* 25: 211–36. https://doi.org/10.1177/0265532207086782.
Spolsky, B. (1995). *Measured Words*. Oxford: Oxford University Press.
Verhallen, M. and Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. *Applied Linguistics* 14: 344–65. https://doi.org/10.1093/applin/14.4.344.
Wesche, M. and Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Breadth vs. depth. *Canadian Modern Language Review* 53: 13–39.
Xue, G. and Nation, I. S. P. (1984). A university word list. *Language Learning and Communication* 3: 215–29.
Zhang, D., Yang, X., Lin, C.-H. and Gu, Z. (2017). Developing a word associates test to assess L2 Chinese learners' vocabulary depth. In D. Zhang and C.-H. Lin (eds.), *Chinese as a Second Language Assessment*. Singapore: Springer.

Section 8

Measurement theory in language testing


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

26
Classical test theory

James Dean Brown

Introduction/definitions

Classical test theory (CTT) was first developed and predominated in measurement circles during the early to middle twentieth century and is "a measurement theory which consists of a set of assumptions about the relationships between actual or observed test scores and the factors that affect these scores, which are generally referred to as error" (Association of Language Testers in Europe, 1998: 138).

> At the heart of CTT is the assertion that an observed score is determined by the actual state of the unobservable variable of interest plus error contributed by all other influences on the observable variable. The actual state of the unobserved variable is its hypothetical true score.
>
> *(DeVellis, 2006: S50)*

Clearly, the notions of error and true scores are central to CTT and therefore to all else that follows in this chapter.

The recognition that all measurement contains errors can be traced back at least as far as Spearman (1904: 76), who observed that individuals created "accidental deviations" in any measurement, which become "variable errors" for groups of individuals. Put in the parlance of CTT, *observed scores* (i.e., the examinees' actual scores on a test) contain *errors* (unsystematic effects due to factors not being measured), which in turn contribute to *error variance* (unsystematic variation in the scores that is due solely to random errors). Such error variance is taken to be random because it can arise from a variety of extraneous, non-systematic sources, which typically have nothing to do with the purposes for which the test was designed: for example, sources of error arise in the environment (e.g., noise, heat, lighting, etc.), administration procedures (e.g., directions, equipment, timing, etc.), scoring procedures (subjectivity, math errors, scorer biases, etc.), test items (e.g., item types, item quality, test security, etc.), the examinees themselves (e.g., health, fatigue, motivation, etc.), and so forth (Brown, 2005: 171–75).

DOI: 10.4324/9781003220756-35

Unless the scores on a particular test are completely random (i.e., 100 percent error), some proportion of the observed score variance that results from the test will be attributable to the construct that the test was designed to measure. *True scores* are hypothetical representations of the scores that would result if there were no errors in measurement. Variation among examinees in their hypothetical true scores is called *true score variance*. Logically, any set of observed scores will be made up of both true score variance and error variance. Let's represent these relationships as follows:

Total test variance = true score variance + error variance

$$Var_{Total} = Var_{Total\ Score} + Var_{Error} \quad (1)$$

This notion that observed score variance is made up of true score variance plus error variance underlies the entire framework of CTT.

Historical perspective

There are two historical perspectives through which to view CTT: the history of testing in education and psychology and the history of language testing.

The history of CTT in education and psychology

By the beginning of the twentieth century, Pearson (1896) had shown that the best value for the correlation coefficient was the covariance of the two sets of numbers divided by the product of the standard deviations for those two sets of numbers. In addition, scientists had widely accepted the notion that errors occurred in measurement. Thus, the stage was set for the beginnings of CTT, the development of which can be traced through a number of landmarks. In 1904, Spearman formulated the correction for attenuation using reliability estimates for the first time. Spearman (1910) and Brown (1910) showed how to calculate what is commonly called split-half reliability adjusted by the Spearman-Brown prophecy formula from a single set of test items. Kuder and Richardson (1937) critiqued existing methods of reliability estimation (split-half and test-retest) and derived the K-R20 and K-R21 formulas. Later, Guttman (1945) showed that reliability estimates from one administration of a test are *lower bounds* estimates (i.e., underestimates) of the actual correlation between observed and true scores. All these CTT trends and others discussed later in the chapter are best summarized in Lord and Novick (1968).

The history of CTT in language testing

Based on Spolsky (1978) and Hinofotis (1981), Brown (2005) briefly traces the history of language testing through three central movements: psychometric/structuralist testing, integrative testing, and communicative testing. Classical test theory (as exemplified by Lado, 1961) dominated all three of those movements but has been overshadowed during the last 40 years or so among expert language testers by additional theories and techniques, especially criterion-referenced testing (CRT), generalizability theory (G theory), and item-response theory (IRT) (see Ockey, this volume, for an overview of the latter). Indeed, the label *classical theory* surfaced and became necessary only in the presence of these alternative theories. From the perspective of the expert language testers doing research using

CRT, G theory, and IRT, *classical* is sometimes used pejoratively to mean *old fashioned*, which, of course, is really not the case, given the rarity of genuine instances of use of CRT, G theory, and IRT in operational testing situations in language programs around the world. Indeed, outside the major test development institutions (e.g., Educational Testing Service, University of Cambridge Local Examinations Syndicate, the Society for Testing English Proficiency, etc.) and the language testing professional organizations (e.g., the International Language Testing Association, the European Association for Language Testing and Assessment, etc.), classical testing theory is probably the dominant psychometric theory actually applied to the problems of language testing in real-world language teaching situations.

Critical issues and topics

Other crucial notions in the development of CTT include the concepts of strictly parallel items and item variance, as well as parallel tests. Let's consider each of these in turn before combining them with the ideas of true score variance and error variance discussed earlier. The phrase *strictly parallel items* describes a set of conditions that were necessary to logically derive CTT procedures for estimating the reliability and standard error of measurement for a test. Thus, this notion underlies much of CTT. The conditions or assumptions of strictly parallel items are that they must

1 be related to some true score (i.e., be unidimensional)
2 be independent of all other items in terms of error
3 be independent of true scores in terms of error
4 have equal variances
5 be equally good indicators of true scores.

As a matter of practice, in piloting and revising language tests under CTT, all items are designed to test a single construct equally well (thus, test developers are striving to meet assumptions 1 and 5). Furthermore, test development and administration procedures are used that minimize the degree to which items' errors might be correlated to each other and to true scores (thus addressing assumptions 2 and 3). Then, by selecting those items for the final version of a test that are most highly correlated with the total scores (and therefore probably most highly correlated with true scores) and within a desired range of item facility (e.g., from 0.30 to 0.70), the test developer maximizes the degree to which items discriminate well, the degree to which they discriminate similarly, the degree to which they produce equal variances (assumption 4), and the degree to which they are equally good indicators of true scores (assumption 5).

Another concept is important to any discussion of CTT: *item variance* is the standard deviation squared of the item responses. In the special case in which items are scored right/wrong and coded 1 for correct and 0 for incorrect, the item variance for each item can be calculated by simply multiplying the proportion of correct answers (symbolized by p) times the proportion of incorrect answers (symbolized by q, which is equal to $1 - p$). Thus, item variance = $p \times q = p\,(1 - p)$ for each item. For example, a difficult item with $p = 0.20$ would mean that only 20 percent of the examinees answered correctly; the proportion of incorrect answers would be $q = (1 - p) = (1 - 0.20) = 0.80$, and the item variance for that item would be $p \times q = 0.20 \times 0.80 = 0.1600$.

The idea of *parallel tests* describes the tests that will result if sets of strictly parallel items are repeatedly developed (to measure a construct) and compiled into tests (with

equal numbers of items). Any two such tests can theoretically be said to be parallel tests. More importantly from a CTT point of view, the average of each examinee's score on an infinite number of such tests, at least theoretically, would provide that examinee's true score.

Once this theoretical idea of examinee true scores is accepted, it is a small step to argue that a correlation coefficient calculated between the observed scores on a single operational test and the true scores for that construct or domain (i.e., the average scores across an infinite number of parallel tests) would indicate the proportions of the total variance in observed scores that are related to true score variance and error variance. Under CTT, it is assumed that total variance $\left(Var_{Total}\right)$ is the sum of true score variance $\left(Var_{True\ Score}\right)$ plus error variance $\left(Var_{Error}\right)$ and, therefore, that the variance that is not accounted for as true score variance must be error variance. Since true scores have no error variance, this error variance must be associated with the operational test. In CTT, reliability can be defined in terms of true score variance and error variance as follows:

$$reliability = \frac{Var_{\substack{True \\ Score}}}{Var_{\substack{True \\ Score}} + Var_{Error}} = \frac{Var_{\substack{True \\ Score}}}{Var_{Total}} \tag{2}$$

Thus, reliability is the true score variance on a test divided by the total variance on that test (which consists of true score variance plus error variance).

Current contributions and use

Often when language testers (especially young ones) talk or write about CTT, they refer to it as though it is a thing of the past. I would argue that CTT is alive and well.

First, many (even most) people who take any university courses in language testing are trained primarily in CTT. Brown and Bailey (2008) show that the vast majority of topics (and probably course hours) involve content within the general area of CTT.

Second, a number of textbooks are available to help language teachers and administrators do language testing. For example, the four most commonly used textbooks written for MA level SLS teacher training programs (according respondents in Brown and Bailey, 2008) appear to be Bachman and Palmer (1996), Brown (2005), Alderson *et al*. (1995), and Hughes (2002). Brown (2005) covers both norm-referenced and criterion-referenced analyses. It is therefore at least 50 percent based on CTT. The other three books, insofar as they cover statistical testing theory, do so from the CTT perspective.

Third, perhaps because of the first and second points mentioned earlier, most language teachers and administrators doing proficiency or placement testing in their programs are doing it on the basis of their training in CTT during MA or PhD programs (again, see Brown and Bailey, 2008).

Fourth, while expert language testers seem to naturally take an interest in CRT, G theory, IRT, and so forth, understanding CTT is clearly a prerequisite for comprehending and using any of the more exotic forms of analysis. Indeed, articles reporting research based on these forms of analysis sometimes start by contrasting the newer analyses with CTT or by explaining the newer concepts with reference to CTT concepts (e.g., Brown and Ross, 1996; Henning, 1984).

CTT, then, appears to be alive and well because it continues to contribute to the actual operational testing in language programs and because it serves as the basis for understanding all aspects of language testing.

Main research methods

The main research methods used in CTT include at least item analysis (item facility, item discrimination, and distractor efficiency analysis), reliability estimates, the standard error of measurement, and various validity analyses (including criterion-related, content, and construct validity strategies). This section will end by stressing the importance of calculating descriptive statistics and attending to the distributions they describe when interpreting any other testing statistics.

Item analyses

Item facility (also known as *IF*, *item difficulty*, *item easiness*, or *p* for proportion correct) is usually expressed as the proportion, or percentage, of examinees who answered a particular item correctly. If 95 percent of the examinees answer an item correctly, that item is viewed as very easy. If only 11 percent of the examinees answer an item correctly, the item is considered very difficult. Ideal items in CTT (which is essentially norm-referenced in focus) would be items of intermediate facility: that is, items between 0.30 and 0.70 in facility or, put another way, items that 30 to 70 percent of the examinees answered correctly. In addition to the fact that testers do not want to wastefully require examinees to breeze through items that are too easy or suffer through items that are too difficult for them, items with difficulties within the range of 0.30 to 0.70 generally discriminate better than items outside that range.

Item discrimination (also called *ID*, *item-total correlations*, *point-biserial correlation coefficients*, *d-value*, or simply r_{pbi}) indicate in one way or another the degree to which the examinee performances on a given item are related to their performances on the test as a whole. Given that the sum of the information provided by all the items on a test is assumed to be more reliable than the information on any single item, examining the *ID* in a form such as item-total correlations indicates the degree to which each item is testing what the more reliable total score is testing. For instance, a point-biserial correlation coefficient (r_{pbi}) of 0.26 would indicate that the particular item in question was not very highly correlated with the total test scores; that is, this item is not very highly related to whatever the ensemble of items is testing. In contrast, an r_{pbi} of 0.87 would indicate that the item is doing very much what the rest of the test is doing. Put another way, item discrimination indicates the degree to which an item is spreading the examinees out similarly to the way the total scores spread out the examinees. Thus, items with high item discrimination indexes are viewed from the classical theory perspective as desirable to include in revised versions of a test. Including items with high discrimination indexes in a revised version of a test will lead to more reliable measurement overall, whereas including items with low discrimination will lead to less reliable measurement.

One simple and quick way to estimate item discrimination (*ID*) is by calculating the *item facility* (*IF*) for the top-scoring group of examinees (e.g., the top 30 percent of examinees based on their total scores), calculating the *IF* for the bottom-scoring group of examinees (e.g., the bottom 30 percent of examinees based on their total scores), and then subtracting the *IF* for the top group minus the *IF* for the bottom group as follows: $ID = IF_{top} - IF_{bottom}$.

For example, if 90 percent of the top group answered an item correctly, but only 20 percent of the bottom group did so, the *ID* would be as follows: $ID = IF_{top} - IF_{bottom} = 0.90 - 0.20 = 0.70$. Such an item seems to discriminate well between the high-scoring and low-scoring examinees.

Distractor efficiency analysis (also known as *distractor analysis*) is used to investigate the degree to which the internal parts of items are working well. This form of analysis is most commonly used with multiple-choice items, so any discussion of this topic will necessarily involve related terminology like *item stem* (the stimulus part of the item at the top), the *options* (the alternative choices presented to the examinee), the *correct answer or choice* (the option that will be scored as correct), and the *distractors* (the options that will be counted as incorrect). The incorrect options are called distractors because they are meant to divert, or pull the students away, from the correct answer if they do not know which is correct. The primary goal of distractor efficiency analysis, then, is to investigate the degree to which the distractors are attracting students who do not know which answer is correct. To do this for a given item, the percentages of students who choose each option are analyzed. If this analysis can also compare the percentages choosing each option in the upper, middle, and lower groups, the analysis will be even more useful. In any case, the goal is to investigate the degree to which the distractors are functioning efficiently.

After dividing the examinees on a test into high, middle, and low thirds based on their total scores, a sound item in CTT would be one for which most (e.g., 100 percent) of the high group select the correct answer, about 50 percent of the middle group, and few or none of the lower group. However, distractor efficiency analysis focuses on how well the other options, the distractors, are doing their jobs. For example, the distractors can only be said to be working efficiently if they are all drawing at least some examinees away from guessing the correct answer without knowing it. Thus, a distractor that nobody selects is contributing nothing to the functioning of the item because it is not attractive enough; in contrast, a distractor that many high group students are selecting may indicate that the distractor is too attractive or, indeed, that the answer key is wrong or even that there are two possible answers and so forth. Common sense examination of the percentages of students selecting each option can lead the test developer to revise and improve the internal functioning of individual items.

Reliability estimates

Recall the earlier CTT argument that test score reliability is the true score variance divided by the total variance (which, remember, consists of true score variance plus error variance). Clearly, it is impossible to develop and administer an infinite number of parallel tests and determine the examinees' actual true scores. Fortunately, CTT provides a number of shortcuts, one of which involves calculating the correlation coefficient between two sets of observed scores on two parallel tests. Then, because these tests both contribute error variance to the correlation coefficient, we know that the resulting correlation coefficient will be lower (i.e., an underestimate) than the correlation coefficients between either of those operational tests and the true scores (which, after all, have no error variance). We therefore interpret the correlation coefficient between the two operational parallel tests as an underestimate of the proportion of true score variance in either of the tests. This set of procedures is sometimes called *equivalent forms reliability estimation* and gives us an idea of the consistency of results across parallel test forms (where true scores are theoretically based on an infinite number of parallel tests).

Similar procedures used to analyze the correlation between two administrations of the same test (say, two weeks apart) are called *test-retest reliability estimation*, which gives us information about the stability of the same test over time (where true scores are theoretically based on an infinite number of administrations of the same test). In reality, even these CTT shortcut methods of estimating test reliability (across forms or across administrations) are so cumbersome (requiring two tests in the one case and two test administrations in both cases) that they are rarely used in practice.

Hence, easier shortcut methods for calculating CTT reliability were developed. One such method has the unwieldy moniker of *split-half reliability (adjusted by the Spearman-Brown prophecy formula)*. Split-half reliability, as it is sometimes known, artificially creates two forms of a test from a single test. Typically, the two forms are created by scoring the odd- and even-numbered items separately. Then the correlation coefficient between the half-test scores for the even-numbered items and those for the odd-numbered items is calculated. The resulting correlation coefficient provides an estimate of the half-test reliability: that is, the reliability of the even-numbered items or the odd-numbered items, but not both combined. The full-test reliability will logically be higher than the half-test reliability because more observations are typically more reliable then fewer observations. Hence the half-test correlation coefficient must be adjusted using the Spearman-Brown prophecy formula for a test of double the half-test length. One formula for this adjustment is as follows: S – B prophecy = $(2 \times r_{half\text{-}test})/(1 + r_{half\text{-}test})$. For example, if the half-test correlation ($r_{half\text{-}test}$) turned out to be 0.50, the S – B prophecy adjustment would be $(2 \times r_{half\text{-}test})/(1 + r_{half\text{-}test}) = (2 \times 0.50)/(1 + 0.50) = 1.00/1.50 = 0.67$. Thus, the split-half reliability adjusted for full-test reliability would be 0.67.

Other internal consistency estimates have been developed within CTT that are relatively easy to calculate. These include at least Kuder-Richardson (K-R) formulas 20 and 21 and Cronbach's alpha.

K-R21 is perhaps the simplest to calculate because it requires only that the tester know the number of items on the test (k), the mean (M), and the standard deviation (S) of the scores as follows:

$$K-R21 = \frac{k}{k-1}\left(1-\frac{M\left(k-M\right)}{kS^2}\right) \tag{3}$$

However, K-R21 assumes that the items on the test have been scored right or wrong and that they are all of equal facility. If those assumptions are tenable, K-R21 gives a reasonable estimate of reliability. For a test with 30 items, with a mean (M) of 15 and a standard deviation (S) of 8, K-R21 would be calculated as follows:

$$K-R20 = \frac{k}{k-1}\left(1-\frac{M\left(k-M\right)}{kS^2}\right) = \frac{30}{30-1}\left(\frac{15\left(30-15\right)}{30\left(8^2\right)}\right) = 1.034\left(1-\frac{225}{1920}\right) \tag{4}$$

$$= 1.034(1 - .117) = 0.913 \approx 0.91$$

K-R20 is a bit more difficult to calculate by hand (though it is fairly straightforward to calculate using a Microsoft Excel spreadsheet on a computer) because it requires that the tester know not only the number of items on the test (k) and standard deviation of the test

scores (S_t) but also the p and q values for all items (i.e., where p = item facility, or the proportion of examinees who answered correctly, and $q = 1 - p$, or the proportion of examinees who answered incorrectly). For example, an item that 40 percent of the examinees answered correctly would have $p = 0.40$, and q would equal $1 - p = 1 - 0.40 = 0.60$. Based on the given definitions of the symbols in the formula, K-R20 is calculated as follows:

$$K-R20 = \frac{k}{k-1}\left(1-\frac{\Sigma pq}{S_t^2}\right) \tag{5}$$

The only tedious part of this equation is the Σpq in the numerator on the right side. To calculate Σpq, the tester needs to multiply p times q for each item (e.g., for the example item in the previous paragraph, $pq = 0.40 \times .60 = 0.2400 \approx 0.24$) and then sum up the pq values for all the items on the test. K-R20 assumes that the items on the test have been scored dichotomously (right or wrong). If that assumption is tenable, K-R20 provides a good estimate of reliability. For a test with 30 items, with $\Sigma pq = 7.5$ and with a standard deviation of 8, K-R20 would be calculated as follows:

$$K-R20 = \frac{k}{k-1}\left(1-\frac{\Sigma pq}{S_t^2}\right) = \frac{30}{30-1}\left(1-\frac{7.5}{64}\right) = 1.034\left(1-0.117\right) = 0.913 \approx 0.91 \tag{6}$$

Cronbach suggested another, more flexible formula (i.e., one that is not restricted to dichotomously scored tests and can therefore be used with weighted items), which is called alpha (α). Cronbach's α requires that the tester know not only the number of items on the test (k) and standard deviation of the test scores (S_t) but also the item variances for each individual item $\left(S_i^2\right)$ (i.e., the standard deviation squared for each item; these values can be calculated fairly simply in an Excel spreadsheet using the = VARP(range) function. With those quantities in hand (and summing up the S_i^2 values for all the items), for a test with 30 items, with $\Sigma S_i^2 = 7.5$, and with a standard deviation of 8, calculating Cronbach's α is accomplished as follows:

$$\alpha = \frac{\boldsymbol{k}}{\boldsymbol{k}-1}\left(1-\frac{\Sigma \boldsymbol{S_i^2}}{\boldsymbol{S_t^2}}\right) = \frac{30}{30-1}\left(1-\frac{7.5}{64}\right) = 1.034\left(1-0.117\right) = 0.913 \approx 0.91 \tag{7}$$

On a test with items that are all scored dichotomously, the results for K-R20 and Cronbach's α will be the same.

Interpreting CTT reliability estimates is straightforward because reliability estimates indicate the proportion of observed score variance on a test that is true score variance. For example, if Cronbach's alpha turns out to be 0.91, that means that 91 percent of the total variance is true score variance, and, by extension, 9 percent is error variance (i.e., 1.00 − 0.91 = 0.09). Expanding on the CTT conception of reliability discussed earlier, the relationships between true score and error variance would be as follows:

$$reliability = \frac{Var_{\substack{True\\Score}}}{Var_{\substack{True\\Score}} + Var_{Error}} = \frac{091}{0.91+0.09} = \frac{0.91}{1.00} = 0.91 \tag{8}$$

Since reliability is often defined as the degree to which a set of test scores measures consistently, we sometimes interpret CTT reliability estimates as indicating the proportion of observed score variance that is consistent. In addition, because all the reliability estimates described here are viewed as underestimates, none of them should overestimate the actual state of reliability in the test scores involved – as long as the assumptions are tenable.

Remember that CTT reliability estimates are calculated from the scores of a particular group of examinees taking a particular set of items. Hence, any reliability estimate is linked to that particular group and that specific set of items. Put another way, we can only make claims about the reliability of a test with reference to that particular group of examinees and that set of items. In practice, testers often cautiously make claims about the probable level of reliability if the test were administered to a *very similar* group of examinees with about the same characteristics and range of abilities.

The standard error of measurement

Reliability estimates like those described in this chapter are one useful CTT strategy for investigating test consistency. Such estimates can be used to express the reliability of a set of test scores in percentage terms, which is useful when we are arguing for the quality of a test (i.e., trying to sell it to a group of stakeholders in one way or another). However, a more concrete strategy for looking at the consistency of a set of test scores, especially when making decisions about who should pass or fail, is called the *standard error of measurement* (*SEM*, not to be confused with the *SEM* for structural equation modeling). The *SEM* provides a concrete estimate in test score values of the amount of unreliable score variation in a set of scores. This statistic is calculated as follows (where S = the standard deviation of the test scores, and $r_{xx'}$ = reliability estimate):

$$SEM = S\sqrt{1 - r_{xx'}} \tag{9}$$

For a set of test scores with an $S = 6.00$ and $r_{xx'} = 0.91$, the *SEM* would be as follows:

$$SEM = S\sqrt{1 - r_{xx'}} = 6.00\sqrt{1 - .91} = 6.00\sqrt{0.09} = 6.00 \times .30 = 1.80 \approx 2 \tag{10}$$

The *SEM* can be used to estimate the average error in score points across all the examinees who took the test. Using the *SEM*, we can estimate with certain probabilities how far examinees' scores are likely to vary by chance alone. We can predict that error alone can cause examinees' scores to vary within a band of plus or minus of one *SEM*: that is, about ± 2 points in the example here (rounded because these scores are only interpretable as whole numbers) with 68 percent accuracy (based on the expected percentages under the normal bell-shaped distribution of test scores). Thus, for a particular examinee whose score is 65, the *SEM* of 2 indicates that her scores could vary by chance alone between 63 and 67 points (65 − 2 = 63; 65 + 2 = 67) 68 percent of the time if she were to take repeated administrations of the test. If we want to be even more confident in predicting where her score is likely to be on repeated administrations, we can go out two *SEM*s (2 + 2 = 4) and say that her scores could vary by chance alone between 61 to 69 (65 − 4 = 61; 65 + 4 = 69) with 95 percent confidence.

An *SEM* of only 2 points indicates that we can expect relatively small fluctuations in the examinees' scores if they take repeated administrations of the test, while an *SEM* of 15 points would indicate larger fluctuations. Generally, the narrower (or smaller) the *SEM*, the narrower the band of random fluctuations will be or, put another way, the smaller the *SEM*, the more consistently the scores will represent the examinees' true abilities. Some testers find *SEMs* to be easier to interpret than reliability estimates because they represent score bands rather than the more abstract percentage of true score or consistent variance interpretations typically applied to reliability estimates.

Validity analyses

Under CTT, validity is typically defined as the degree to which a test is measuring what it claims to measure. Three types of validity arguments are used in CTT: criterion-related validity, content validity, and construct validity.

Criterion-related validity arguments are built in two ways. First, the tester can administer a new test (e.g., English Proficiency Test X) and give the same group of test takers a well-respected test, or *criterion measure*, of the same construct (e.g., TOEFL iBT, see https://ets.org/toefl/ibt/about) and then calculate a correlation coefficient between the two measures. Doing so investigates the degree to which the new test is producing scores that are spread out very much like the scores on the criterion measure (this form of criterion-related validity is also known as *concurrent validity*). The second approach is to administer the new test (e.g., Academic Ability Test X) and gather subsequent post-test information about the criterion construct (e.g., academic ability as measured by grade-point average of the examinees in the two years following the test administration) and then calculate a correlation coefficient between the two measures to investigate the degree to which the new test is producing scores that predict the variation in the academic ability construct as measured by grade-point average (this form of criterion-related validity is also known as *predictive validity*).

Content validity arguments are built by asking content experts to judge, in one way or another, the degree to which the items measure the construct that the test designer is claiming the items assess. For example, a panel of experts in communicative language teaching might be assembled to judge each of the tasks on a communicative English proficiency test using a rating scale ranging from 5 (high degree of match to the construct) to 1 (no match) in each of three categories (authenticity of the language input, unpredictability of the language input, and amount of productive language required).

Construct validity arguments are built by setting up an experiment to investigate the degree to which the new test is measuring what it claims to measure. The experiment can take numerous forms: (1) a *differential groups study*, in which the test is administered to two groups (one group that is known to have relatively high levels of the construct and another group that is known to have relatively low levels of the construct) and the results are compared in terms of the degree to which the results for the group with the construct are superior to those for the group lacking the construct; (2) an *intervention study*, in which the test is administered as a pre-test and post-test to a group of examinees who are taught or otherwise led to acquire the construct to demonstrate that the scores increase after the examinees acquire the construct; (3) a *multi-trait multimethod study*, in which data from tests of different constructs (e.g., reading, listening, and writing) are gathered using three very different testing methods (e.g., multiple-choice, graded classroom assignments, and self-assessment) for a group of examinees so that the pattern of all possible correlation coefficients can be compared to see if correlations are higher among tests of the same *trait*

(i.e., the same construct) as opposed to among tests using the same *method* (i.e., the same testing method); and (4) a *factor analysis study*, in which data from tests of different constructs (e.g., reading, listening, writing, and speaking) are gathered from different test batteries (for example, the TOEFL iBT [see https://ets.org/toefl/ibt/about] and *Eiken Grade 1* [see https://www.eiken.or.jp/eiken/en/eiken-tests/]) taken by the same examinees so that the results indicate whether a four-factor solution (one for each construct) makes sense and, if so, whether the loadings for the subtests suggest that each of the four factors represents a different construct.

In the last three decades, studies of test validity have moved historically from the three-pronged approach described here to a unified theory of validity and then further to the expanded set of validity responsibilities assigned to testers by the Messick model (e.g., Messick, 1989, 1996; see also Chapelle and Lee, this volume). Nonetheless, the strategies developed under CTT can still serve well as part of the evidential basis for validity (though not so much for the consequential basis that Messick argued for).

Importance of descriptive statistics and normal distribution in CTT

It is important to recognize that calculating and examining the descriptive statistics (i.e., the mean, standard deviation, range, etc.) are crucial steps in interpreting all of the other testing statistics discussed in this chapter. The bottom line is that correlational statistics like item discrimination indexes, reliability estimates, as well as the numerous validity arguments that are dependent on correlation coefficients (i.e., both types of criterion-related validity, as well as construct validity based on multi-trait/multimethod or factor analysis studies) are highly reliant on (indeed, they all assume) normal distribution of scores on whatever tests are involved. Thus, as with most statistical studies, it is crucial to present the descriptive statistics that underlie all the other more exotic statistics whenever those other statistics are to be interpreted, understood, or explained. Note that well-designed criterion-referenced tests that have been carefully developed to fit a particular curriculum may not – indeed, most often should not – produce the normal distribution in their scores at the beginning or end of instruction. For this reason, separate statistical analyses outside CTT have been developed for CRTs (see Brown, 2005; Brown and Hudson, 2002).

Recommendations for practice

Given the general bias in most fields for doing research and publishing in new, exotic areas of analysis, it seems likely that most important psychometric developments in language testing will take place within the frameworks of CRT, IRT, G theory, SEM, and Messick's expanded view of validity. Nonetheless, particularly at the local level in areas of proficiency and placement testing, where the target classes and examinees have fairly stable characteristics, CTT will probably continue to thrive. Testers working at that level should consider the advantages and disadvantages of CTT as a framework for developing their tests and decide for themselves if CTT will be adequate for their needs.

Advantages of CTT

There are numerous advantages to working within the CTT framework when doing test development. First, the concepts of CTT are widely taught and understood. Second, CTT concepts are relatively easy to learn, use, and explain. The very useful descriptive statistics

(e.g., the mean, standard deviation, range, etc.) and item analysis statistics (especially item facility and item discrimination) can all be calculated fairly easily in spreadsheets or in widely available computer programs. Third, the CTT model fits many measurement needs, particularly for the development of proficiency and placement measures that can aid admissions, program comparisons, and placement decisions. Fourth, right or wrong, CTT accommodates and allows interpretation of examinee scores of 0 percent and 100 percent and item facility estimates of .00 and 1.00 – results that do occur in the real world. Such examinee scores (a.k.a. person ability estimates) and item difficulty estimates are generally considered misfitting in IRT models.

Disadvantages of CTT

However, working within CTT carries certain disadvantages as well. First, CTT tests tend to be long and necessarily have homogeneous items. Second, examinees taking tests developed through CTT may have to slog through numerous items that are far too easy or too difficult for them as individuals. Third, the results on CTT tests only apply to the sample at hand or very similar samples. Fourth, CTT results only apply to the current selection of items. Fifth, because of its dependence on the normal distribution, CTT is only useful for developing norm-referenced tests. Sixth, because of the correlational nature of item discrimination, reliability, and some validity estimates, CTT items and tests often end up being sensitive to differences between the extremes on the scale. Seventh, while in actuality the measurement errors on a test vary across the range of all possible scores on a CTT test (i.e., the SEM is generally larger at or near the mean and becomes increasingly smaller as scores get further from the mean in both directions), the standard error of measurement that is estimated in CTT is an *average* across that range.

Caveats

Given the disadvantages listed in the previous paragraph, testers using CTT should keep at least the following caveats in mind while using it:

1 Don't write tests that are too long (i.e., pilot and revise for shorter tests).
2 Don't forget to use item analysis to fit a given test as closely as possible to the range of abilities in the examinees being tested.
3 Don't continue to use a particular CTT test (without re-validating it) if the examinees' abilities are changing on average from semester to semester or year to year.
4 Don't assume that items in different forms of a test are the same because they look similar (be sure to pilot and equate test forms).
5 Don't use CTT for criterion-referenced purposes because CRTs are not designed to spread students out and are therefore not referenced to a normal distribution.
6 Don't interpret CTT item discrimination estimates, reliability estimates, and validity statistics as characteristics of the test itself, but rather as characteristics of that set of items when administered to the examinees in the particular distribution involved.

Future directions

The disadvantages of CTT listed here appear to be more numerous than the advantages. However, from the local perspective, the disadvantages may, in most cases, be more theoretical

than real. For example, I have item analyzed tests using IRT (Rasch analysis, to be specific) and CTT techniques and found little practical difference in the item-level test-development decisions that I would make based on the two models (see also Henning, 1984). All in all, I predict that CTT is sufficiently accurate, easy to learn, and practical to continue in use for years to come in real (especially local) testing situations.

I further believe that CTT item statistics and distractor efficiency analysis will continue to provide useful feedback to item writers and test developers about their items and their test specifications. In particular, I cannot imagine that distractor efficiency analysis will be replaced by the new forms of analysis. Indeed, some IRT software packages acknowledge this possibility by providing CTT item statistics and distractor efficiency analyses right along with their IRT difficulty, discrimination, and guessing parameter estimates (e.g., the *Xcalibre* software described at https://assess.com/xcalibre/).

Since CTT excels at showing test developers which items and/or tasks discriminate when they are administered to a given group of examinees, developers are tempted to focus only on and select only those items or tasks that fit the abilities of the particular population and spread them out into a nice, neat, normal distribution. After all, doing so often produces high reliability estimates for the resulting scores. However, in the interest of test validity, it has long been important to temper the use of item statistics by close analysis of item content and quality. In view of the increasing importance in recent years of validity concerns, I predict that item specifications will become progressively more important, especially in classroom- and program-level diagnostic, progress, and achievement testing. As a consequence, the roles of item/task content and item/task writers will rise in importance, and decisions will increasingly be informed by common-sense aspects of CTT like item facility and distractor efficiency analysis in ways that will at least supplement the newer forms of analysis like IRT. Such trends may foster research and debate on the authority of item/task writers, especially in terms of the need for clear item specifications in all testing situations and the role of item/task writers in creating and judging the match of test items to test specifications, as well as the match of test specifications to construct definitions or curriculum content.

True, technology has opened up many new possibilities, some or all of which may lead some testers to abandon CTT in favor of IRT analyses. Holland and Hoskens (2003: 123) view "classical test theory as a very general version of IRT, and commonly used IRT models as detailed elaborations of CTT for special purposes," and it is these special purposes that naturally nudge language testers away from CTT toward IRT models. Such special purposes include but are not limited to the following: item banking, automated test construction, sample and test equating, and computer-adaptive testing.

Regardless of all these special purposes and tendencies toward newer forms of analysis, it appears that CTT will live on as the basis for the training we give teachers and administrators to help them develop and understand local institutional (especially proficiency and placement) tests. As a result, CTT will serve as the base knowledge needed by language testers who want to learn, understand, and use the other, more specialized forms of analysis discussed in this chapter. In that sense, CTT is the cake, and all else is the icing.

Further reading

Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*, new edn. New York, NY: McGraw-Hill.

This book provides an introduction to language testing for language teachers and teachers in training. To do so, it covers both norm-referenced CTT and criterion-referenced testing and provides straightforward explanations of all concepts and clear instructions for calculating all the related statistics by hand or using an Excel spreadsheet (with written instructions only).

Carr, N. T. (2011). *Designing and Analyzing Language Tests*. Oxford: Oxford University Press.
This book also presents a general introduction to language testing. It focuses primarily on CTT and provides explanations for how to calculate all the related statistics by hand but also provides effective video modules that walk readers through the process of using an Excel spreadsheet to calculate the related statistics.

Hauenstein, C. E. and Embretson, S. E. (2018). Classical test theory. In B. B. Frey (ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. Thousand Oaks, CA: Sage.
This encyclopedia entry provides a short and accessible introduction to key theoretical issues in CTT, including true-score and error variances, the four most commonly used reliability estimates, the standard error of estimate, item analysis, validity, and the limitations of CTT.

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement Issues and Practice* 16: 8–14.
This article offers a relatively short and clear historical overview of the development of CTT beginning with background from the beginning of the twentieth century and then coverage of the key achievements in CTT over the years, discussion of how CTT was formalized with references from 1923 to 1968, and then concluding remarks focusing on the problems and limitations of CTT.

References

Alderson, C. J., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University.

Association of Language Testers in Europe. (1998). *Studies in Second Language Testing 6: Multilingual Glossary of Language Testing Terms*. Cambridge, UK: Cambridge University Press.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford, UK: Oxford University Press.

Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*, new edn. New York, NY: McGraw-Hill.

Brown, J. D. and Bailey, K. M. (2008). Language testing courses: What are they in 2007. *Language Testing* 25: 349–83. https://doi.org/10.1177/0265532208090157.

Brown, J. D. and Hudson, T. (2002). *Criterion-Referenced Language Testing*. Cambridge, UK: Cambridge University Press.

Brown, J. D. and Ross, J. A. (1996). Decision dependability of item types, sections, tests and the overall TOEFL test battery. In M. Milanovic and N. Saville (eds.), *Performance Testing, Cognition and Assessment*. Cambridge, UK: Cambridge University Press.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology* 3: 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x.

DeVellis, R. F. (2006). Classical test theory. *Medical Care* 44: S50–S59.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika* 10: 255–82.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing* 1: 123–33. https://doi.org/10.1177/026553228400100201.

Hinofotis, F. B. (1981). Perspectives on language testing: Past, present and future. *Nagoya Gakuin Daigaku Gaikokugo Kyoiku Kiyo* 4: 51–59.

Holland, P. W. and Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika* 68: 123–49.

Hughes, A. (2002). *Testing for Language Teachers*. Cambridge, UK: Cambridge University Press.

Kuder, G. F. and Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika* 2: 152–260. https://doi.org/10.1007/BF02288391.

Lado, R. (1961). *Language Testing: The Construction and Use of Foreign Language Tests*. New York, NY: McGraw-Hill.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher* 18: 5–11. https://doi.org/10.3102/0013189X018002005.

———. (1996). Validity and washback in language testing. *Language Testing* 13: 241–56. https://doi.org/10.1177/026553229601300302.

Pearson, K. (1896). Mathematical contributions to the theory of evolution-III. Regression, heredity and panmixia. *Philosophical Transactions*, *A* 187: 252–318. https://doi.org/10.1098/rsta.1896.0007.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* 15: 72–101. https://doi.org/10.1037/11491-005.

———. (1910). Correlation calculated from faulty data. *British Journal of Psychology* 3: 171–95. http://proquest.umi.com/login/athens?url=www.proquest.com/scholarly-journals/correlation-calculated-faulty-data/docview/1293688112/se-2?accountid=11979.

Spolsky, B. (1978). Introduction: Linguists and language testers. In B. Spolsky (ed.), *Advances in Language Testing Series: 2*. Arlington, VA: Center for Applied Linguistics.

27

Item response theory and many-facet Rasch measurement

Gary J. Ockey

Introduction/definitions

Item response theory (IRT) is an approach used to estimate how much of a latent trait an individual possesses. The theory aims to link individuals' observed performances to a location on an underlying continuum of the unobservable trait. Because the trait is unobservable, IRT is also referred to as latent trait theory – the literal meaning of *latent* is "hidden." An example of a latent trait is a test taker's English reading ability. The construct of reading ability is not observable, so measuring this ability requires observable performances designed to assess it. For instance, test takers could read short passages and then answer multiple-choice questions designed to measure their comprehension of the passages. IRT can then be used to relate these observable performances – scores on the multiple-choice reading test items – to test takers' underlying reading abilities. IRT can be used to link observable performances to various types of underlying traits. For instance, it can be used to connect individuals' observed or self-reported anxiety about an assessment to their underlying levels of test anxiety.

The following definitions are helpful for understanding IRT concepts as discussed in the chapter. *Latent variables* are unobservable traits like second language listening ability, which influence observable behaviors, such as individuals' responses to items on a second language listening assessment. The term *latent variable* is used synonymously with the terms *construct*, *factor*, and *underlying trait* throughout the chapter. A *model* is a mathematical equation in which independent variables are combined to optimally predict dependent variables. Various IRT models, some of which are discussed in this chapter, are used to estimate individuals' underlying traits on language ability constructs. *Parameter* is used in IRT to indicate a characteristic about a test's stimuli. For instance, a one parameter (1PL) model mathematically defines only one characteristic of a test item.

Historical perspective

The concepts and mathematics that underlie IRT are not new. Linacre contends that the foundations of item response theory "go back to ancient Babylon, to the Greek philosophers, to the adventurers of the Renaissance" (Bond and Fox, 2015: x). Current IRT practices can be

 DOI: 10.4324/9781003220756-36

traced back to two separate lines of development. Thurston's (1925) "A method of scaling psychological and educational tests" is the paper often cited as the "intimations" of IRT for one line of development. The paper describes how to place items designed to assess children's mental achievement on an age-graded scale (Bock, 1997). Roughly 30 years later, Fredrick Lord (1952) provided the foundations of IRT as a measurement theory by outlining assumptions and providing detailed models. Lord and Novick's (1968) monumental textbook *Statistical Theories of Mental Test Scores* outlined the principles of IRT that they and others had developed in the previous two decades. Due to Lord's access to large data sets as an employee of the Educational Testing Service, the authors were able to apply developing IRT principles to real data sets.

George Rasch (1960), a Danish mathematician, is also credited with developing IRT principles. The primary focus of his research was to use probability to separate test taker ability and item difficulty. Rasch inspired Benjamin Wright, who was a professor at the University of Chicago, to teach these principles in his graduate courses. Wright became an avid proponent of Rasch's approach, and he and his graduate students are credited with many of the developments of the family of Rasch models.

Although the two lines of IRT development are mathematically equivalent and have led to similar practices, they have at least one major difference. Rasch models are prescriptive. If data do not fit the model, the data must be edited or discarded. The other IRT line promotes a descriptive philosophy. Under this view, a model is built that best describes the characteristics of the data. If the model does not fit the data, the model is adapted until it can account for the data.

IRT became prominent in the language testing literature in the 1980s. The first article published in the journal *Language Testing* was written by Grant Henning (1984). His article, "Advantages of latent trait measurement in language testing," discusses the advantages of IRT as compared to classical test theory (CTT). The following year, three papers in an issue of *Language Testing* (Davidson and Henning, 1985; Henning *et al.,* 1985; Woods and Baker, 1985) further established IRT in the language testing field.

While language assessment researchers and practitioners have benefited from both lines of IRT research, the majority have used Rasch models. About a decade after IRT appeared in the journal *Language Testing*, a particularly influential book on the subject was written by Tim McNamara (1996), *Measuring Second Language Performance*. The book provides an introduction to many-facet Rasch measurement (MFRM) and FACETS (Linacre, 1987–2019), software used for estimating ability on performance-based assessments. Soon after the publication of McNamara's book, studies using MFRM began to appear in the language testing literature. Some of these studies are mentioned in the Current contributions and research section. McNamara and two of his colleagues later updated the book (McNamara *et al*., 2019) by situating an introduction to the Rasch model and the FACETS program in current thinking about fairness in language assessment. Thomas Eckes (2015) wrote another influential book, *Introduction to Many-Facet Rasch Measurement*, which guides researchers through the interpretation of FACETS output. These books have led to increased use of IRT models, particularly the Rasch model, in the language assessment arena.

Critical issues and topics

Conceptualization of IRT

The purpose of IRT is to link test takers' observed performances, such as scores on a multiple-choice listening test, to a location on an underlying continuum of the latent trait,

such as listening ability. This is made possible by relating the probability of test takers' correct responses to items and the ability of the test taker.

IRT assumes an underlying ability continuum with the least amount of ability at the far left and the most at the far right. Each test taker has an ability, represented on the logit scale, which can be located at a point along this continuum. Test takers with low ability estimates have negative estimates, and test takers with high ability estimates have positive logit values. This continuum can be seen in Figure 27.1, which is further described in the Item characteristic curves section. Logits are also defined in that section.

Observable performances aimed at assessing the ability of interest are used to locate test takers' abilities on this latent continuum. Correct and incorrect responses of 20 test takers on 15 multiple-choice items designed to assess listening comprehension are presented in Table 27.1. These responses represent the observable performances used to locate test takers' listening abilities on an underlying listening ability continuum. It should be noted that in practice, more test takers would be needed to use IRT techniques.

In Table 27.1, "1" indicates a correct response and "0" an incorrect response, so, for example, Keiko, the first test taker, answered items 3, 1, 4, 10, and 12 correctly and all of the other items incorrectly.

The test takers and items have been ordered based on their patterning of correct and incorrect responses to help demonstrate the theory underlying IRT. Progressing from left to right in the table, items become more difficult, and progressing from top to bottom, test takers possess more listening ability. The bottom left-hand corner is filled with mostly correct responses because this is where test takers with the most listening ability are answering the easiest items. In contrast, the top right-hand corner is filled with incorrect responses

Table 27.1 Test taker response on a multiple-choice listening test

Item															
Test Taker	3	1	5	4	7	12	8	10	14	13	9	2	6	15	11
Keiko	1	1	0	1	0	1	0	1	0	0	0	0	0	0	0
Sally	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Steve	1	1	0	1	1	1	0	1	0	0	0	0	0	0	0
Pierre	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
Don	1	1	1	1	1	0	0	0	1	0	1	0	0	0	0
Maia	1	1	1	1	1	1	0	0	1	1	0	0	0	0	0
Trang	1	1	1	0	0	1	1	1	0	0	1	1	0	0	0
Srey	1	0	1	1	1	1	1	0	1	1	0	0	0	0	0
Junko	1	1	1	0	1	1	1	0	1	0	1	0	1	0	0
Lee	1	1	1	1	1	1	1	0	0	0	1	1	0	0	0
Sam	1	1	1	1	0	1	1	0	1	0	1	0	1	0	0
Yasu	1	1	1	1	1	1	1	0	1	1	0	0	0	0	0
Tara	1	1	1	1	1	0	1	1	0	0	0	0	1	1	1
Inma	1	1	1	1	1	1	0	0	0	1	1	1	1	1	0
May	1	1	1	1	1	0	1	1	1	1	0	1	0	1	0
Anna	1	1	1	1	1	0	1	1	0	0	0	1	1	1	1
Ning	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1
Mina	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1
Jack	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
Ali	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

because this is the area in the table where test takers with the least amount of listening comprehension are answering the most difficult items. It should be noted that the pattern is not completely predictable. It might better be described as probabilistic since not all responses in the matrix would have been predicted from the pattern.

A practical advantage of probabilistic modeling should be pointed out at this juncture. Response patterns found in the data that are highly unlikely can be identified. Test takers or items that produce unlikely response patterns are often referred to as *misfitting* (Eckes, 2015). Once identified, misfits can be investigated, and decisions can be made about how to handle them. For instance, if a test taker correctly answers most hard items but incorrectly answers many easy items, it may be that the test taker was unfairly distracted during the exam and should be given an opportunity to retake it.

The least predictable responses in Table 27.1 are in a strip from the top left-hand corner to the bottom right-hand corner. In this section of the matrix, it appears that the responses are correct or incorrect in a random manner. In these cases, test takers' listening abilities are at about the same level as the difficulty of the items. Thus, roughly half of the time, the test takers respond correctly, and the other half of the time, they respond incorrectly. Put another way, these test takers have about a .5 probability of answering these items successfully. By capitalizing on these probabilities, a test taker's ability can be defined by the items that are at this level of difficulty. This is central to IRT because it means that test taker ability and the difficulty of items can be placed on the same latent ability scale.

The one parameter Rasch model

Relating test taker ability to the difficulty of items makes it possible to mathematically model the probability that a test taker will respond correctly to an item. The most fundamental IRT model, the Rasch or one parameter (1PL) logistic model (where item discrimination/slope is set at 1.0 for the Rasch model and based on the average item discrimination/slope for most other one parameter models) is presented as Equation 27.1.

$$\text{Probability of correct answer} = \frac{\exp\left(\text{test taker ability} - \text{item difficulty}\right)}{\left(1 + \exp\left(\text{test taker ability} - \text{item difficulty}\right)\right.} \tag{27.1}$$

Readers should focus on the relationship between the probability of the correct answer, test taker ability, and item difficulty. Item difficulty is subtracted from test taker ability to arrive at the probability of the test taker answering the item correctly. The rest of the equation, including *exp*, the exponential function, is based on probability statistics and is beyond the scope of this introductory chapter. However, to demonstrate this relationship, an example calculation is provided. De Ayala (2009) provides an explanation of the mathematics underlying IRT models.

The example will predict the probability of Melanie, a new test taker with a listening ability of 1on the underlying logit ability scale answering an item with a difficulty of 0, an item of moderate difficulty. The number 1 would replace test taker ability in the top and the bottom of the right-hand side of the equation, and 0 would replace item difficulty on the top and bottom of the right-hand side of the equation. To make the calculation, it is necessary to begin by subtracting Melanie's ability estimate from the difficulty of the item in both the numerator and the denominator of the right-hand side of the equation: that is, 1 – 0 equals 1.

The exponential function (available on many calculators) of 1 is then calculated to give a value of 2.72 in the numerator and 1 + 2.72 in the denominator. The 2.72 in the numerator is then divided by the 3.72 in the denominator to give a probability of .73, indicating that Melanie has about a 73 percent chance of answering an item with a difficulty level of 0 correctly. It's not surprising that Melanie has a good chance of answering the item correctly because her listening ability is greater than the difficulty of the item.

Item characteristic curves (ICCs)

The relationship between the probability of a test taker answering an item correctly and the test taker's ability can also be presented in graphic form in an item characteristic curve (ICC). These graphs can be used to understand IRT principles as well as differences among various IRT models (Ockey and Choi, 2015). In addition, ICCs are commonly used to help test developers understand how specific items function on an assessment. Being able to interpret ICCs is important for making decisions about which items to maintain, modify, or discard for a particular assessment instrument.

ICCs present the characteristics of each item as trace lines on probability and ability continuums. As previously discussed, the ability of the test taker and the difficulty of the item are on the same logit scale. ICCs for three test items (hard, moderate, easy) from the 1PL Rasch model are presented in Figure 27.1.

The furthest left line, the lightest line, represents one item; the dashed line another; and the furthest right line, the solid line, another. The probability of a test taker correctly responding to an item is presented on the vertical axis. This scale ranges from 0 probability at the bottom to an absolute probability of 1.0 at the top. The horizontal axis displays the estimated ability level of test takers in relation to the item difficulty, with least at the far left

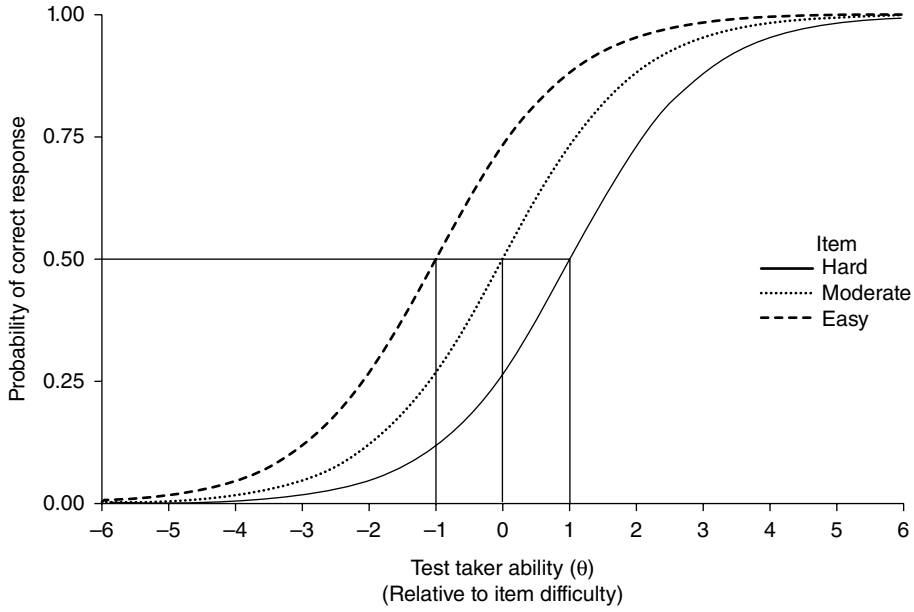


Figure 27.1 ICCs for three items from the 1PL Rasch model

and most at the far right. The measurement unit of the scale is a logit, which is short for log odds and based on probability statistics. Odds are defined as the probability of success divided by the probability of failure, and the logit is the natural logarithm of odds. One logit is equal to the distance along the measurement scale that increases the odds of success by 2.72 (Eckes, 2015). The logit scale is set to have a center point of 0, and a test taker with an ability of 0 logits would have a .50 probability of correctly answering an item with a difficulty level of 0 logits.

ICCs express the relationship between the probability of a test taker correctly answering an item and a test taker's ability. As a test taker's ability level increases, moving from left to right along the horizontal axis, the probability of correctly answering an item increases, moving from the bottom to the top of the vertical axis. For instance, for the item represented by the dotted (middle) line, a test taker with an ability level of 0 on the logit scale would have a .5 probability of answering the item correctly. This can be seen by drawing a vertical line from the 0 ability level on the horizontal axis to the ICC and then noting where this intersection occurs on the vertical axis. Lines have been drawn in Figure 27.1 from the solid line to the horizontal axis and the vertical axis to illustrate this point.

The three ICCs shown in Figure 27.1 were chosen because they demonstrate some characteristics of ICCs of the 1PL Rasch model. First, the ICCs are somewhat S-shaped, meaning the probability of a correct response changes considerably over a small ability level range. For example, the ICC represented by the solid line indicates that the probability of a correct response increases quite gradually for test takers with abilities at the low end, below -1, and high end, above +3, of the scale. Test takers with abilities ranging from below -1 have less than a .2 probability of answering the item correctly, whereas, for test takers with ability levels in the middle of the scale, between roughly -1 and +3, the probability of correctly responding changes from quite low, below .2, to quite high, nearly .9.

Inflection points are at the location on ICCs where the slope of the line is at a maximum and a test taker's chances of answering an item correctly are .5 (for the 1PL and many other models). Inflection points are indicated by the thin dotted lines that have been drawn from the middle of the ICCs – that is, half the distance from the upper and lower extremes – to the vertical and the horizontal axes. The inflection point also indicates the difficulty of the item. The thin dotted vertical lines drawn from the ICCs inflection points to the horizontal axis indicate the difficulties of the three items. The light line has a difficulty level of -1 logits; the dotted (middle) line, 0 logits; and the solid line is the hardest item, with a difficulty level of 1 logit.

Knowing the location of an inflection point can assist test developers in increasing the measurement properties of the test. Information about all the items on a test are often combined and presented in test information function (TIF) plots. The TIF indicates the average item information at each ability level. The TIF can be used to help test developers locate areas on the ability continuum where there are few items. Items can then be written that target these ability levels. For instance, if the TIF indicated that few items discriminate well at the -2 to -1 logit ability level on a listening test, item writers could craft additional items targeted at this level.

Two and three parameter models

In two parameter models (2PLs), unlike 1PLs, the shapes of the ICCs can vary based on each item's particular discrimination properties. Stated differently, items are not all modeled (or assumed) to discriminate equally. The equation for the 2PL logistic model is as follows:

Probability of correct answer =

$$\frac{\exp\left[\text{item discrimination}\left(\text{test taker ability} - \text{item difficulty}\right)\right]}{\left(1 + \exp\left[\text{item discrimination}\left(\text{test taker ability} - \text{item difficulty}\right)\right]\right.} \tag{27.2}$$

Compared to Equation 27.1, Equation 27.2 is identical except item discrimination has been included in the model. Item discrimination refers to the slope of the ICC. With the 1PL, the discrimination parameter is fixed, and for all items, the probability of a correct answer is assumed to be equally related to test taker ability. With the 2PL, the relationship for each item between the probability of a correct answer and the listening ability of the test taker can be different: that is, the slopes of the ICCs can be different.

The three parameter model (3PL), which includes a parameter to control for guessing, is also often encountered in language assessment research. When test takers have a significant chance of guessing the correct answer, such as in a multiple-choice format, the 3PL may be appropriate. The equation for this model is as follows:

Probability of correct answer = guessing + (1 – guessing)

$$\frac{\exp\left[\text{item discrimination}\left(\text{test taker ability} - \text{item difficulty}\right)\right]}{\left(1 + \exp\left[\text{item discrimination}\left(\text{test taker ability} - \text{item difficulty}\right)\right]\right.} \tag{27.3}$$

As can be seen, a guessing parameter has been added to the model presented in Equation 27.2. The differences in the 1PL, 2PL, and 3PL are evident when the three ICCs in Figure 27.2 are compared. All have item difficulties of 0 on the logit scale.

In Figure 27.2, the solid line presents an ICC that was estimated with a 1PL. The estimate of person ability from this item is based only on the difficulty of the item and whether or not the item was answered correctly. The discrimination (slope) and guessing parameters are not included in the model, so the item is constrained to have a certain slope (1.0 for the Rasch model) and a guessing parameter of 0. Because it does not include a discrimination parameter, the shape of the ICC follows the S-shaped pattern of 1PL ICCs, and because it does not have a guessing parameter, it crosses the vertical axis at 0.

The dotted line represents an item that used a 2PL for estimation, which models the difficulty of the item as well as discrimination or slope of the line. This ICC does not follow the S-shaped pattern because the slope of the line has been included in the model or, put another way, allowed to vary according to the actual relationship between the probability of a correct response and the test taker's ability. When the 2PL is used, ICCs for each item will have varying slopes, whereas, when the 1PL is used, all items will have the same slopes. Because the 2PL does not include a guessing parameter, the ICC for this item also crosses the vertical axis near 0.

Information about the actual relationship between the probability of a correct response and test taker ability can also be useful for test developers. Items that have steep slopes are generally considered better than ones that gradually increase across a wide ability range. This is because items with steep slopes discriminate well at certain ability-level locations. Items that rise gradually provide little information about where a test taker should be located on the underlying ability continuum and are therefore not very useful. Items

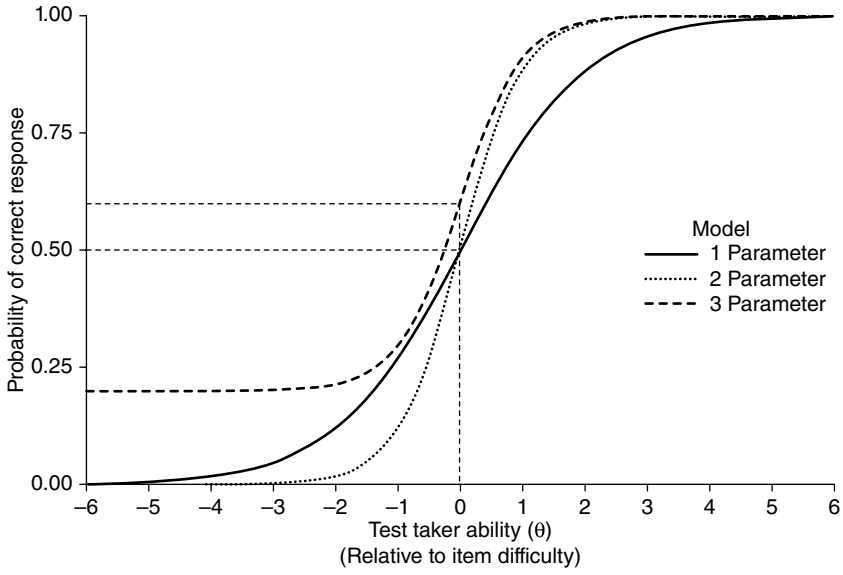


Figure 27.2 ICCs for 1PL, 2PL, and 3PL

with steep slopes can be selected for use, and less discriminating items can be modified or discarded.

Like the other two lines in Figure 27.2, the light line represents an ICC of an item with a difficulty of 0 on the logit scale. However, unlike the solid line, it includes a discrimination parameter, and, unlike both the solid and dashed lines, it includes a guessing parameter, making it a 3 parameter model (3PL). Because it models guessing, the line is not constrained to cross the vertical axis near 0. In the figure, it crosses the vertical axis well above 0, at a little below .25, indicating that even test takers with no listening ability are able to correctly respond to the item nearly 25 percent of the time. This is because with multiple-choice items, even test takers with little or no ability on the construct to be assessed have a probability of a correct response of roughly .25 on a test with four answer options. It should be noted that although item difficulty occurs at the inflection point, it is not at .5 as is the case with the 1PL and 2PL. In this example, it is at about .6, which is indicated by the thin horizontal dotted line that connects the ICC to the vertical axis.

Estimation procedures

Person ability estimation is determined through an iterative process and for most IRT models is based on the pattern of correct and incorrect responses. For instance, in Table 27.1, Keiko, the first test taker, answered items 3, 1, 4, 10, and 12 correctly and all the other items incorrectly. The estimation of Keiko's ability is conceptually the same as asking which ability level has the highest probability of producing Keiko's pattern of correct responses. The question is answered via an iterative process that involves multiple steps. In the first step, the probability of each response in the pattern is calculated based on one of the IRT formulas, such as Equation 27.2. In the second step, the probability of the response pattern is

estimated by multiplying the probabilities of correctly answering each item. In step 3, steps 1 and 2 are repeated for a range of ability levels. That is, points along the ability continuum are selected, and the likelihood of each response pattern producing each ability estimate is compared. The process is iterative, and an ability estimate is produced when the desired degree of precision has been reached (de Ayala, 2009).

Many-facet Rasch measurement model

The many-facet Rasch measurement (MFRM) model has been used in the language testing field to model and adjust for various assessment characteristics on performance-based tests. MFRM is a one parameter model (with the discrimination parameter fixed to 1.0), but because test taker ability and item difficulty are treated as measurement facets, other characteristics of the measurement process, such as raters and scales, can be included in the model (Bond and Fox, 2015). When raters assign scores for performances-based assessments, the scores may be affected by various factors (see also Galaczi and Lim, this volume). For example, test takers may be required to write an academic essay to demonstrate their academic writing ability. The essays could then be rated on a five-point scale. In large-scale assessments, it may be necessary to use different raters, prompts, and times of the day for the assessment, and as a result, the scores may be affected by factors like rater severity, the difficulty of the prompt, the time of day that the test is administered, or the interaction of two or more of these factors. MFRM can be used to identify and estimate the magnitude of such effects and adjust the scores to compensate for them. The MFRM model used for modeling test taker ability, rater severity, and item difficulty for a rating scale can be expressed as follows:

Probability of rating =

$$\frac{\exp\left(\text{test taker ability} - \text{item difficulty} - \text{rater severity} - \text{rating step difficulty}\right)}{1 + \exp\left(\text{test taker ability} - \text{item difficulty} - \text{rater severity} - \text{rating step difficulty}\right)} \quad (27.4)$$

Like the models for items scored as either right or wrong, item difficulty is subtracted from test taker ability to arrive at a rating. The difference between this MFRM and the 1PL Rasch model for items scored as correct or incorrect is that the severity of the rater and the rating step difficulty are also subtracted from the test taker ability. Rater severity denotes how strict a rater is in assigning scores to test takers while rating step difficulty refers to how much of the ability is required to move from one step on a rating scale to another. For example, on a five-point writing scale with 1 indicating least proficient and 5 most proficient, the level of ability required to move from a rating of 1 to 2, or between any two scales, would be the difficulty of rating step. Figure 27.3 presents category response curves (CRCs) when rating scales are used for a five-point scale.

CRCs are analogous to ICCs. The probability of assignment of a rating on the scale – the five-point scale in this example – is shown on the vertical axis, and test taker ability level is shown on the horizontal axis. Each rating on the five-point scale appears as a hill. Progressing from left to right, the far left peak indicates a 1 on the five-point scale, and the far right peak indicates a rating of 5. The numbers 1 to 5 have been written above the curve for the writing scale that they represent. By comparing the point on each rating category to the

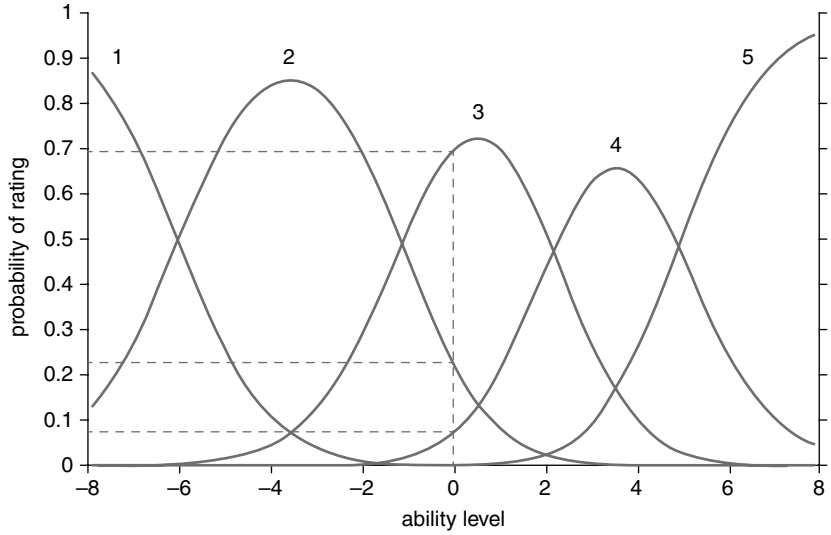


Figure 27.3 Category response curves for writing example

vertical and horizontal axes, it is possible to determine the probability of a test taker with a particular ability being assigned a particular rating on the five-point scale. Figure 27.3 includes dotted lines for an ability level of 0 on the logit scale to each rating category that has a probability of being assigned for this ability level. A test taker with an ability level of 0 would have a near 0 probability of a rating of 1, a little above a .2 probability of a rating of 2, about a .7 probability of a rating of 3, a little less than a .1 probability of a rating of 4, and a near 0 probability of a rating of 5.

The five-point writing scale in Figure 27.3 indicates that a score of 2 is commonly assigned since it extends the farthest along the horizontal axis. Ideally, rating categories are highly peaked and equivalent in size and shape to each other. Test developers can use the information in the CRCs to revise rating scales or inform rater training. For instance, faced with the CRC in Figure 27.3, it may be appropriate to revise the 2 rating scale to include less ability in the descriptor or to amend rater training procedures with the aim of reducing the ability range assigned for a 2.

Current contributions and research

IRT is currently used for various purposes in the language testing arena. Probably its most commonly reported use over the past three decades is for investigating rater and task characteristics and their effects on various types of performance-based assessments, most commonly through the lens of MFRM. For instance, MFRM has been used to investigate the effects of rater bias (Eckes, 2005; Kondo-Brown, 2002), rater training (Weigle, 1998), rater feedback (Elder *et al*., 2005; Knoch, 2011), task difficulty (Fulcher, 1996), task bias (Aryadoust, 2016; Batty, 2018; He *et al.,* 2013), and rating scale effectiveness (Bonk and Ockey, 2003; Li, 2019).

Applications of computer adaptive testing (CAT) are not commonly reported in the language assessment literature, likely because of the large number of items and test takers

required for its feasibility. However, it is used in some large-scale language assessments and is considered one of the most promising applications of IRT. Computer adaptive tests are designed to deliver assessments that are tailored to the ability levels of test takers. The approach capitalizes on the fact that items near test takers' ability levels are most useful for ability estimation. A computer is programmed to deliver items increasingly closer to the test takers' ability levels. In its simplest form, if a test taker answers an item correctly, the IRT algorithm assigns the test taker a more difficult item, whereas if the test taker answers an item incorrectly, the next item will be easier. The test is complete when a predetermined level of precision of locating the test taker's ability level has been achieved. Sawaki provides further explanation of CAT in Chapter 31.

Language testers also use IRT techniques to identify and understand possible differences in the way items function for different groups of test takers. Differential item functioning (DIF), which can be an indicator of biased test items, exists if test takers from different groups with equal ability do not have the same chance of answering an item correctly. IRT DIF methods compare ICCs for the same item in the two or more groups or conditions of interest. When the ICCs are different, the items are flagged for DIF. Some IRT DIF studies can be found in the language assessment literature. Ferne and Rupp (2007) discuss DIF techniques and early studies that were conducted in language testing. Ockey (2007) reports on the use of IRT DIF techniques to identify items on a National Assessment of Educational Progress math test that functioned differently for English learners and non–English learners. Koyama *et al*. (2016) demonstrate the use of DIF to identify items that function differentially across video-mediated listening assessments that include varying amounts of item preview.

Others have used an MFRM framework to investigate DIF. For instance, Aryadoust (2016) shows that it is possible to use MFRM bias analyses procedures to identify DIF (gender and area of study) when peers evaluate their classmates' oral presentations. Batty (2018) demonstrates an innovative use of a DIF analysis in an MFRM framework to show that the use of video in listening assessments (as compared to audio only) leads to implicit items becoming differentially easier than explicit items.

Recommendations for practice

When deciding whether or not to use IRT for language assessment projects, assumptions about the data necessary for appropriate application, along with the advantages and disadvantages of using IRT techniques for the intended purpose, should be considered.

Assumptions underlying IRT models

For the results of an IRT analysis to be meaningful, the data must conform to assumptions that underlie the theory. An exhaustive list of the assumptions of IRT will not be provided here. Instead, a few of the major ones are described. One assumption relates to dimensionality and, for most IRT models, unidimensionality. In a unidimensional data set, a single ability can account for the differences in scores. For example, a second language listening test would need to be constructed so that only listening ability underlies test takers' responses to the test items. A violation of this assumption would be the inclusion of an item that measured both the targeted ability of listening and reading ability not required for listening comprehension.

A second assumption underlying IRT, referred to as local independence, is that responses to each item on a test are independent of responses to all other items. This means that each item should be assessed independently of all other items. The assumption of local independence could be violated on a reading test when the question or answer options for one item provide information that may be helpful for correctly answering another item about the same passage.

Another assumption underlying IRT, sometimes referred to as certainty of response, is that test takers make an effort to demonstrate the level of ability that they possess when they complete the assessment. Test takers must try to answer all questions correctly because the probability of a correct response in IRT is directly related to their ability. This assumption is often violated when researchers recruit test takers for a study, and there is little or no incentive for the test takers to offer their best effort.

It is important to bear in mind that almost all data will violate one or more of the IRT assumptions to some extent. It is the degree to which such violations occur that determines how meaningful the resulting analysis is (de Ayala, 2009). Diagnostic tests should therefore be conducted and reported on each assumption so stakeholders can make informed decisions when placing trust in the reported estimates. De Ayala (2009) provides a description of statistical approaches for how to assess assumptions across various types of assessment data for particular contexts.

Sample size

Although not generally conceived of as an underlying assumption, a requirement of the effective use of IRT is a sufficient sample size. Many factors should be considered when determining the number of cases needed for a sound analysis. In general, smaller samples provide less accurate parameter estimates, and models with more parameters require larger samples for accurate estimates. High-stakes tests require more accurate estimates than low-stakes tests, and users must decide how much inaccuracy they are willing to tolerate. As a general rule, de Ayala (2009) recommends that the starting point for determining sample size for dichotomously scored items should be a few hundred. More or less may be appropriate given the context of the assessment. In MFRM studies, the focus should be on the number of data points rather than the sample size. It could be that results from 40 test takers is sufficient if there are many ratings for each of them. Because the larger the number of data points for a given estimate, the more accurate it is, MFRM users should pay particular attention to the number of data points and the relevant standard errors when interpreting findings. For example, if the speaking proficiency of 300 test takers was judged by two raters, each test taker's score would be subject to a great amount of error, while the estimate of the rater's level of severity would have little error. Researchers need to be careful to interpret the findings accounting for these differences in sample sizes for estimation of the different facets in the analysis.

Advantages of IRT

IRT has advantages over CTT (see Brown, this volume, for an overview of CTT). First, if ability estimates are drawn from the population of interest, they are group independent. This means that ability estimates are not dependent on the particular group of test takers who complete the assessment. Second, IRT can be used to aid in designing instruments that target

specific ability levels based on the TIF. For instance, an assessment for the purpose of English certification of international airline pilots would require a specific cutscore targeted to separate qualified and unqualified individuals. Using IRT item difficulty parameters makes it possible to design items with difficulty levels near the desired cutscore, which would increase the accuracy of decisions at this crucial ability location. Third, IRT can provide information about various aspects of the assessment process, including items, raters, and test takers, which can be useful for test development. For instance, raters who have inconsistent rating patterns, are too lenient, or are biased when rating particular tasks can be identified. These raters can then be provided with specific feedback on how to improve their rating behavior.

Disadvantages of IRT

In the early 1990s, it was believed that the use of IRT would become increasingly prevalent in the language testing field (Garret, 1991). For the next decade or so, however, the use of IRT grew very modestly. Over the past decade, however, the use of IRT has begun to blossom in the field, particularly the use of MFRM. This growth has coincided with a desire for more authentic assessments, ones that require test taker performances and accompanying judgments of these performances. Language assessment programs are commonly teaching IRT courses, and many of the new generation of language assessment researchers are armed with IRT knowledge and expertise in using software, which effectively uses IRT principles to analyze data. McNamara and Knoch (2012) provide a discussion of MFRM's battle to be accepted by language assessment researchers from its introduction in the early 1990s to its general acceptance by 2010. Nonetheless, IRT may not have reached the standing that most would have predicted three decades ago. This is likely because of the need for large sample sizes and the expertise needed to effectively use IRT principles and software.

Future directions

The use of IRT is likely to continue its recent growth trend. IRT software programs are becoming increasingly user friendly – e.g., FACETS (Linacre, 1987–2019) – and accessible texts for a language assessment audience with limited statistical background, e.g., *Statistical Analyses for Language Testers* (Green, 2013), *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (Eckes, 2015), and *Fairness, Justice, and Language Assessment* (McNamara *et al*., 2019), are increasingly available.

The use of IRT in the language assessment field has largely been through the lens of MFRM for aiding in the understanding of various facets of the human rating process. As computer scoring becomes more prevalent for evaluating performance assessments, such as writing and oral communication, it is likely that the role of IRT in the language assessment field will change. During the transition period from human to computer scoring, which is where the field is now, IRT is likely to play an increasing role in aiding in the comparison of human- versus computer-scored performance assessments. Once computer scoring is the norm, assuming this is how the future plays out, it is likely that IRT will be used more for ensuring psychometric properties of tasks and items prior to test administration. This is common in high-stakes assessment contexts now, but it is likely to become more common in lower-stakes contexts.

The role of IRT is unlikely to decrease for language assessment researchers and practitioners in the near future. It is therefore recommended that people in the field develop their knowledge in this area.

Further reading

Bond, T. C. and Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd edn. Mahwah, NJ: L. Erlbaum.
The authors provide a transparent introduction to the Rasch model. The book is written for an audience without advanced mathematical knowledge. Various examples are described, and many are accompanied by interpreted computer output. An emphasis of the book is to convince readers that the 1PL Rasch model is appropriate for all situations.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. New York, NY: Peter Lang.
In this book, Eckes masterfully demonstrates appropriate uses of MFRM and the FACETS software program for language assessment practitioners and researchers. Eckes makes challenging information transparent for a non-technical language assessment audience. This book can be used as a textbook for language assessment graduate courses.

Green, R. (2013). *Statistical Analyses for Language Testers*. Basingstoke: Palgrave Macmillan.
In this book, Green begins with an overview of classical test theory (CTT) statistical analyses procedures before providing a light introduction to item response theory (IRT). Green does a nice job of comparing principles of IRT with those of CTT. This book is appropriate for readers who have little or no background knowledge in statistical approaches to language assessment analyses and want a basic knowledge of both CTT and IRT for language assessment.

McNamara, T., Knoch, U. and Fan, J. (2019). *Fairness, Justice, and Language Assessment*. Oxford: Oxford University Press.
This book, which is a revised version of McNamara's (1996) *Measuring Second Language Performance*, is a must-read for language assessment researchers who are interested in performance-based assessments and/or MFRM. The book is both conceptual and practical. It provides a very clear introduction to the principles underlying the Rasch model as well as interpretations and explanations of output from the FACETS program. It demonstrates how to apply MFRM and the FACETS software program on contemporary issues of fairness in language assessment.

References

Aryadoust, V. (2016). Gender and academic major bias in peer assessment of oral presentations. *Language Assessment Quarterly* 13: 1–24. https://doi.org/10.1080/15434303.2015.1133626.

Batty, A. (2018). Investigating the impact of nonverbal communication cues on listening item types. In G. J. Ockey and E. Wagner (eds.), *Assessing L2 Listening: Moving Towards Authenticity*. Philadelphia, PA: John Benjamins, 161–75.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice* 16: 21–33. https://doi.org/10.1111/j.1745-3992.1997.tb00605.x.

Bond, T. C. and Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*, 3rd edn. Mahwah, NJ: L. Erlbaum.

Bonk, W. and Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing* 20: 1, 89–110. https://doi.org/10.1191/0265532203lt245oa.

Davidson, F. and Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing* 2: 164–79. https://doi.org/10.1177/026553228500200205.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guilford Press.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly* 2: 197–221. https://doi.org/10.1207/s15434311laq0203_2.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement*. New York, NY: Peter Lang.

Elder, C., Knoch, U., Barkhuizen, G. and von Radow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly* 2: 175–96. https://doi.org/10.1207/s15434311laq0203_1.

Ferne, T. and Rupp, A. (2007). Synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly* 4: 113–48. https://doi.org/10.1080/15434300701375923.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing* 13: 23–51. https://doi.org/10.1177/026553229601300103.

Garret, N. (1991). Technology in the services of language learning: Trends and issues. *The Modern Language Journal* 75: 74–101. https://doi.org/10.1111/j.15404781.2009.00968.x.

Green, R. (2013). *Statistical Analyses for Language Testers*. Basingstoke: Palgrave Macmillan.

He, T. H., Gou, W. J., Chien, Y. C., Chen, I., Shan, J. and Chang, S. M. (2013). Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment. *Psychological Reports* 112: 469–86. https://doi.org/10.2466/03.11.PR0.112.2.469-485.

Henning, G. (1984). Advantages of latent trait measurement in language testing. *Language Testing* 1: 123–33. https://doi.org/10.1177/026553228400100201.

Henning, G., Hudson, T. and Turner, J. (1985). Item response theory and the assumption of unidimensionality for language tests. *Language Testing* 2: 41–154. https://doi.org/10.1177/026553228500200203.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing* 28: 179–200. https://doi.org/10.1177/0265532210384252.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing* 19: 3–31. https://doi.org/10.1191/0265532202lt218oa.

Koyama, D., Sun, A. and Ockey, G. J. (2016). The effects of item preview on video-based multiple-choice listening assessments. *Language Learning & Technology* 20: 148–65. http://llt.msu.edu/issues/february2016/koyamasunockey.pdf.

Li, S. (2019). Variations in rating scale functioning in assessing speech act production in L2 Chinese. *Language Assessment Quarterly* 16: 271–93. https://doi.org/10.1080/15434303.2019.1648473.

Linacre, J. M. (1987–2019). *Many Facets Rasch Measurement, Version 3.83.0*. Chicago, IL: Mesa Press.

Lord, F. M. (1952). *A Theory of Test Scores*. Psychometric Monograph No. 7. Richmond, VA: Psychometric Corporation.

Lord, F. M. and Novick, M. (1968). *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

McNamara, T. (1996). *Measuring Second Language Performance*. New York, NY: Addison Wesley Longman Limited.

McNamara, T. and Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing* 29: 555–76. https://doi.org/10.1177/0265532211430367.

McNamara, T., Knoch, U. and Fan, J. (2019). *Fairness, Justice, and Language Assessment*. Oxford: Oxford University Press.

Ockey, G. J. (2007). Investigating the validity of math word problems for English language learners with DIF. *Language Assessment Quarterly* 4: 149–64. https://doi.org/10.1080/15434300701375717.

Ockey, G. J. and Choi, I. (2015). Item response theory. In C. A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*. Oxford: John Wiley and Sons, Inc.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.

Thurston, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology* 16: 433–51. https://doi.org/10.1037/h0073357.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing* 15: 263–87. https://doi.org/10.1177/026553229801500205.

Woods, A. and Baker, R. (1985). Item response theory. *Language Testing* 2: 117–40. https://doi.org/10.1177/026553228500200202.

28

Reliability and dependability

Xun Yan and Jason Fan

This chapter introduces the concept of reliability in language testing research and practice. It provides a non-technical discussion of (1) conceptualizations of reliability in relation to the interpretation of test scores, (2) operationalizations of reliability for both selected- and constructed-response items, (3) factors influencing reliability for both standardized tests and classroom-based assessments, and (4) the relationship between reliability, validity, and fairness. This chapter is also intended to complement the introduction to different measurement theories applied to testing language proficiency presented in the other chapters of this section (Brown, this volume; Ockey, this volume).

In psychometrics, reliability is conceptualized as the necessary but insufficient condition of validity (AERA *et al*., 2014; Bachman, 1990). That is, validity cannot exist without reliability; consequently, reliability has been considered a psychometric quality central to language testing research and practice. However, more recent conceptualizations suggest that the relationship between reliability and validity is more complex than previously construed. That is, to some extent, validity can exist without (psychometric) reliability (Moss, 1994, 2004), and, as will be discussed later, reliability and validity can even be in conflict with each other (Spolsky, 1995). In this chapter, while we acknowledge this complex relationship, reliability will be presented as an integral component of validity as well as fairness. That is, if a test is unreliable, the inferences we draw from its scores are unlikely to be valid (Kane, 1992) or fair (McNamara *et al*., 2019). Approaches to estimating reliability can be utilized to gather evidence to support multiple layers of inferences in the validity or fairness argument of a test (Knoch and Chapelle, 2018; Xi, 2010).

Measurement, error, and reliability

Measurement is the assignment of a number to a characteristic of an object, an event, or a person, which can be compared with other objects, events, or persons. Measurement is deeply embedded in the idea of comparison. As we have the need to compare properties across objects or across time and space within the same object, we need to develop a system that would allow us to measure these properties. In research, we tend to make a distinction between two large categories of measurement: namely, physiological measurement and

DOI: 10.4324/9781003220756-37

psychological measurement. To some extent, this distinction coincides with the division between natural sciences and engineering and social and behavioral sciences. The fields of natural sciences and engineering tend to quantify properties of objects through unified physiological measurement (e.g., measuring temperature with a thermometer, weight with a scale, or length with a ruler). This is perhaps the canonical definition of measurement to many, as it is arguably the basis of all scientific inquiries. The idea that quantification is the way to understanding was memorably expressed by Kelvin in 1883:

> [W]hen you can measure what you are speaking about, and express it in numbers you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.
>
> *(Kelvin, quoted in Stellman, 1998: 1973)*

Reliability is concerned with measurement error. In fact, all measurements have error, though to different degrees. Reliability is conceptualized as the degree to which a measurement lacks measurement error. The unified physiological measurements tend to be consistent across conditions and susceptible to very little measurement error. For example, if you measure your weight with a scale and you step on the scale several times, the results will be practically the same. That is an indication of high measurement reliability. However, when you measure your weight in the morning versus at night, the weight might show some discernible differences. Yet the differences are so small that they could be ignored for practical purposes. Thus, reliability is rarely a concern for physiological measurements. However, social and behavioral sciences focus on the measurement of psychological constructs (e.g., personality, depression, language proficiency). Psychological constructs are observable in real life; however, unlike physiological constructs (e.g., weight and height), they are not easily quantified, nor do they have unified scales or standards of measurement that are easily generalizable across contexts and disciplines: e.g., weight can be measured in kilograms (the metric system) or pounds (the imperial system), and the metric and imperial systems can be mathematically converted.

Measuring a person's language proficiency is quite different from measuring, say, his or her body temperature, which can be performed with a thermometer. Although there are different types of thermometers, any type is able to provide a sufficiently consistent measurement of body temperature for medical purposes. However, when it comes to language proficiency, the "tools" (i.e., language tasks) used to assess language proficiency might differ substantially in type, format, delivery mode, and quality. For example, to measure reading skills, one can use reading comprehension or cloze tests; among reading comprehension tasks, one can use either multiple-choice or short-question formats; moreover, the same set of reading comprehension tasks can be delivered either on paper or a computer platform. These differences might lead to inconsistencies in the proficiency scores they generate. If the test scores are used for high-stakes decisions (e.g., college admissions), the inconsistencies in test scores might have severe consequences on the test takers. Thus, reliability becomes a central concern to research in language testing and assessment.

In the following sections, we will introduce the conceptualizations of reliability and different approaches to operationalizing it. We will then talk about factors that influence the reliability of language tests. Finally, we will discuss the relationship between reliability, validity, and fairness.

Conceptualizations and operationalizations of reliability

The definition of reliability in measurement is somewhat counterintuitive. When we consider something or someone "reliable" in common parlance, the term is synonymous with *trustworthy* or *accurate*. Feldt and Brennan (1989) point out that "in everyday conversation, 'reliability' refers to a concept much closer to the measurement concept of *validity*"; that is, it implies that the person is able to give accurate information or solve a problem. However, in language testing, reliability has the narrower meaning of "consistency." As Feldt and Brennan (1989) stated, "quantification of the consistency and inconsistency in examinee performance constitutes the essence of reliability analysis" (p. 105).

In fact, the concepts of "accuracy" and "consistency" do not conflict with each other. Reliability is canonically defined as the degree to which a test lacks measurement error or the extent to which test takers' scores reflect the relative abilities (i.e., accuracy). In practice, if a reliable test is repeatedly administered to the same test taker, it will produce the same or similar results. The degree of score fluctuation across all test administrations indicates the level of measurement error or the lack of reliability. This is also the definition of *test-retest reliability*, the canonical form of reliability. The smaller the fluctuation, the more reliable the test. In addition, on a reliable test, test takers of different abilities should consistently receive different scores.

The estimation of reliability originates from the assumption that the score a test taker receives on a test (i.e., observed score) consists of two components: namely, the true score (i.e., the score the test taker deserves in theory) and the measurement error. As test takers of different ability levels will receive different test scores, the differences in test scores across individuals can be computed in terms of *score variance*. Reliability (or *reliability coefficient*) is computed as the proportion of the score variance that is associated with variability in test takers' true ability and not other factors. A reliability coefficient is typically described as a correlation (or correlation-like) coefficient. There are multiple types of reliability. In the following section, we will introduce five commonly used reliability coefficients: namely, *test-retest reliability*, *parallel-form reliability*, *split-half reliability*, *internal consistency*, and *rater reliability*. Depending on the type of reliability being analyzed, what is correlated with what will change; however, all reliability coefficients are interpretable in the sense of repeated measurement described earlier. A perfectly reliable test would have a reliability coefficient (r) of 1. In addition, while the first four types of reliability tend to apply to tests with selected-response items, rater reliability applies to constructed-response tests or items involving rater judgment.

Reliability for selected-response items

Test-retest reliability assumes repeated administration of the same test to the same group of test takers, which is expressed in terms of the correlation between test takers' scores on both test administrations. The closer the correlation is to 1, the higher the test reliability. While the interpretation of test-retest reliability is straightforward, it has three major flaws. First, we normally do not administer the same test twice to the same group of test takers. It is possible to recruit volunteers to take the same test twice, but it is generally difficult to get a sufficiently large number of test takers in order to obtain a good estimate of test-retest reliability. Large-scale tests might have a sizable group of repeaters (those who take the test more than once), and it is possible to calculate test-retest reliability using their test

scores. However, repeaters represent a special subpopulation of test takers; typically, test takers retake a test in order to achieve higher scores for admissions or other purposes (i.e., the scores they receive in their first attempt tend to be on the low end of the scale). Thus, the repeaters might represent a restricted range of proficiency levels on the test. As will be discussed in the section on factors influencing reliability later, the restricted range will likely lead to a weaker correlation between scores on repeated testing, thereby underestimating the actual test-retest reliability. Second, test-retest reliability is highly susceptible to memory or practice effect. That is, if the test takers memorize the content of a test, it is highly likely that the test will produce different scores on the second administration, especially when the test has high stakes. Third, even if test-taking strategy is less of a concern here, the time interval between the two test administrations also influences test-retest reliability. If learning takes place in between test administrations, the scores will fluctuate, even if the test has little measurement error, leading to an underestimation of the reliability of the test.

The shortcomings of test-retest reliability can be addressed by two alternative types of reliability: namely, parallel-form reliability and split-half reliability. Parallel-form reliability, as the name suggests, is expressed through the correlation between test scores on two comparable forms of the same test. This can address the memory or practice effect, if the two tests are given to the same group of test takers. However, in operational testing, it is still less practical to ask the same group of test takers to take the test twice. It is possible to randomly assign the parallel forms to different test takers and compute the correlation between their scores; however, the legitimacy of this type of reliability relies on a rather strong assumption that the test takers taking the two forms are comparable in the distribution of abilities. To create these circumstances would require random assignment of test takers to the parallel forms and a large sample of test takers.

In comparison with parallel-form reliability, split-half reliability does not require repeated testing. Instead, a test is split into two equal halves, and the correlation is computed between test takers' scores on the two halves. This type of reliability relies on the assumption that test items on the same test measure the same abilities. If this assumption is largely true, the two halves can function as two separate tests, which can overcome the shortcomings of repeated testing in test-retest reliability. However, how to split a test can make a big difference in the comparability of the two halves. One way is to split the test right in the middle. This can be problematic if the items in the first half measure different subskills from those in the second half. Split-half reliability is also a problem in the case of speeded tests in which many test takers are not able to finish all the test items on time. Even when the test is not speeded, test takers might perform worse on the second half of the test due to fatigue. Alternatively, it is more desirable to split the test by even-odd numbers, which could, to some extent, address the problem of test speededness and fatigue. However, a potential problem with this approach is that some items may be locally dependent; in other words, test takers' responses to these items may depend on whether they can get the preceding item(s) correct. In that case, the test items cannot be simply split by even-odd numbers. For these reasons, the use of split-half reliability has been decreasing in language testing as well as other fields where measurement reliability is examined.

Internal consistency is another type of reliability that is similar to but an improvement on split-half reliability. In fact, split-half reliability is also an estimate of internal consistency, as it assumes that items that measure the same construct should be correlated with one another, thereby making the whole test internally consistent. The reliability coefficients that are typically used to estimate internal consistency include Cronbach's alpha and Kuder-Richardson formulas (KR-20 and KR-21). The formulas used to compute these coefficients

are less straightforward to explain (see Brown, this volume, for further detail). However, in contrast to split-half reliability, which looks into the correlation between two halves of the test, internal consistency can be conceived as examining the correlation between every pair of test items by computing the *covariance* among scores on individual items and then obtaining the "average correlation" among all pairs. Thus, compared to all other forms of reliability discussed earlier, internal consistency is a less biased estimate of test reliability. As a result, internal consistency is interpreted as the extent to which items are measuring the same construct, or whether the test as a whole is able to discriminate among test takers who have mastered the assessed language construct and those who have not. For this reason, internal consistency has also been called an index of *homogeneity* (Gulliksen, 1950: 220). Internal consistency has the great practical advantage of requiring response data from a single operational test administration, unlike test-retest or parallel-form reliabilities. Thanks to their practicality, internal consistency estimates of reliability are by far the most commonly used. However, it is important to point out that a language test tends to have different sections, each of which targets a slightly different subskill. One can reasonably expect that scores on items in different sections tend to be less correlated than items within the same section. If the items across sections are included indistinguishably in the computation of internal consistency of a test, it will likely result in a low reliability coefficient. To address this problem, it is common practice to compute internal consistency for each section of a test in addition to an overall reliability estimate for the whole test. By doing so, test developers and users are able to obtain more fine-grained information about the reliability of a test.

Closely associated with test reliability is the concept of standard error of measurement (SEM). SEM is a transformation of reliability in terms of test scores. While reliability refers to the consistency of scores of a group of test takers, the SEM shows the impact of reliability on the score of an individual. It indicates how close a test taker's score is likely to be to their "true score." This impact is often expressed in terms of a confidence interval, which is a range of possible scores within which a test taker's true score is likely to fall. This range is computed based on the value of SEM (see Brown, this volume, for examples).

Reliability for constructed-response items

The forms of reliability discussed earlier are suitable to estimate the reliability of tests that consist largely of selected-response items because the computation of those coefficients tend to require a large number of test items. Thus, they are more often used on reading or listening tests. In contrast, speaking or writing tests tend to have very few items. It is less feasible to compute reliability indices introduced earlier in this chapter. These tests often use constructed-response items, which require human judgment. Thus, rater reliability becomes the major type of reliability for speaking and writing assessments (see Davis, this volume; Galaczi and Lim, this volume).

One can examine both intra- and inter-rater reliability. Intra-rater reliability is concerned with the internal consistency of individual raters, which examines whether each rater will assign the same or similar scores to the same examinee (if rated more than once) or examinees with the same or similar abilities. In comparison, inter-rater reliability indicates the extent to which multiple raters can reach consensus in their ratings on the same test takers. In this case, it is helpful to consider raters as comparable test items as each rater's evaluation represents an independent measurement of the test taker's speaking or writing ability. Although it is less feasible to ask test takers to complete multiple speaking or writing tasks in one sitting, it is relatively easier to ask multiple raters to evaluate the same test response at

once. If the raters are "consistent" with one another, they should produce the same or similar ratings for each test taker. When a testing agency hires a group of raters, it is important to assign raters such that any pair of raters would share some common test responses in their rating, although each of them can rate a different set of test responses. This will allow for the examination of inter-rater reliability.

There are several estimates for rater reliability. Stemler (2014) categorizes these estimates into three types: namely, consensus estimates, consistency estimates, and measurement estimates. Consensus estimates of interrater reliability examines the extent to which two raters reach exact agreement in their score assignment. This type of estimate tends to take the form of percentage-agreement figures. Sometimes, if it is difficult to reach a high level of exact agreement between raters, adjacent scores will also be counted as consensus (e.g., on a five-point rating scale, the scores assigned by the two raters are only one point apart). Another commonly used type of consensus estimate is Cohen's kappa. The kappa statistic is based on percentage-agreement figures but also accounts for the likelihood of agreement by chance; that is, even if the scores of raters fluctuate for a range of reasons, there is still an estimable level of likelihood that they will agree by chance. While Cohen's kappa is an improvement over percentage-agreement figures in representing reliability, it has also been criticized as being an overly conservative estimate of rater reliability as it assumes that all raters would play the guessing strategy.

Consistency estimates of rater reliability are less stringent than consensus estimates. They are based on the assumption that it is not really necessary for two raters to always reach exact agreement, so long as they are each consistent in rank ordering the examinees according to their relevant abilities. This type of rater reliability can be examined by correlations (e.g., Pearson's *r*, Spearman's rho, and Cronbach's alpha), instead of agreement percentage, in score assignment between two raters. If two raters rank order test takers in similar ways, the correlation between their scores will be close to 1; in contrast, if their rankings differ substantially, their correlation will be close to 0. It is also possible to obtain a consistency estimate for a group of raters using Cronbach's alpha. In that case, individual raters are treated as individual test items in the case of internal consistency for tests with selected-response items. The interpretation of this coefficient in this case becomes the extent to which the raters function interchangeably when rating the test performances.

Measurement estimates of rater reliability are more complex than the first two types of estimate. As the behavior of human raters is influenced by a range of psychological factors, an understanding of how these factors influence their ratings can help in devising approaches to correct for those factors, thereby enhancing rater reliability. Measurement estimates utilize sophisticated statistical models to perform a fine-grained analysis of rater performance and the relationships among these factors influencing their performance. There are two types of analyses that are commonly performed on rater performance: namely, many-facets Rasch measurement theory (see Ockey, this volume) and generalizability theory. We will only provide a conceptual understanding of them here. The Rasch model, named after Georg Rasch, is a family of psychometric models for analyzing test scores that assume the scores test takers receive to be a function of the tradeoff among the test takers' ability, the item difficulty, and the severity of the raters, as well as other potential factors in an assessment context. This model can provide fine-grained information about the nature of each factor and thus has been frequently utilized in language testing research to analyze individual raters' behavior. Generalizability theory relies on a similar set of assumptions about the factors influencing test scores, including the raters. However, in contrast with the Rasch model, generalizability theory examines the proportion of score variance associated

with differences in rater behavior and provides an estimate of the number of raters that a test needs in order to achieve a desired level of rater reliability. Because of the complex nature of rater behavior, rater performance and reliability are among the most researched areas in performance assessment (McNamara and Knoch, 2012; Fan and Knoch, 2019).

Factors influencing test reliability

An important note about reliability, or validity, is that it is not an inherent property of a test itself. First, reliability and validity are more concerned with the scores and the inferences about the test takers we draw from the scores. As a result, reliability and validity are context dependent, which means they can be influenced by a range of factors related to test takers, test administration conditions, and test uses. These factors influence test takers' scores and inevitably create measurement error. However, there is a difference between errors that can be explained and controlled for (i.e., systematic errors) and errors that cannot (i.e., random errors). Figure 28.1 illustrates four major sources of error which are discussed by Feldt and Brennan (1989: 107): namely, individual variation, test situation, rater variation, and instrumental variation. While the first two sources of error are more difficult to control or systematically account for, the third and fourth sources could be seen as potentially affecting the very nature of the ability that is being tested and thus impact on the reliability as much as the validity of the test. We will elaborate on these two sources.

The impact of raters (markers, evaluators) is prominent on constructed-response tasks. In language testing, constructed-response tasks are most frequently used in speaking and writing tests. Rater behavior can be influenced by a number of factors, and their impacts can manifest in different ways (known as rater effects). Specifically, raters can differ in severity. For example, in language testing, it is possible that two raters are able to rank order the test takers in the same way, but one rater is consistently more lenient than the other, therefore assigning higher scores to every performance that he or she happens to rate. If two test takers of the same speaking or writing ability are rated by a harsh and a lenient rater, respectively, it is likely that they will receive different scores. Sometimes, even the

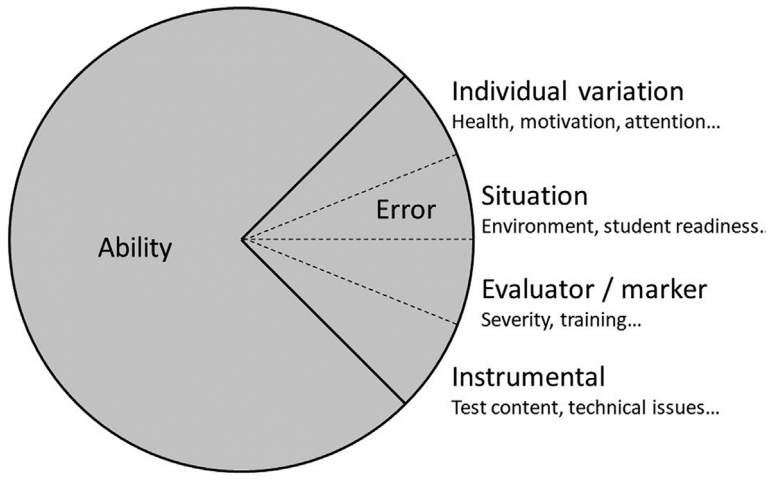


Figure 28.1 Some sources of error in a test

same rater can show differential severity to different test takers. This is especially common in speaking assessment, where raters are often asked to evaluate the pronunciation of the test takers (see Nakatsuhara *et al*., this volume). Second language speakers can show distinct accents that deviate from the native norms, and their levels of accentedness might impact the intelligibility of their speech and require listener effort in order to comprehend the meaning of their speech. However, if raters happen to be familiar with the accents of a particular group of test takers (e.g., the rater and test takers share the same first language background, the rater speaks the first language of the test takers as a second language, or the rater has been exposed frequently to speakers of that first language), they might rate that group of test takers more favorably. Similarly, a rater's reliability is influenced by his or her experience in rating and rater training, and their rating behavior might also fluctuate over time and across rating conditions. To combat these rater effects, it is important to provide ongoing rater training to align raters to the scale and monitor their performance over time (Yan and Ginther, 2017).

The impact of instrumental factors on test reliability can be observed on both selected-response and constructed-response tests. For selected-response tests, test length makes a difference. In general, the more items a test has, the more opportunities that test takers have to demonstrate their ability, and thus the more reliable the test is. Unfortunately, in most cases, test developers have little choice in deciding how long a test can be. However, careful control of the content and format of the test items or tasks can enhance the reliability of a test. When writing selected-response items, especially multiple-choice items, it is important to make items less susceptible to guessing. More importantly, test developers need to strive to create items that will consistently elicit the use of only the target knowledge, skill, or ability, and nothing else, to distinguish the answer key from the distractors. To that end, the content of the test items should be presented in language at appropriate complexity level and not be biased toward or against any subgroups of test takers. The two multiple-choice items in Figure 28.2, both of which are designed to assess test takers' knowledge of the word *ferry*, serve to demonstrate this point. Item 1 seems to function well to be included in an ESL placement test for a university in New York. Test takers need to know that the Statue of Liberty is on an island and that a ferry is a means of public transportation that travels on water. While the meaning of the word *ferry* is the target knowledge, the item also requires the test takers to have the background knowledge about New York. This should work well for students admitted into universities in New York, but not when it is administered to English as a foreign language (EFL) learners who lack this background knowledge. In this case, a content

1. You need to take a _____ to get to the Statue of Liberty.
   A. bus
   B. ferry
   C. taxi
   D. train

2. You need to take a _____ to get to the island.
   A. bus
   B. ferry
   C. taxi
   D. train

Figure 28.2 Two sample multiple-choice items measuring vocabulary

or fairness review would help item writers or test providers avoid such potentially biased items. In contrast, Item 2 would not require such background knowledge and instead only require knowledge of the word *island* from the test takers. Similarly, constructed-response items, especially integrated speaking or writing tasks, tend to use source material (e.g., text, audio, picture, or graph). While it is difficult to predict how test takers go about completing the task, the source materials should be carefully controlled in terms of content and complexity to allow test takers to demonstrate their levels of language performance.

Although it belongs to situational factors, a test's reliability is often influenced by range restriction in test takers' proficiency. The impact of restricted range can be easily illustrated (see Figure 28.3). We can take the canonical form of reliability, test-retest reliability, as an example. Suppose we administer a multiple-choice reading test (with a total score of 100) to 25 test takers, and their scores range between 70 and 100. Figure 28.3 is a scatterplot that plots the test takers' scores at Time 1 on the x-axis and scores at Time 2 on the y-axis. When the scores are plotted, each dot represents an individual test taker, and its coordinates represent their scores on the two test administrations. A strong correlation between the two times (evidence of high reliability) would show that all the dots are located roughly around a straight line with a 45-degree slope. This would be the case in Figure 28.3a, and the correlation coefficient is 0.81. However, if we truncate the sample of test takers and only take those who scored between 80 and 90 on both test administrations, we would end up with a much weaker correlation (r = .53), and the correlation is visualized in Figure 28.3b. What this suggests is that it is easier to achieve higher reliability if a test is designed for and administered

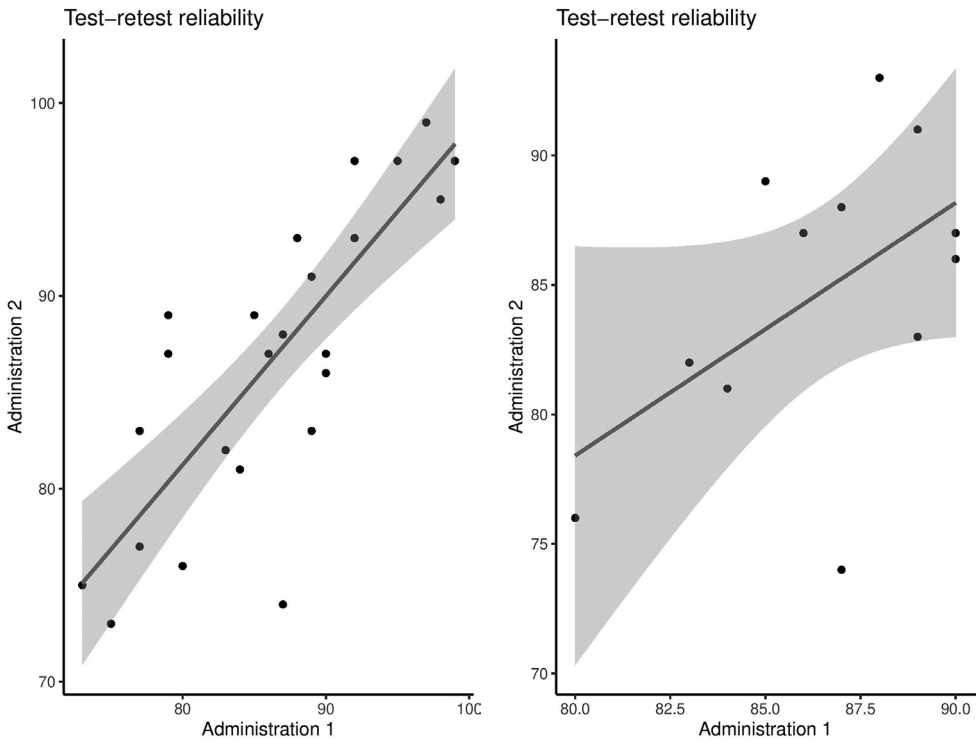


Figure 28.3 Scatterplots of test scores on two administrations

to test takers of a wide range of proficiency levels. While the restriction in range is less likely to happen in large-scale tests, local language tests, especially post-admissions tests or those designed for a subpopulation of test takers, might be easily susceptible to this problem. In fact, this is a common problem in educational research, where the academic and/or language abilities of the students in university contexts tend to be restricted in range due to admissions or other types of selection procedures (i.e., *direct range restriction*, Wiberg and Sundstrom, 2009). Direct range restriction leads to an underestimation of the correlation between the ability and other variables. Therefore, on local post-admissions language tests, it is more difficult to achieve high reliability. Take a typical English as a second language (ESL) placement test as an example. This kind of test is very common at universities where English is the medium of instruction (EMI), commonly used to place ESL students into different levels of language-support courses. While students who are admitted into the university tend to have scores above a certain cutoff on TOEFL iBT or IELTS, the purpose of the placement test is to make finer distinctions among them. Thus, it takes higher-quality items to make that distinction.

The concept of dependability

The concept of *dependability* is closely related to that of reliability. However, the two terms differ in that the reliability estimates discussed earlier apply to norm-referenced tests, whereas dependability is more applicable to criterion-referenced tests. Taken literally, norm-referencing means interpreting a learner's performance relative to other learners, while in criterion-referenced testing, test scores are interpreted relative to some external criterion, such as national or international standards of language performance (e.g., Common European Framework Reference, China's Standards of English). While the reliability of norm-referenced tests is concerned with consistency of rank ordering test takers, dependability refers to the consistency of classification: for example, assigning letter grades to test takers or classifying learners into masters or non-masters of a domain of knowledge. Usually, a master is defined as a student who knows the material or has mastered the ability being tested while a non-master is a student who fails to demonstrate such mastery.

The dependability of a test is estimated based on agreement in classification between repeated testings, similar to the calculation of consensus estimates of rater reliability mentioned previously. There are several approaches to estimating dependability. They generally fall into one of three categories: threshold loss agreement, squared-error loss agreement, and domain score dependability. The formulas for the latter two types of dependability are explained in detail in Brown (1989) and Brown and Hudson (2002). Here, we will explain two commonly used approaches to estimating threshold loss agreement: namely, agreement coefficient and kappa coefficient. These two approaches are straightforward to calculate and interpret and can be used in most large-scale and classroom-based assessment situations.

The canonical approach to estimating dependability is through agreement coefficient (often noted as P_0), which is a percentage figure of agreement. The calculation can be easily illustrated, as in the example presented in Table 28.1. Suppose two parallel forms of an achievement test are administered to a group of N students in a grammar course in an ESL program. Based a predetermined cut point on the total scores, students can be classified into masters or non-masters of the target grammatical structures in the course, and only those who are classified as masters will advance into the next level of ESL courses. The cut points are usually determined by the purpose of the test or through standard setting procedures.

Table 28.1 Classification of test takers on two test administrations

		Administration 2		*Total*
		Mastery	*Non-mastery*	
Administration 1	Mastery	A	B	A+B
	Non-mastery	C	D	C+D
Total		A+C	B+D	A+B+C+D = N

In Table 28.1, we first count the number of students who are classified as masters in both administrations (cell A) and then count the number of students who are classified as non-masters in both administrations (cell D). Thus, A+D would give us the number of students on whom the two forms have consensus in classification. However, in order to compute the agreement coefficient, we also need the total number of students to get the percentage of agreement. Thus, the agreement coefficient (P_0) can be expressed as (A+D)/(A+B+C+D). The higher the percentage figure, the more dependable the test is.

However, since the classification has only two levels, it is possible that the two tests might have the same classification by chance. Thus, it makes sense to account for agreement by chance. This is exactly what kappa coefficients do. The formula used to calculate kappa coefficient (k) is $k = (P_0 – P_{chance})/(1 – P_{chance})$, where P_{chance} is agreement by chance. The formula used to calculate agreement by chance is $P_{chance} = [(A+B)(A+C) + (C+D)(B+D)]/N^2$. Because kappa removes the agreement by chance, it is a more conservative estimate of dependability. We can use an example to compare the values of agreement and kappa coefficients. Suppose the two forms of the test are administered to 100 (N) students in the ESL program. Among them, 59 (A) students were classified as masters and 18 (D) as non-masters by both test forms. In addition, 12 (C) students were classified as masters in the first test form but non-masters in the second, and 11 (B) students were classified as non-masters in the first test form but masters in the second. (The same information is also presented in Table 28.2.) Thus, the number of students receiving the same classification (A+D) is 59 + 18 = 77; the agreement coefficient is 77/100 = .77. In order to calculate kappa coefficient, we need to first calculate the proportion of agreement by chance (P_{chance}). Given the formula provided earlier, P_{chance} in this case would be $[(59+11)(59+12) + (12+18)(11+18)]/ 100^2 = 0.58$, which is quite high. If we plug P_0 and P_{chance} into the kappa formula, then K = (0.77–0.58)/(1–0.58) = 0.45. Thus, we can see that the kappa coefficient is considerably smaller than the agreement coefficient. In practice, threshold lost agreement statistics can be used for two test administrations, evaluators (e.g., two teachers), and even researchers. Consequently, it is widely used to estimate dependability or consistency in a range of contexts (e.g., the classification accuracy of automated scoring system, inter-coder reliability in qualitative analysis), although there are also a variety of transformations of the kappa statistics.

Reliability, validity, and fairness

As we discussed earlier, by and large, reliability can be considered as a necessary but insufficient condition of validity. That is, the score of a reliable test should not be taken as a correct or accurate measure of what the test purports to measure. We can demonstrate this point with the same example of the weight scale. Suppose you are using a bathroom scale with

Table 28.2 An example of classification of test takers for computing kappa coefficient

		Administration 2		*Total*
		Mastery	*Non-mastery*	
Administration 1	Mastery	59	11	70
	Non-mastery	12	18	30
Total		71	29	100

a dial screen, which measures your weight mechanically with springs. After years of use, the scale will underestimate your weight because the springs become loose and less elastic, and it will give you the same result no matter how many times you step on the scale. This consistent result reflects reliability, but the reported weight is not accurate; it is invalid. Just like a broken weight scale, a language test might include items with incorrect keys or hire very harsh raters, and these conditions might produce reliable test scores, but the inferences we draw from the scores will not accurately reflect the language abilities of the test takers. Nonetheless, a valid test must have acceptable reliability, without which the results can never be meaningfully interpreted.

However, it is important to note that the picture we have presented of the relationship between reliability and validity is a simplified one. This section presents some critical views on the relationship between the two concepts. In contrast to the traditional view that reliability is a precursor to validity, there has also been a view seeing reliability as being in conflict with validity. Spolsky (1995: 5) identifies "two major ideologies" underlying testing: a traditional "humanist-scepticist descriptive approach," associated with the British style of testing and relying typically on essays, and a "rationalist-empiricist approach," associated with the United States and with multiple-choice tests. He finds both problematic:

> To oversimplify, with the traditional examination we think we know what we are assessing, but remain happily or unhappily uncertain about the accuracy or replicability of our assessment; with the modern examination, we are sure enough of our measurement, but are, or should be, uncertain as to what exactly we have measured.

To some extent, this view makes sense. In principle, items that are very similar to one another (in terms of target construct, format, and administration condition) will show strong internal consistency. Thus, if the goal is only to achieve high psychometric reliability, items included in a test need to be controlled to be as similar as possible to one another. Technically, this would be fine if we have a clear definition of the construct and know exactly how we measure it. However, language proficiency is a multi-faceted and fluid psychological construct. As the larger field of applied linguistics and TESOL has grown, the conceptualization of language proficiency has expanded to include more and varied subcomponents, each of which requires slightly different language tasks to measure. Despite the conceptual expansion, researchers and practitioners in the field do not have an agreed-upon operational definition of language proficiency. In practice, it is safer to include a variety of tasks that could cover different subcomponents of this construct. This is true whether the target construct is overall language proficiency (in which case one might include items that tap into reading, listening, speaking, or writing subskills) or specific language subskills (e.g., to measure speaking skills, one might evaluate the quality of speaking performances in a

number of criteria such as pronunciation, fluency, morphosyntax, and coherence). However, the variety of tasks and/or scoring criteria will likely lead to unique score variance unassociated with test takers' ability, different types of tasks might show weaker correlations than tasks of the same kind, and the unique score variances associated with task type will reduce the reliability of the test. On the other hand, if one prioritizes psychometric reliability and carefully controls the types of tasks used in a language test, he or she would run the risk of underrepresenting the construct as the test would weed out items that correlate less highly with the others. This, as Ennis (1999) points out, is potentially a serious threat to the validity of a test as it leads to a progressive narrowing of what is tested, without explicit consideration of how this narrowed construct impacts the content of the test. A classic way of narrowing the testing focus is to restrict the range of task types used and select items primarily based on psychometric quality – the discrete multiple-choice test format that Spolsky questioned.

That said, it is certainly no longer the case that language testing can be neatly classified into the reliable but invalid or the valid but unreliable. Moreover, the evolution of the concept of validity in language assessment has shifted focus from the purely psychometric quality of tests to the justification of score interpretation or test use (Messick, 1989; Kane, 1992, this volume). Current conceptualizations of validity (Chapelle *et al*., 2010; Chapelle and Lee, this volume) view reliability as an integral aspect of validity. Thus, in practice, the relationship between reliability and validity can be viewed more legitimately as an optimization problem. The aim should be to optimize both for meaningful score interpretation and use, rather than prioritize one over the other.

Moss (1994) pushed the question further by asking, "Can there be validity without reliability?" Her concern is about "the influence of measurement *practices* on the quality of education," arguing more recently that "conventional operationalizations of reliability in the measurement literature . . . unnecessarily privileged certain assessment practices over others" (Moss, 2004: 245, emphasis in original). As an alternative approach to psychometric reliability more suitable for large-scale testing, she endorses hermeneutics, which can cover a wide range of assessment practices, especially classroom-based assessments. She defines hermeneutics as

> a holistic and integrative approach to interpretation of human phenomena that seeks to understand the whole in light of its parts, repeatedly testing interpretations against the available evidence until each of the parts can be accounted for in a coherent interpretation of the whole.
>
> *(Moss, 1994: 7)*

In response to Moss's question, Mislevy (1994) defined reliability more broadly: "if by reliability we mean credibility of evidence, where credibility is defined as appropriate to the inference, the answer is no, we cannot have validity without reliability" (Mislevy, 1994: 11; see also Mislevy, 2004). However, he also acknowledges that sometimes in assessment practice, there is not a clear cut between what can be considered reliability or validity evidence, and this creates problems for estimating reliability in the psychometric sense. Instead, he suggests viewing estimations of reliability as collecting converging evidence to support score inferences under the same argument-based validity framework (Kane, 1992; this volume).

There is a similarly complex relationship between reliability and fairness. Fairness has been conceptualized in both broad and narrow senses (see Walters, this volume). In the

narrow sense, fairness is an integral part of test quality, which specifies that a fair test should have comparable psychometric quality across comparable test taker subgroups. In the broad sense, fairness is inseparable from validity and social justice (Xi, 2010; Yan *et al*., 2019). On social justice, McNamara and Ryan (2011) make a convincing argument from examples of language tests for immigration purposes that the narrow and broad senses of fairness might not necessarily coincide in language testing practices. Depending on the practical context of use, a fair test (in terms of psychometric quality) can lack justice; similarly, tests that are socially just might violate psychometric principles that are crucial to reliability. This relationship is somewhat similar to the relationship between reliability and validity. Situating the issue within the context of Rasch measurement models, McNamara *et al*. (2019) argue that information obtained from Rasch analysis – typically used to evaluate rater, item, and test reliability – can be conceptualized under the fairness framework, providing fairness evidence for language tests (see also Fan and Knoch, 2019). The evolving conceptualizations of reliability, validity, and fairness shifted the focus away from reliability as a central concern and, instead, brought these three concepts closer than ever before.

Future directions

The previous discussion indicates several respects in which current assessment practices are coming under pressure to change. We would like to suggest two possible directions for future research in the conceptualization and operationalization of reliability.

The first area of future research lies in rater reliability. As performance-based assessment has become the standard practice in both standardized and classroom-based assessment contexts, the importance of rater reliability cannot be overstressed. Human raters need rater training and repeated practice to internalize a rating scale and evaluate language performances more reliably. However, rater reliability cannot be achieved after only a one-time training session. Moreover, while some raters are able to achieve satisfactory reliability after a period of rater training, other raters might struggle with rating and/or rating scales despite repeated training. Thus, it is important to understand how raters learn to rate language performances and why certain rating behaviors persist. Two strands of research can help answer these questions: namely, longitudinal analysis of changes in rater behavior and fine-grained analysis of rater cognition. Thus far, few studies have investigated these questions, with the exception of two studies. Lim (2011) is a longitudinal study that tracked the performance of 11 certified raters for the MELAB test over three periods of 12 to 20 months. What he found was that rater reliability does not improve in a linear fashion. Instead, rater performance fluctuates over time. This fluctuation might last for a certain duration of time before rater performance settles down and becomes consistent and stable. Future research might explore possible causes of such fluctuations or developmental patterns of rater behavior. Another study that suggests the potential benefits of examining rater cognition and development is Winke and Lim (2015), which used eye-tracking technology to measure ESL essay raters' cognitive processes in applying the rating rubric developed by Jacobs *et al*. (1981; reproduced in Weigle, 2002). They found that raters' attention to analytic rubric subcomponents was associated with their inter-rater reliability estimates. Raters who had higher inter-rater reliability on different analytic criteria tend to have common attentional foci on those components on the scale; disagreeing raters did not show the same pattern of attention foci and instead read different parts of the rubric to justify their scores. The findings of this study imply that reliability must be viewed not simply as a technical quality of a test but as a developmental process that can connect language testing research with other related

disciplines (e.g., learning and cognitive sciences). More studies in this direction are needed to examine whether there are some universal patterns in rater reliability and development.

The second area of research we would like to recommend is for a better conceptualization and operationalization of reliability for classroom-based or alternative assessments. The reliability and dependability indices discussed in this chapter tend to be quantitative in nature, which could be challenging to implement in classroom contexts where language teachers are conducting the assessment. Testing researchers should seek complementary roles for quantitative and qualitative methods. While the quantitative psychometric information of language tests can be systematically evaluated and compared, it is challenging to have teachers, who typically do not have strong statistical skills, incorporate these approaches in classroom-based assessment (Shavelson, 2008). However, it is also important to develop awareness among teachers that formative assessment is assessment as well, and both its methods and its effect on teaching and learning require empirical validation (Bennett, 2010). Thus, formative assessment is the area where we see a clear need for language testing researchers to explore new forms of operationalizing test reliability. One approach that is worthy of exploring is to embed the examination of test reliability in the context of language assessment literacy. The importance of reliability and how reliability can be assessed differently in the classroom should be communicated to teachers in a way that is closely linked to their daily assessment contexts. For example, it is a common attitude or stance regarding assessment among teachers and teacher educators that teachers should have the freedom to hold assessment close to their own values and beliefs; they should not be standardized in terms of how they evaluate students' performance, as long as learning takes place among the students. However, standardization and freedom do not necessarily form a relationship of opposition. A teacher should be granted freedom in terms of instructional approach and materials selection. However, when multiple sections of the same class are taught by different instructors, when students need to be placed or advance into different levels of courses, or when the instructional program is required to achieve certain teaching or learning goals, standardization in assessment procedures becomes important for program evaluation purposes. In another example, we might frequently hear doubts about large-scale tests from teachers: for example, "I know this student well. His test score is not valid at all because he is a great student in my class." While this remark can be interpreted and rebutted in many different ways, what this phenomenon reflects is that teachers interact extensively with the students and can repeatedly observe student performance across contexts. Through repeated observation, teachers are able to make a more reliable assessment of students' ability. (This demonstrates the concept of reliability.) Both examples discussed here can potentially be used for assessment literacy training and serve as ways of ensuring and estimating reliability for classroom-based assessments. Language testing professionals should continue to promote engagement of language assessment in language teaching and learning.

Further reading

In this chapter we have provided a general overview of measurement, reliability, and dependability. While we introduced different approaches to estimating reliability and dependability, this chapter is mainly aimed at a conceptual understanding. Here are some suggestions for further reading on issues related to reliability and dependability. The suggested readings offer a combination of statistical reasoning behind reliability coefficients as well as its operationalizations in practical contexts.

Brown, J. D. (1989). Criterion-referenced test reliability. *University of Hawai'i Working Papers in ESL* 1: 79–113.
This article explains in detail the definitions and operationalizations of reliability in both norm-referenced and criterion-referenced tests, with a focus on the latter. It is a good introductory source on test reliability.

Dimova, S., Yan, X. and Ginther, A. (2020). *Local Language Testing: Design, Implementation, and Development*. New York, NY: Routledge.
This book does not focus on reliability per se. However, it provides a non-technical account of assessment principles, including item and rater reliability, through assessment activities in local contexts. This book is useful for language teachers, program coordinators, and novice language testing researchers who are interested in obtaining a general understanding of reliability in local assessment practices.

Krzanowski, W. J. and Woods, A. J. (1984). Statistical aspects of reliability in language testing. *Language Testing* 1: 1–20.
This article discusses the statistical properties of commonly used reliability estimates and explains how those estimates can be derived from ANOVA models.

McNamara, T., Knoch, U. and Fan, J. (2019). *Fairness, Justice and Language Assessment*. Oxford: Oxford University Press.
This book offers a practical account of how analyses of item, rater, and test reliability using Rasch measurement models can provide evidence for test fairness. Different from most articles or books on test reliability, this book integrates psychometric reliability in building fairness argument for language tests.

Myford, C. M. and Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement* 4: 386–422.

Myford, C. M. and Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement* 5: 189–227.
These two articles provide a conceptual explanation of rater effects and illustrate how these rater effects can be examined through many-facet Rasch measurement models. It is a useful source as an introduction to Rasch models and rater reliability.

References

AERA, APA, & NCME. (2014). *Standards for Educational and Psychological Testing*. Washtington, DC: AERA.

Alderson, J. C. (2010). Cognitive diagnosis and q-matrices in language assessment: A commentary. *Language Assessment Quarterly* 7: 96–103. https://doi.org/10.1080/15434300903426748.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

Bennett, R. (2010, February 18). *Formative Assessment: A Critical Review*. Cambridge, UK: Cambridge Assessment Network Seminar.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement* 38: 295–317. https://doi.org/10.1111/j.1745-3984.2001.tb01129.x.

Brown, J. D. (1989). Criterion-referenced test reliability. *University of Hawai'i Working Papers in English as a Second Language* 8: 79–113.

Brown, J. D. and Hudson, T. (2002). *Criterion-Referenced Language Testing*. Cambridge, UK: Cambridge University Press.

Chapelle, C. A., Chung, Y. R., Hegelheimer, V., Pendar, N. and Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing* 27: 443–69. https://doi.org/10.1177/0265532210367633.

Coe, R. J. (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch model. *Oxford Review of Education* 34: 609–36. https://doi.org/10.1080/03054980801970312.

Coleman, J. A. (2004). Modern languages in British universities: Past and present. *Arts and Humanities in Higher Education* 3: 147–62. https://doi.org/10.1177/1474022204042684.

Council of Europe. (2001). *Common European Framework of Reference for Languages, Teaching, Learning, Assessment*. Cambridge, UK: Cambridge University Press.

Ennis, R. H. (1999). Test reliability: A practical exemplification of ordinary language philosophy. In R. Curren (ed.), *Philosophy of Education*. Urbana, IL: The Philosophy of Education Society, 242–48.

Fan, J. and Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? *Papers in Language Testing and Assessment* 8: 117–42.

Feldt, L. S. and Brennan, R. L. (1989). Reliability. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: Macmillan, 105–46.

Ferris, T. L. J. (2004). A new definition of measurement. *Measurement* 36: 101–9. https://doi.org/10.1016/j.measurement.2004.03.001.

Frederiksen, N., Mislevy, R. J. and Bejar, I. (eds.). (1993). *Test Theory for a New Generation of Tests*. Hillsdale, NJ: Erlbaum.

Gulliksen, H. (1950). *Theory of Mental Tests*. New York, NY: Wiley.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V. and Hughey, J. (1981). *Testing ESL Composition: A Practical Approach*. Rowley, MA: Newbury House.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin* 12: 527–35. https://doi.org/10.1037/0033-2909.112.3.527.

Knoch, U. and Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing* 35: 477–99. https://doi.org/10.1177/0265532217710049.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing* 28: 543–60. https://doi.org/10.1177/0265532211406422.

McNamara, T. and Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing* 29: 555–76. https://doi.org/10.1177/0265532211430367.

McNamara, T., Knoch, U. and Fan, J. (2019). *Fairness, Justice and Language Assessment*. Oxford, UK: Oxford University Press.

McNamara, T. and Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly* 8: 161–78. https://doi.org/10.1080/15434303.2011.565438.

Messick, S. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: American Council on Education and Macmillan, 1–103.

Mislevy, R. J. (1994). *Can There be Reliability Without 'Reliability'?* Princeton, NJ: Educational Testing Service.

———. (2004). Can there be reliability without 'reliability'? *Journal of Educational and Behavioral Statistics* 29: 241–44.

Mislevy, R. J., Almond, R. G. and Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centred Design*. Research Report RR-03-16. Princeton, NJ: Educational Testing Service.

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher* 23: 5–12. https://doi.org/10.3102/0013189X023002005.

———. (2004). The meaning and consequences of 'reliability.' *Journal of Educational and Behavioral Statistics* 29: 245–49.

Shavelson, R. J. (2008). Guest editor's introduction. *Applied Measurement in Education* 21: 293–94. https://doi.org/10.1080/08957340802347613.

Spolsky, B. (1995). *Measured Words: Development of Objective Language Testing*. Oxford, UK: Oxford University Press.

Stellman, J. M. (1998). *Encyclopedia of Occupational Health and Safety*, 4th edn. Geneva, Switzerland: International Labor Office.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approached to estimating integrated reliability. *Practical Assessment, Research & Evaluation* 9: 1–11. https://doi.org/10.7275/96jp-xz07.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge, UK: Cambridge University Press.

Winke, P. and Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing* 25: 38–54. https://doi.org/10.1016/j.asw.2015.05.002.

Xi, X. (2010). Aspects of performance online graph description tasks: Influenced by graph familiarity and different task features. *Language Testing* 27: 73–100. https://doi.org/10.1177/0265532209346454.

Yan, X., Cheng, L. and Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examinee L1 background on scores of an ITA speaking test. *Language Testing* 36: 207–34. https://doi.org/10.1177/0265532218775764.

Yan, X. and Ginther, A. (2017). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang and A. Ginther (eds.), *Assessment in L2 Pronunciation*. Routledge, 67–88.

29

Scoring performance tests

Evelina D. Galaczi and Gad S. Lim

Introduction

Key considerations

Performance tests simulate real-world language use and typically involve test takers in the production of spoken or written texts, usually of some complexity and length. Such tests – also known as "direct" and "rater-mediated" tests – are scored by raters and are distinct from "indirect" tests, which principally focus on receptive knowledge and are usually scored without rater judgment through discrete items. A useful comparison has been drawn between the measurement of sports achievement and performance assessment: in sports, some disciplines can be readily measured – e.g., how many seconds an athlete takes to run 100 meters – while others involve the subjective judgment of expert judges, as in the evaluation of gymnastics routines (Pollitt, 1991; Spolsky, 1995). Performance tests clearly fall within the latter category. To extend this analogy further, the instruments used for the measurement of time or distance are interval scales with equal base units of seconds or millimeters; they are calibrated and by and large accepted as a global standard. Performance assessment does not enjoy the same level of psychometric precision, nor does it rest on universally accepted standards and scales. These inherent features of performance assessment make it one of the most challenging and problematic areas of (language) assessment, as it has to balance the need to elicit rich evidence of language ability and at the same time evaluate it systematically and consistently, in order to uphold the "foundational areas" of validity, reliability, and fairness of assessment (AERA *et al*., 2014).

The assessment *tasks* used naturally play a fundamental role in performance assessment. However, no matter how well crafted the tasks are, the language constructs tapped into are not truly realised without the *scoring criteria* used for their evaluation. Also referred to as "marking schemes" and "scoring rubrics," scoring criteria and associated descriptors are now widely seen as the operational definition of the test construct and an explicit description of the evidence sought to evaluate the elicited performance (Bejar, 2012; Fulcher, 2003; Mislevy *et al*., 2003; Weigle, 2002). Another key element of performance assessment are the *raters* who score the elicited performance. The key considerations and issues surrounding

DOI: 10.4324/9781003220756-38

scoring criteria and raters will be our primary focus in this chapter. With its "principles and *practice*" approach, the chapter is also intended to complement Fulcher's (2012) insightful "principles and *philosophy*" chapter in the first edition of this *Handbook*. Related discussions of L2 performance assessment tasks can be found in Knoch (this volume) and Nakatsuhara *et al*. (this volume) covering writing and speaking assessment.

Raters

Human judgment is intrinsic to scoring, making performance tests susceptible to conscious or unconscious rater idiosyncrasies, which may affect the psychometric quality of the rating. These "rater effects" (Myford and Wolfe, 2003) include raters who are inconsistent or who are inappropriately harsh or lenient, ultimately impacting test validity, reliability, and fairness. Raters could differ from each other in a number of ways – their decision-making styles (Baker, 2012); demographic backgrounds, experience, and training (Barkaoui, 2011; Lim, 2011; Şahan and Razi, 2020); and performance over time (Congdon and McQueen, 2005). Raters might also respond differently to the practical demands of their physical environment and the volume of ratings they need to get through – e.g., develop idiosyncratic "fast and frugal rules" (Bejar, 2012: 6) to get through more ratings faster. All these factors may cause "construct irrelevant variance," i.e., variations in marks unrelated to the test construct. Rater effects could be minimised through well-constructed tasks, scales, and criteria; through rater training, monitoring, and feedback on rating tendencies; and through multiple marking and/or statistical adjustment of marks.

Scoring criteria

Alongside raters, the evaluation instruments they use play an equally central role in performance assessment. The most widespread instruments for scoring performance tests are *rating scales*. They are typically composed of a number of hierarchical levels/bands for one or more *assessment criteria* (e.g., fluency, pronunciation), with *band descriptors* characterising the notable features of performances under those criteria across a number of levels of ability. Rating scales can be adapted to meet the needs of different groups: e.g., test constructors, examiners, or test takers (Alderson, 1991), through different lengths of scales, language of the band descriptions, and numbers of band descriptors. For example, some early self-assessment scales only defined the two endpoints, leaving the levels in between undefined, thus reducing the cognitive load for their lay users (Oscarson, 1980, 1989). A more recent example – the CEFR Self-Assessment Grid (Council of Europe, 2001) aimed at language learners – is a relatively short scale (six levels) and has simple language suitable for non-specialist users. In contrast, rater-oriented scales can include technical terms and more nuanced distinctions between different levels of ability, which raters are trained to understand and apply.

Rating scales can be *holistic* or *analytic* and, less frequently, *primary trait* or *multiple trait* (Hamp-Lyons, 1991). Holistic scales require raters to make one global judgment about all relevant aspects of the performance and assign a single score, whereas analytic scales identify a number of assessment criteria, each marked independently. Holistic scoring, it is argued, promotes more natural interpretation of the performance on the part of raters (Huot, 1993), lower cognitive demand on raters (Xi, 2007), and more efficient scoring and ease of reporting (Davies *et al*., 1999). In contrast, analytic scoring is seen as engendering a more systematic and focused evaluation process because it breaks down the construct of interest

into greater detail. Primary-trait/multiple-trait scales, seen as distinct from holistic/analytic scales, are more fine grained in nature and are particularly suitable for task-based assessment contexts, especially when building learner profiles (Fulcher, 2012; Hamp-Lyons, 1991, 2016). In the case of analytic and multiple-trait approaches to assessment, the collection of multiple marks is also seen as potentially enhancing reliability, since multiple observations lead to increased measurement precision (Barkaoui, 2011).

Assessment scales can also be task specific, sometimes referred to as *part scoring*, in which each test part is scored, as opposed to *full-test scoring*, in which an overall score/scores are given to performance across the whole test (Khabbazbashi and Galaczi, 2020). The part-scoring approach is well suited to assessing whether specific task-related features have been mastered. Part scoring can be holistic or analytic, although it may be too cognitively demanding for raters to assign a full range of analytic marks to each test part, especially in the real-time demands of speaking assessment (Taylor and Galaczi, 2011), or not possible to do so reliably in cases of tasks which elicit relatively limited performance samples.

While rating scales may be the most prominent approach to scoring performance tests, they are by no means the only option or, indeed, the best option in certain contexts. Alternative approaches to scoring have been developed, such as empirically derived, binary-choice, boundary definition scales (EBBs), performance decision trees (PDTs), and checklists.

EBBs consist of a set of binary questions which are hierarchically arranged. Raters make a series of yes/no choices to arrive at a final score, as exemplified by Upshur and Turner (1995) and Hirai and Koizumi (2013). It is argued that EBBs are better than rating scales at identifying the cut point between acceptable and unacceptable performances. On the other hand, the questions developed for EBBs tend to be task specific and not easily transferrable to other tasks, which limits their generalisability.

PDTs are based on a set of guiding questions about performance features, which are based on learner/test taker performances (Fulcher *et al*., 2011). For example, the PDT evaluation of interactional competence in service encounters (as illustrated by Fulcher *et al*., 2011) would be based on analysis of relevant interactions and would consist of a set of decisions about aspects of interactional competence raters need to make: e.g., the use of transition boundary markers or backchannelling, with each aspect accruing points. The fundamental dual aim of PDTs is to capture the breadth and depth of actual performance as seen in contexts of language use – "thick description," in Fulcher's (1996) terms – and to simplify the process for raters through focused step-by-step guidance.

Lastly, checklists involve specific criteria broken down into can-do statements, which can be evaluated on a binary or non-binary basis. The value of using checklists is that they go further than rating scales in terms of explicitly defining fine-grained criteria and achievement targets. They are, therefore, well suited to providing feedback on performance, and since they are most often "local" in their application, as opposed to the "global" context of rating scales in national/international tests, checklists can better "align the assessment to the realities of the classroom" (Harsch and Seyferth, 2020: 12).

Historical perspectives

The history of L2 performance assessment can be traced through evolving conceptualisations of test constructs and understanding of raters and scales. We will only briefly touch on developments in test constructs and instead will focus on the evolution of psychometric awareness of issues around raters and scales, which ultimately led to the development of

more sophisticated and nuanced conceptualisations of validity arguments in performance assessment.

Bejar (2012) traces one of the initial developments in performance assessment to Fechner's (1897, cited in Bejar, 2012) work on evaluating traits, which in his case involved standardizing judgments of aesthetics in art. That same decade, Edgeworth, a statistician, investigated the influence of chance in the British civil service examination scoring. He pessimistically concluded:

> I find the element of chance in these public examinations to be such that only a fraction – from a third to two-thirds – of the successful candidates can be regarded as safe, above the danger of coming out unsuccessfully if a different set of equally competent judges had happened to be appointed.
>
> *(Edgeworth, 1890: 653)*

A few decades later, a key contribution to the understanding of performance measurement came through the work of Thorndike in the USA. Working in the domain of writing assessment, Thorndike set out to create an objective scale for marking essays. He recommended using a panel of judges to agree on which essay samples constitute a specific level, in order to arrive at "one point on the scale thereby defined" (Thorndike, 1912: 292). Scholars in the UK were skeptical about the use of Thorndike's scale, as it was seen as too contextually dependent on schools in the USA: "There is, in fact, no help for it," wrote a British academic. "We must, for the present at least, fall back upon the method of personal impression" (Ballard, 1923: 216, quoted in Fulcher, 2012). These promising developments at the beginning of the twentieth century took place outside language assessment. In L2 assessment, rater judgment was still the only method of assigning scores on performance tests.

Alongside these broad developments in educational measurement at the start of the twentieth century, L2 assessment was additionally under the influence of pedagogic movements – the direct method (with its focus on oral fluency) and the reform movement (with its focus on phonetics and pronunciation) – as well as socio-political forces such as two world wars, the expansion of global travel, and the internationalisation of education, which resulted in growing awareness of the need for "foreign languages for global war and global peace" (Kaulfer, 1944: 150). This in turn led to more widespread use of performance tests, especially speaking tests, and the foregrounding of relevant psychometric issues (Fulcher, 2003; Spolsky, 1995; Weir *et al*., 2013). This is the context in which the first high-stakes L2 test to contain a speaking component alongside a writing component was administered in 1913 – the Certificate of Proficiency in English, offered by the University of Cambridge Local Examinations Syndicate, UK. The speaking component was an hour long and consisted of a dictation, reading-aloud task, and – crucially for performance assessment – a conversation with an examiner. In this pioneering performance test, full trust was placed on individual rater judgments, and no assessment scales or criteria were explicitly defined to guide raters to a common standard (Fulcher, 2003; Weir *et al*., 2013). Other speaking tests later followed, as seen in the Lower Certificate in English exam released in 1939 in the UK, the US Army Specialized Training Program launched in 1943, and the US Foreign Service Institute (FSI) proficiency test introduced in the 1950s (see Fulcher, 2003; Spolsky, 1995 for more detailed discussions).

The first concerns with the unique challenges posed by L2 performance assessment are seen in the work of Roach in the 1940s in the UK in the domain of L2 speaking. Roach focused on the positive impact of joint examining by two examiners on keeping standards

comparable across individuals and on the need for more precise definitions of standards to guide raters (Roach, 1936, 1945). In highlighting issues around norm referencing rather than examining to an external standard, he cautioned that, in the absence of an external criterion, candidates would set the standard, since each individual's score would be affected by the ability of the others in the group. Standards and associated scores, therefore, would be relative to the ability of candidates in each group and not based on a shared universal criterion: "[T]here is no guarantee that standards acquired by Examiner X through experience in South America will correspond with those learnt by Examiner Y in the Middle East" (Roach, 1945: 8).

A decade later, similar test reliability challenges were faced in the United States during the development and administration of the FSI test, such as the use of inter-rater, intra-rater, and test-retest reliabilities and the relationship of speaking test scores to other measures of language competence (Clark, 1978). The FSI test is further important for our discussion because it was a significant first attempt in L2 testing to define a criterion external to the test candidature itself. It had a six-point holistic scale which included descriptors defining distinct levels of proficiency and evolved to include a list of five criteria (accent, comprehension, fluency, grammar, and vocabulary) as a check on the holistic score – an important first step toward using analytic scales in speaking assessment (Fulcher, 2003).

This brief and, by default, selective historical narrative highlights the evolution in thinking about psychometric concerns in performance assessment. These conceptual and practical developments in thinking paved the way for current conceptualisations of score inference and interpretation and for frameworks of language testing and associated validity debates. Two notable examples are Weir's (2005) socio-cognitive framework for test validation, in which "scoring validity" is posited as part of the validity of a test, and argument-based validity arguments, as seen in the "scoring" inference in Kane's (2006) work. These influential frameworks have direct relevance for performance assessment and are a culmination of over a century of developing thought in the area of performance tasks, scales, raters, and overall views of validity.

Critical issues and topics

"Strong" and "weak" performance tests

In performance tests, test takers perform communicative tasks similar to those in the real world. Thus, it is possible to evaluate test takers according to successful completion of these tasks: what McNamara (1996) calls "strong" performance assessments. However, most language tests with a performance component are "weak" performance tests, evaluating test takers primarily on linguistic criteria and not on actual real-life task completion. This fails to take full advantage of the unique evidence that performance tests bring. As Fulcher (2012) notes, "strong" performance assessments can answer important questions, such as "Is this person able to guide pilots through air space safely and get them onto the ground without incident? Is that person capable of explaining a medical condition to a patient, discussing and selecting alternative treatments, and writing a case history?" (p. 387).

In recent years, there has been work to expand the use of task-based assessment (Norris *et al*., 2002) as well as the identification of "indigenous criteria" (Jacoby, 1998) in specific-purposes testing (e.g., Elder *et al*., 2013), and these developments have helped move performance assessment toward the strong end of the continuum. One such example comes from the healthcare sector, where the Occupational English Test (OET) speaking component

simulates a consultation between a health professional and a patient. Alongside traditional linguistic criteria, the assessment criteria for the test include clinical communication criteria such as "relationship building," "understanding and incorporation the patient's perspective," and "information gathering."

Choosing contextually appropriate scoring instruments

Discussions about the scoring of performance tests often begin and end with rating scales. As we noted earlier, rating scales meet the requirements of particular assessment contexts, whereas other assessment contexts may be better served by different scoring approaches. Rating scales are useful for large-scale, summative assessment contexts, where reducing performance to one or a few numbers or indicators is desirable. Where assessment is primarily meant to inform teaching and learning, on the other hand, having just one or a few numbers is less helpful, and relative terms that are often used in rating scales such as "good" or "somewhat accurate" are quite vague, making them difficult to work with (Barnwell, 1986; Matthews, 1990; Wilds, 1979). In such contexts, checklists with items that cover the different aspects of interest, as well as clear yes-no judgments about attainment of them, can be more useful. Conversely, checklists, EBBs, and PDTs – with their criteria presented discretely – may not be the best for high-stakes testing contexts, as it would be very easy for test takers to game the test by exclusively providing the specific features identified in the criteria.

On the really large-scale end of the spectrum, the European Survey on Language Competences aimed to identify comparable performances across several different languages. While the research team initially started with an analytic rating scale for the writing test, it quickly became clear that simply translating scale descriptors into different languages would not result in equivalent instruments or outcomes (Jones, 2009). In the end, a comparative judgment approach was adopted (Bramley, 2005), in which writing performances were evaluated against exemplar scripts at a range of levels so as to produce a rank ordering (European Commission, 2012). In this case, sample performances became the "criteria."

Thus, it is important to carefully consider the context in which the performance assessments are being used, as the selection of an appropriate scoring instrument is crucial to producing meaningful and useful outcomes.

Advantages and limitations of scoring approaches

Holistic and analytic approaches to rating performance scales are not without their critics. Hamp-Lyons (2016) criticises holistic scoring, likening it to building a house with one brick. White (1984) puts forward a contrasting view, portraying holistic scales as an antidote to "analytic reductionism" (p. 400). Such differing viewpoints are not surprising. To reiterate one of our earlier points, one scoring approach is not inherently superior to another. Instead, it's more critical to ensure that the scoring approach used is well fitted to the purpose and context of the test.

The choice of scoring approach also reflects the philosophy of what is considered important in test performance – e.g., part scoring, as argued by Khabbazbashi and Galaczi (2020), attributes more importance to variation in performance across tasks and aligns more closely to task-centred approaches to construct definition (Brown *et al*., 2002; Norris *et al*., 2002). Analytic scoring, in contrast, focuses on performance against different linguistic criteria and has closer affinity with trait-based construct definitions (Chapelle, 1998).

Developing assessment criteria

The methodologies for developing assessment criteria are sometimes described as employing either an intuitive or an empirical approach. The former refers to using expert judgment, whereas the latter is data driven. This terminology implies that one of these approaches, the empirical one, is better and therefore really the only way to construct scales. It is perhaps more helpful to describe the development of scoring criteria according to their starting point. Typically, developers either start by developing descriptors based on collective expert knowledge and available body of research or start by looking at performance samples (Council of Europe, 2001), subsequent to which a range of qualitative and quantitative methods can and should be undertaken to validate and provide empirical backing for decisions. In practice, these different approaches are often used together in an iterative process of development and validation.

Where descriptors (or checklist items) are the starting point, a key advantage is that developers can ensure that all important aspects of the construct are appropriately accounted for in the scoring criteria, some of which can be missed if starting with samples which may not have a range of salient performance aspects. A key caveat, on the other hand, is that starting with descriptors means that the validation claim of the criteria rests heavily on the experience and expertise of the developers. This may be sufficient in certain assessment contexts but may require much more empirical backing in other contexts. In the case of the illustrative descriptors in the Common European Framework of Reference, which were drawn from existing rating scales, the validation included the use of Rasch analysis to show that a large group of teachers can consistently rank order them, allowing the descriptors to be placed at appropriate levels (North, 2000). However, the use of human judgment to validate human judgment has been criticised as being circular (Fulcher, 2009; Hulstijn, 2007).

Performance samples are an alternative starting point for developing scoring criteria, and this is true for rating scales as well as for other approaches such as EBBs and PDTs. With this approach, there is a link between performances and criteria from the beginning, which can make it easier to establish validation claims (Fulcher *et al*., 2011). Still, what is observed of the performance samples depends on who is doing the observing, so the weakness of subjectivity ascribed to the other approach is present in this approach too. Even with automated scoring systems, in which large numbers of performances are processed by machine in order to detect the underlying criteria, the typical starting point for the machine algorithms remains the scores assigned to those performances by human raters. Human subjectivity, therefore, is necessarily a part of the test development process, and while it is something to be aware of and account for, it does not necessarily detract from and can potentially contribute to the validity of test instruments and outcomes (Moss, 1994).

Limitations of rating scales

It is important to recognise that rating scales take the form that they do in part to account for human rater attentional limitations, as well as other contextual limitations (e.g., amount of money and time available for the rating). Even though the possible evaluation criteria are myriad (Broad, 2003), many analytic scales tend to have the same number of analytic criteria (Lim, 2012), and the number of score points for those criteria seem to have "the magical number seven, plus or minus two" (Miller, 1956: 81), roughly corresponding with the number of pieces of information we can hold in short-term memory.

Some design features of scales also seem to be more the result of historical inertia than empirical insights. For example, analytic scales are often displayed as perfect symmetrical grids, in which each assessment criterion has the same number of performance levels/bands and associated band descriptors. And yet there is no necessary reason for the different criteria in an analytic scale to each have the same number of score points. Indeed, the evidence shows that raters can distinguish more levels of some criteria and fewer levels of others (e.g., Isaacs *et al*., 2015). Scales with the same number of score points across criteria, apart from being less than sound from a measurement perspective, can also falsely convey notions of natural ability levels or that various aspects of that ability develop in lockstep. Indeed, there is limited evidence that progression on many rating scales bears any relation to progression in second language acquisition (Brindley, 1998; Fulcher, 2008). Users may be better served by scales which have varying numbers of score points for different criteria as appropriate.

The limited number of criteria and the short length of descriptors also goes against the need for adequate construct coverage, so rating scales are necessarily a compromise between competing requirements which cannot be jointly maximised. The same issue also applies to checklists, EBBs, and PDTs to some extent.

Role of rater effects

In our discussion so far, rater effects have been portrayed as undesirable and to be minimised. This perspective is squarely based in a measurement paradigm of test validity and is not the only way to conceptualise the role of raters' individual differences. Arguments that put the validity of the individual rater's perspective centre stage (Brown, 2012; Erdosy, 2004; Moss, 1994) challenge this psychometric perspective. The divergent opinions of individual raters, these arguments go, need to be seen as part of validity arguments and not socialised out of raters in service of high psychometric indices. Such a rater-primary conceptual approach is well suited to classroom-based environments in which teachers'/examiners' individual and potentially divergent views would be a valuable element in the development of learner profiles. At the same time, it is important not to lose sight of the implications of this approach in standardised performance assessment. While it would be irresponsible not to minimise rater effects through appropriate mechanisms in large-scale performance assessment, it would be equally undesirable to swing to the opposite extreme of engendering "rating hyper-homogeneity" (Ross, 2012: 229) or "inculcating mindless conformity" (Linacre, 1989: iii).

The role of rater effects is, therefore, not a clear-cut conceptual point but is part of a tension between psychometric convergence and individual divergence. Indeed, it is worth keeping in mind that even machine ratings of performance – though often portrayed as bias-free and the ultimate standard of objectivity – are not immune to the influence of rater effects. In most cases, it is human judgments that provide the basis for the development and evaluation of the algorithms that ultimately provide the scores (Bejar, 2012; Caliskan *et al*., 2017; Yapo and Weiss, 2018).

Score interpretation

Whichever type of scoring instrument one uses, whether involving human raters or computer algorithms, there is the task of interpreting what these observations and scores mean in terms of the test constructs of interest. As we argued earlier, strong performance assessments may help with this somewhat by reducing the amount of inference and interpretation

needed, but there remains the question of how generalisable those results are. If one or more facets of the task were to be changed, would the outcome still be the same for the test taker? (See Jin, this volume, for a discussion on test specifications and mechanisms for keeping inferences stable across test forms.) In the end, in the practice of (performance) assessment, there is no escaping from the need to make inferences or from the need to build arguments in support of a particular interpretation of assessment outcomes.

Current contributions and research

This section will focus on three areas that have seen a growing body of research in the past decade: the use of research designs based on multiple sources of evidence in developing assessment scales, more in-depth understanding of rater cognitive processes, and developments in automated evaluation of performance assessments.

Mixed-methods scale development

Recognition of the need for empirically driven assessment scales is not new – over three decades ago, Fulcher (1987) started advocating for this approach, which was later included in the influential CEFR (2001) descriptor development guidance. What is new is the widespread recognition that such an approach is now the norm. Published research in this area is gaining ground, with more frequently available accounts of the process, challenges, limitations, and affordances encountered in scale development. One example is a study reported by Banerjee *et al*. (2015), which focuses on the review and revision of analytic scales for a large-scale English language writing test. A range of data sources were used to inform decisions: a literature-informed taxonomy of text features which distinguish different proficiency levels, corpus-based evidence on those text features which best predict writing proficiency levels, and rater perceptions about scale components. Banerjee *et al*.'s work was guided by theoretical models (through the available body of research), prediction (through corpus evidence), and usability (through rater perceptions). Based on such broad evidence, the research team could suggest features to be included in the band descriptors and identify the personal constructs raters employed which helped or hindered the reliability of their judgments. Such findings both informed the assessment categories and band descriptors in the scale and provided targeted areas for rater training. In addition to providing an example of the broad body of evidence gathered, this study also highlighted a range of considerations for bridging research and practice in the development and implementation of scales, such as the number of levels to include in the scale, how to define levels and keep distinctions between adjacent levels clear, and the difficulties of the "perfect grid" approach to scale design. Other examples of scale development based on multiple data sources can be found in Harsch and Seyferth's (2020) account of the development of CEFR-based checklists for assessing writing proficiency, Galaczi and Khabbazbashi's (2016) overview of the multistage process of developing analytic scales for a speaking test, and Hirai and Koizumi's (2013) description of the development of EBBs for a storytelling classroom assessment.

Raters' cognition and decision-making processes

An area of growing interest in performance assessment is that of rater cognition and decision making. It focuses on understanding what goes on in the minds of raters as they rate and, it is argued, is central to validity arguments underlying performance tests (Bejar, 2012).

Lumley's (2005) model of the rating process was an influential attempt to highlight the complexity of rater decision-making in writing assessment. His model – essentially a matrix – includes three stages of rating (reading, scoring, and conclusion) with transversal dimensions cutting through – instrumental (the visible part of the process), interpretation (the cognitive dimension of the process), and institutional (the purpose for measuring the performance) levels. Lumley's model is valuable in that it highlights the inter-related nature of the social and cognitive dimensions of rating.

A different but related line of research is offered by Baker (2012), who investigated decision-making rater styles such as rational, intuitive, dependent, avoidant, and spontaneous. The author speculated that raters' preferred decision-making style may influence their rating behavior because certain styles may have an affinity with certain types of scales or features of scales. Barkaoui (2011) argued along similar lines: namely, that the type of rating scale might impact raters' decision-making behaviour and, more recently, Fairbairn and Dunlea (2017) speculated that "rational, strategic, and reflective raters would work better with analytic scales and intuitive, non-strategic, impulsive, and spontaneous raters would prefer holistic scales" (p. 10). Similarly, Zhang (2016) linked raters' cognitive and metacognitive strategies with rating accuracy and found that there was a relationship between them: those raters who were better at integrating information from essays were more accurate in their scoring.

A study from music assessment provides interesting findings as well regarding the use of machine learning in the automated classification of (music) raters in a number of rater types. In this respect, Wesolowski (2019) proposed a method combining Rasch measurement to estimate rater measures and a "random forests" supervised machine-learning algorithm to classify raters into distinctive groups. The result was the grouping of raters into three distinct groups with 95 percent accuracy: syntactical, expressive, and mental representation rater types. Such innovative AI-driven methods could provide focused findings of how raters interact with scoring instruments and improve our understanding of rater differences.

Automated scoring

Automated scoring of performance tests (see Xi, this volume) is now firmly part of assessment practice, especially in large-scale, standardized testing. These engines are typically trained on large samples of texts that have been previously scored. Not constrained by human attentional limitations, these technologies can identify a large number of features or combinations of features (often numbering into the thousands) that distinguish between different performances, all of which can be appropriately weighted and factored into the final score. They also can rank order performances into as many levels as there are performances and can provide relevant feedback if desired (Attali and Burstein, 2006; Burstein and Chodorow, 1999; Yannakoudakis *et al*., 2011, 2012). These engines have now gone beyond looking at surface features to consider more complex aspects of constructs, such as coherence (e.g., Wang *et al*., 2017), but are generally still limited in their ability to deal with extended spontaneous speech and interactive speech.

Recommendations for practice

Through a practical lens, it is useful to approach the complex process of performance assessment via a design/develop/operationalize cycle. Here we provide a number of suggestions following this cycle. They are not intended as a checklist to follow, but rather as a starting

point highlighting key questions for consideration by both developers and users of performance assessments.

Design stage

- *Fit-for-purpose scoring instruments*: The most fundamental initial consideration is to consider what scoring instruments will be most suitable, given the purpose and context. For example, is an overall score needed, a profile with a number of scores, fine-grained task-based evaluations, or diagnostic recommendations?
- *Scoring criteria*: What scoring criteria will be most appropriate? Is there theoretical justification for them? Are they scalable (i.e., capable of differentiating levels of performance quality)?
- *Scoring model*: What scoring model will ensure suitable reliability? Are multiple marking, statistical adjustments of scores, and/or ongoing training/monitoring possibilities?
- *Practical constraints*: What constraints need to be taken into account, in terms of numbers of raters needed, rater training, and frequency/length of assessment?

Development stage

- *Evidence for the design decisions*: Multiple sources of quantitative and qualitative evidence need to be gathered (e.g., theoretical insights from relevant literature, language data from performances, rater perceptions, rater scoring behavior, rater cognitive processes, test taker perceptions, and other key stakeholder views).
- *Multi-stage iterative development*: The development of assessment instruments should go through at least two or three stages: a prototyping stage, a pilot stage, and a field trial stage (see Kremmel *et al*., this volume). Even in low-stakes settings, a prototyping or pilot stage is indispensable in order to test out assumptions about the design principles and data gathering. More complex-scale development projects would necessitate several development stages.

Operational stage

- *Robust implementation*: This is a critical stage for successful performance assessment since it underlies the transition from research into practice. Key considerations centre on the examiners/teachers who will be delivering the assessment: e.g., what kind of training will they need to be able to engage in reliable assessments? What monitoring will need to be in place to evaluate the success of the training? (see Davis, this volume).
- *Ongoing validation and monitoring*: The reliability of performance assessment is not a one-off judgment, but an ongoing process involving continuous gathering of evidence to support the validity, reliability, and fairness of the assessment. Potential areas to focus on are rater reliability indices, multi-facet Rasch measurement to identify individual examiner effects, examiner/teacher observations, and consultations to gather additional evidence about scoring practices.

Conclusion and future directions

We end with the most essential take-aways and a few projections for the future of performance assessment. First, it should be evident that no instrument for scoring performance

is inherently good or bad. Test purpose, use, and context should shape the type of scoring approach used. Second, raters – with their universe of human variables affecting the measurement – are an integral part of assessment, and it is a "fact of life" (McNamara, 1996) that there will be variability within and between them because they are human beings. That variability needs to be minimised, but rater differences also need to be seen as part of validity arguments supporting tests. Third, assessment scales and raters have by necessity been presented here as discrete, but in reality, they are interconnecting elements in the complex ecosystem of assessing performance.

Technology stands to bring innovations in the domain of rater cognition and to supplement current research on examiner perceptions, which could be susceptible to self-report bias. Data from eye-tracking and electro-encephalography (EEG) technologies has the potential to make the subconscious process of raters visible when reading scripts and listening to speech. Eye tracking focuses on a person's eye movements, whereas EEG measures the electrical activity in a person's brain as a result of listening or reading. The latter is a more advanced tool in capturing real-time cognitive processes: unlike in eye tracking, "where we can only see that something *is* more difficult to process, EEG allows us also to see *why* it is, i.e. because of syntax, semantics, prosody, etc." (Schmidt and Pastorino, forthcoming).

As for automated scoring, there are still limitations, mainly in terms of the narrower constructs that scoring engines can currently deal with. Hybrid machine/human scoring in various configurations, e.g., double marking or complementary marking, in which the machine and human focus are on different tasks and/or assessment criteria, is one solution for the meantime. However, the technology is developing very quickly, and the use of deep learning techniques means that these engines will be able to handle more complex performances, as well as get at underlying features and aspects of performances that human raters cannot, and then the challenge will be for humans to validate score meanings (see Xi, this volume).

The widespread use of technology and the ubiquitous availability of devices and data could also reshape the single-test-event generalizability dilemma in performance assessment. We see a future vision of assessment where personal smart devices collect, record, and store instances of L2 performance and create a learner/test taker profile (their personal corpus), which includes instances of "weak" and "strong" performances, becomes richer over time and provides AI- and human-informed information about language competence.

Further reading

The four suggestions here present recent empirical investigations into a range of performance assessment issues and would be useful extensions of this chapter.

Crusan, D. (2015). The use of rubrics to assess writing: Issues and challenges. *Assessing Writing* 26: 1–82.
A special issue focusing on the development and revision of assessment scales for a range of contexts.

East, M. and Cushing, S. (2016). Innovation in rubric use: Exploring different dimensions. *Assessing Writing* 30: 1–76.
A special issue offering discussions on different contexts of use of assessment scales.

Kuiken, F. and Vedder, I. (2014). Assessing oral and written L2 performance: Raters' decisions, rating procedures and rating scales. *Language Testing* 31: 285–392.
A special issue exploring rater-focused questions.

Wind, S. A. and Engelhard, G. (2019). Rater-mediated assessments. *Journal of Educational Measurement* 56: 473–663.
A special issue focusing on raters from fields beyond L2 assessment.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson and B. North (eds.), *Language Testing in the 1990s*. London: Modern English Publications/British Council, 71–86.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning and Assessment* 4: 1–31. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly* 9: 225–48. https://doi.org/10.1080/15434303.2011.637262.

Ballard, P. B. (1923). *Mental Tests*. London, England: Hodder and Stoughton.

Banerjee, J., Yan, X., Chapman, M. and Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing* 26: 5–19. https://doi.org/10.1016/j.asw.2015.07.001.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice* 18: 279–93. https://doi.org/10.1080/0969594X.2010.526585.

Barnwell, D. (1986). *A History of Foreign Language Testing in the United States*. Tempe, AZ: Bilingual Press.

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice* 31: 2–9. https://doi.org/10.1111/j.1745-3992.2012.00238.x.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement* 6: 202–23.

Brindley, G. (1998). Describing language development? Rating scales and SLA. In L. F. Bachman and A. D. Cohen (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research*. New York, NY: Cambridge University Press, 112–40.

Broad, B. (2003). *What We Really Value: Beyond Rubrics in Teaching and Assessing Writing*. Logan, UT: Utah State University Press.

Brown, A. (2012). Interlocutor and rater training. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. London: Routledge, 413–25.

Brown, J. D., Hudson, T. D., Norris, J. M. and Bonk, W. (2002). *An Investigation of Second Language Task-Based Performance Assessments*. Honolulu, HI: University of Hawai'i Press.

Burstein, J. and Chodorow, M. (1999). *Automated Essay Scoring for Nonnative English Speakers*. Retrieved from www.aclweb.org/anthology/W99-0411.pdf.

Caliskan, A., Bryson, J. J. and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–86. https://doi.org/10.1126/science.aal4230.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman and A. D. Cohen (eds.), *Interfaces Between Second Language Acquisition and Language Testing Research*. New York, NY: Cambridge University Press, 32–70.

Clark, J. L. D. (1978). Interview testing research at educational testing service. In J. L. D. Clark (ed.), *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service, 211–29.

Congdon, P. J. and McQueen, J. (2005). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement* 37: 163–78. https://doi.org/10.1111/j.1745-3984.2000.tb01081.x.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of Language Testing*. Studies in Language Testing volume 7. Cambridge: UCLES/Cambridge University Press.

Edgeworth, F. Y. (1890). The element of chance in competitive examinations. *Journal of the Royal Statistical Society* 53: 644–63. Retrieved from www.jstor.org/stable/2979547.

Elder, C., McNamara, T., Woodward-Kron, R., Manias, E., McColl, G., Webb, G. and Pill, J. (2013). *Toward Improved Healthcare Communication: Development and Validation of Language Proficiency Standards for Non-Native English Speaking Health Professionals*. Melbourne: University of Melbourne.

Erdosy, M. U. (2004). Exploring validity in judging writing ability in a second language: A study of four experienced raters of ESL compositions. In *TOEFL Research Report No TOEFL-RR-70*. Princeton, NJ: Educational Testing Service.

European Commission. (2012). *First European Survey on Language Competences: Technical Report*. Luxembourg: Publications Office of the European Union.

Fairbairn, J. and Dunlea, J. (2017). *Speaking and Writing Rating Scales Revision Technical Report*. London: British Council Assessment Research Group.

Fechner, G. T. (1897). *Kollektivmasslehre*. Leipzig, Germany: Wilhelm Engelmann.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *English Language Teaching Journal* 41: 287–91. https://doi.org/10.1093/elt/41.4.287.

———. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13: 208–38. https://doi.org/10.1177/0265532209359514.

———. (2003). *Testing Second Language Speaking*. London: Longman Pearson Education.

———. (2008). Criteria for evaluating language quality. In E. Shohamy and N. H. Hornberger (eds.), *Encyclopedia of Language and Education*, *Language Testing and Assessment*, 2nd edn., vol. 7. New York, NY: Springer, 157–76.

———. (2009). Test use of political philosophy. *Annual Review of Applied Linguistics* 29: 3–20. https://doi.org/10.0.3.249/S0267190509090023.

———. (2012). Scoring performance tests. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. Abingdon and New York, NY: Routledge, 378–92.

Fulcher, G., Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing* 28: 5–29. https://doi.org/10.1177/0265532209359514.

Galaczi, E. D. and Khabbazbashi, N. (2016). Rating scale development: A multi-stage exploratory sequential design. In J. Creswell, A. Moeller and N. Saville (eds.), *Second Language Assessment and Mixed Methods Research. Studies in Language Testing*, vol. 43. Cambridge: UCLES/Cambridge University Press, 208–32.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (ed.), *Assessing Second Language Writing in Academic Contexts*. Norwood: Ablex, 241–76.

———. (2016). Farewell to holistic scoring? *Assessing Writing* 27: A1–A2. https://doi.org/10.1016/j.asw.2015.12.002.

Harsch, C. and Seyferth, S. (2020). Marrying achievement with proficiency – Developing and validating a local CEFR-based writing checklist. *Assessing Writing* 43: 1–15. https://doi.org/10.1016/j.asw.2019.100433.

Hirai, A. and Koizumi, H. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Language Assessment Quarterly* 10: 398–422. https://doi.org/10.1080/15434303.2013.824973.

Hulstijn, J. A. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *Modern Language Journal* 91: 663–67. https://doi.org/10.0.4.87/j.1540-4781.2007.00627_5.x.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson and B. A. Huot (eds.), *Validating Holistic Scoring for Writing Assessment*. Cresskill, NJ: Hampton Press, 206–32.

Isaacs, R., Trofimovich, P., Yu, G. and Munoz Chereau, B. (2015). Examining the linguistic aspects of speech that most efficiently discriminate between upper levels of the revised IELTS pronunciation scale. *IELTS Research Reports*, *2015–4*.

Jacoby, S. (1998). *Science as Performance: Socializing Scientific Discourse Through the Conference Talk Rehearsal*. Unpublished doctoral dissertation. University of California, Los Angeles.

Jones, N. (2009). A comparative approach to constructing a multilingual proficiency framework: Constraining the role of standard-setting. In N. Figueras and J. Noijons (eds.), *Linking to the CEFR Levels: Research Perspectives*. Arnhem: CITO/EALTA, 35–44.

Kane, M. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: American Council on Education, Praeger Publishers, 17–64.

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal* 28: 136–50. https://doi.org/10.1111/j.1540-4781.1944.tb04835.x.

Khabbazbashi, N. and Galaczi, E. D. (2020). A comparison of holistic, analytic, and part marking models in speaking assessment. *Language Testing*, 1–28. https://doi.org/10.1177/0265532219898635.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing* 28: 543–60. https://doi.org/10.1177/0265532211406422.

———. (2012). Developing and validating a mark scheme for writing. *Research Notes* 49: 6–10.

Linacre, J. M. (1989). *Many Facet Rasch Measurement*. Chicago, IL: MESA Press.

Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective*. Frankfurt: Peter Lang.

Matthews, M. (1990). The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations. *English Language Teaching Journal* 44: 117–21. https://doi.org/10.1093/elt/44.2.117.

McNamara, T. F. (1996). *Measuring Second Language Performance*. London: Longman.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81–97. https://doi.org/10.1037/h0043158.

Mislevy, R. J., Almond, R. G. and Lukas, J. F. (2003). *A Brief Introduction to Evidence-Centered Design. Research Report RR-03–16*. Princeton, NJ: Educational Testing Service.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher* 23: 5–12. https://doi.org/10.3102/0013189X023002005.

Myford, C. M. and Wolfe, A. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement* 4: 386–422.

Norris, J. M., Brown, J. D., Hudson, T. and Bonk, J. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing* 19: 395–418. https://doi.org/10.1191/0265532202lt237oa.

North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. Frankfurt: Peter Lang.

Oscarson, M. (1980). *Approaches to Self-Assessment in Foreign Language Learning*. Oxford: Pergamon.

———. (1989). Self-Assessment of Language Proficiency: Rationale and Applications. *Language Testing* 6: 1–13. https://doi.org/10.1177/026553228900600103.

Pollitt, A. (1991). Giving students a sporting chance: Assessment by counting and judging. In C. Alderson and B. North (eds.), *Language Testing in the 1990s*. London: Macmillan, 46–59.

Roach, J. O. (1936). The reliability of school certificate results. *Overseas Education: A Journal of Educational Experiment and Research in Tropical and Subtropical Areas* 7: 113–18.

———. (1945). *Some Problems of Oral Examinations in Modern Languages: An Experimental Approach Based on the Cambridge Examinations in English for Foreign Students*. University of Cambridge Local Examinations Syndicate Report Circulated to Oral Examiners and Local Examiners.

Ross, S. (2012). Claims, evidence, and interference in performance assessment. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. Abingdon and New York, NY: Routledge, 223–33.

Şahan, Ő. and Razi, S. (2020). Do experience and text quality matter for raters' decision-making behaviours? *Language Testing*, 1–22. https://doi.org/10.1177/0265532219900228.

Schmidt, E. and Pastorino, C. (forthcoming). Eye-tracking and EEG in language assessment. In G. Yu and J. Xu (eds.), *Language Test Validation in a Digital Age. Studies in Language Testing Volume 52*. Cambridge: UCLES/Cambridge University Press.

Spolsky, B. (1995). *Measured Words*. Oxford: Oxford University Press.

Taylor, L. and Galaczi, E. (2011). Scoring validity. In L. Taylor (ed.), *Examining Speaking: Research and Practice in Assessing Second Language Speaking. Studies in Language Testing Volume 30*. Cambridge: UCLES/Cambridge University Press, 171–233.

Thorndike, E. L. (1912). *Education: A First Book*. New York, NY: Macmillan.

Upshur, J. and Turner, C. E. (1995). Constructing rating scales for second language tests. *English Language Teaching Journal* 49: 3–12. https://doi.org/10.1093/elt/49.1.3.

Wang, X., Evanini, K., Zechner, K. and Mulholland, M. (2017). Modeling discourse coherence for the automated scoring of spontaneous spoken responses. *Proceedings of the 7th ISCA Workshop on Speech and Language Technology in Education*, 132–37.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). Limitations of the Common European framework for developing comparable examinations and tests. *Language Testing* 22: 281–300. https://doi.org/10.1191/0265532205lt309oa.

Weir, C. J., Vidaković, I. and Galaczi, E. D. (2013). *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012. Studies in Language Testing Volume 37*. Cambridge: UCLES/Cambridge University Press.

Wesolowski, B. C. (2019). Predicting operational rater-type classifications using Rasch Measurement theory and Random Forests: A music performance assessment perspective. *Journal of Educational Measurement* 56: 610–25. https://doi.org/10.1111/jedm.12227.

White, E. M. (1984). Holisticism. *College Composition and Communication* 35: 400–9.

Wilds, C. (1979). The measurement of speaking and reading proficiency in a foreign language. In M. L. Adams and J. R. Firth (eds.), *Testing kit: French and Spanish*. U.S. Department of State, Arlington, VI: Foreign Services Institute, 1–12.

Xi, X. (2007). Evaluating analytic scoring for the TOEFL® academic speaking test (TAST) for operational use. *Language Testing* 24: 251–86. https://doi.org/10.1177/0265532207076365.

Yannakoudakis, H., Briscoe, T. and Alexopoulou, T. (2012). *Automating Second Language Acquisition Research: Integrating Information Visualization and Machine Learning*. Proceedings of the EACL 2012 joint workshop of LINGVIS & UNCLH, 35–43. Retrieved from www.aclweb.org/anthology/W12-0206.pdf.

Yannakoudakis, H., Briscoe, T. and Medlock, B. (2011). *A New Dataset and Method for Automatically Grading ESOL Texts*. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 180–89. Retrieved from www.aclweb.org/anthology/P11-1019.pdf.

Yapo, A. and Weiss, J. (2018). Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. Retrieved from scholarspace.manoa.hawaii.edu/bitstream/10125/50557/paper0670.pdf.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing* 27: 37–53. https://doi.org/10.1016/j.asw.2015.11.001.

Section 9

Technology in language testing


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

30

Validity and the automated scoring of performance tests

Xiaoming Xi

Introduction

Since the 1990s, we have seen an increasing use of performance-based tasks in large-scale testing (Bennett, 1993), which has helped generate greater demand for more efficient, affordable, and reliable scoring approaches. Increasingly, automated scoring is becoming a part of the solution to meet this demand. For example, e-rater is used as a quality control on the scores assigned by human raters for the Analytical Writing section of the Graduate Record Examinations (GRE) General Test used for admission into English-medium graduate schools (Powers *et al.,* 2002); the Pearson Test of English (PTE) Academic, a test of English proficiency required for participation in English-medium post-secondary education, uses automated scoring alone to score both written and spoken responses (Pearson, 2019).

Research into the automated scoring of constructed responses began more than five decades ago when Page (1966) investigated the use of automated scoring for scoring essays with a view to providing instant, reliable writing score feedback to students and to relieving classroom teachers of the burden to score a large number of essays. This initial work was in suspension for almost three decades, partly due to limited computing power and access to computers, a major obstacle in applying the research to real-world practice. Automated scoring research regained momentum in the early 1990s with advances in computer technologies and a rising demand for more efficient procedures for scoring essays in large-scale standardized testing (Page, 1994; Page and Dieter, 1995; Page and Peterson, 1995; Shermis *et al*., 1999). With the increasing prevalence of computer-based tests that use constructed-response tasks and the growing maturity of computer scoring technologies, automated scoring is gaining in popularity. The growing use of automated scoring underscores the need for investigating and building shared understanding of relevant validity issues, which is the focus of this chapter.

In contrast to the general positive acceptance of automated scoring in computational linguistics and educational measurement, it has met strong resistance from many language and literature teachers and professionals from the outset (Coombs, 1969; Macrorie, 1969). In an early review of Page and Paulus's essay scoring program, Project Essay Grade, an English professor mockingly noted that students would be likely to write to the computer

DOI: 10.4324/9781003220756-40

and produce "Engfish," boring and dull prose (Macrorie, 1969). Hostility toward automated scoring arose again once automated scoring regained momentum in the 1990s (Roy, 1993). Fitzgerald (1994) wrote passionately in his response to Roy, who had argued for the use of a computer scoring for writing placement purposes:

> The recognition that "writing" is a rhetorical act inseparable from its content, contexts, and purposes has been hard-won. It would be a mistake now to undo these admittedly partial victories in the name of seemingly "new" but, no doubt, "authoritative" technologies.
>
> *(Fitzgerald, 1994: 17)*

Deep skepticism from the teaching profession persisted as automated scoring systems evolved. As automated scoring rose to greater prominence after the turn of the century, Cheville (2004) raised the following question while reflecting on the growing influence of automated essay scoring and feedback tools on the learning and instruction of writing:

> How might the context of high-stakes assessment privilege technologies that actually impoverish students' understanding of language conventions and writing?
>
> *(Cheville, 2004: 47)*

In the field of second language testing, computer technologies have helped transform conceptual thinking and practice about test delivery, definition of test constructs, measurement models, and task design (Chalhoub-Deville, 2001; Jamieson, 2005; Ockey, 2009; Sawaki, this volume). The reactions to automated scoring technologies, however, have been mixed from language testers; the general sentiment can be characterized as *cautiously optimistic* (Xi, 2010). On the one hand, language testers understand the challenges of training, monitoring, and retaining a large number of human raters for large-scale testing programs and are appreciative of the potential benefits scoring technologies could offer in terms of producing faster and more reliable scores. On the other hand, they are keenly aware of the limitations of automated scoring technologies, which put constraints on the ways in which the technologies should be used and the contexts in which they will be appropriate. Language testers, many of whom have language teaching experience, also empathize with teachers' concerns that computer scoring technologies might erode good practice in language learning and teaching.

The vastly different ways and contexts in which automated scoring has been implemented have fueled some of the controversy over automated scoring. In some applications, the use of automated scoring may be easier to justify, whereas in others, it may be strongly disputed. Ignoring the differences in how automated scoring is applied is only going to lead to undifferentiated resistance to automated scoring. Consequently, a major challenge is to establish what validity means for different contexts and uses of automated scoring. A general framework geared toward practice with flexibility to accommodate specific applications of automated scoring is much needed. Language testers typically have interdisciplinary training in linguistics, measurement and testing theories, and statistics. Thus, they are well-positioned with the requisite knowledge and skills to address the challenge of engaging experts in adjacent fields to advance research on validity frameworks for automated scoring and to make recommendations for best practice.

This chapter will chart the evolution of validity theories and practice in relation to automated scoring for performance-based language tests. It reviews historical perspectives on

the validity of automated scoring, discusses the current trends and issues in both conceptual and empirical work with a view toward establishing practical guidelines for best practice, and points to some future directions.

Historical perspectives

The quest for the validity of automated scoring has drawn on evolutions in validity paradigms (see Chapelle and Lee, this volume; Kane, this volume). Initially, automated scoring validity research focused primarily on the agreement between human and automated scores (Page, 1966). This was consistent with the dominant view in the 1950s and 1960s that conceptualized validity as the extent to which test scores are related to criterion scores that are considered the gold standard (Cureton, 1951). Although both the system and the validation approach may seem primitive against the backdrop of today's landscape of automated scoring, Page's work was pioneering in that it opened a whole new world of possibilities for utilizing computer technologies.

The human-machine score comparison, although not the sole criterion for evaluating an automated scoring system, continues to be a highly relevant and intuitive criterion both in academic and lay audience publications. This initial approach to validation has been improved or expanded in several ways. A wide array of evaluation statistics that are complementary in nature have been combined for human-automated score comparison (Yang *et al*., 2002), including agreement rate, correlation, kappa, and comparisons of means and standard deviations. Research has also emerged that compares the generalizability of automated versus human scores (Clauser *et al*., 2000).

Since automated scoring research was reenergized in the 1990s, a new generation of validity research has examined the relationship between automated essay scores and independent measures of language or other abilities, as well as human essay scores (Attali *et al*., 2010; Landauer *et al*., 1998; Petersen, 1997; Powers *et al*., 2002; Weigle, 2010). The purpose of this work is to understand the extent to which automated and human scores relate to other measures of similar or distinct constructs. These studies have provided insights into whether automated scores reflect constructs similar to human scores.

The effect of automated scoring on test-taker behavior is an important consideration from a validity perspective. Powers and his colleagues conducted an innovative study on how e-rater may be fooled by writing experts (Powers *et al*., 2002). The aim of the research was to improve the quality of e-rater and to understand the interactions between the scoring engine and the writer. They concluded that computer scoring alone would not be advisable because even the state-of-the-art technologies can be manipulated in ways that could undermine their fidelity to the writing construct. Their findings also helped enhance future versions of e-rater, which have included mechanisms to flag and exclude from scoring responses that exhibit common techniques used to fool e-rater (e.g., repeating the same paragraph many times, exceedingly long essays).

Critical issues and topics

Validity research on automated scoring has gradually moved beyond human-automated score agreement, as noted earlier. However, as Yang *et al*. have pointed out, previous work has followed a piecemeal approach and focused on one or more of the following three areas: (1) comparing scores produced by automated scoring systems and by human scorers; (2) examining patterns of relationships among automated scores, human scores, and scores

on external measures; and (3) understanding the scoring processes that automated scoring systems employ (Yang *et al*., 2002). These different aspects of validity could potentially be marshaled to support an argument for using automated scoring in an assessment; however, a framework is needed to organize them in a coherent manner. This framework should guide practitioners to determine the critical evidence needed, keeping in mind the way in which automated scoring is implemented and the targeted use of the automated scores. It should also provide a mechanism for synthesizing and evaluating existing evidence for using automated scoring in a specific assessment context.

Although validity research on automated scoring has broadened in scope, an underexplored yet critical area of research is the consequences of automated scoring, especially how test consequences can be integrated into an overall validity argument. Important topics include the acceptance of automated scoring by users; the effect of the use of automated scoring on test-taking strategies; the impact of automated scoring on how scores are perceived, interpreted, and used; the impact on how the test is prepared; and, more broadly, the impact on how languages are taught and learned.

Current contributions and research

Conceptual work

When automated scoring is integrated into an assessment, its most immediate impact seems to be on the accuracy of the resulting scores. This is also the aspect of automated scoring that has attracted the most research. However, because automated scoring introduces unique challenges that are typically not associated with human scoring (e.g., failure to assess essential qualities of speaking performance, susceptibility to cheating, constraining the kind of tasks that can be used), the effects of automated scoring may extend beyond the accuracy of scores and be evident in the generalizability, interpretation, and use of the scores.

On the conceptual front, Bennett and Bejar's work (1998) represents the earliest attempt to position automated scoring in the larger context of an assessment, rather than seeing it merely as a substitute for human scoring. They argued that while automated scoring should be designed in alignment with other aspects of the assessment process, including construct definition, test and task design, test-taker interface, and reporting methods, it influences design decisions related to all these aspects. Bennett and Bejar's conceptual approach is particularly useful for driving the development of a valid computerized assessment that involves automated scoring. By seeing automated scoring as a dynamic component in a computerized assessment system consisting of interrelated elements, their framework emphasizes the importance of evaluating the scoring system in the context of a validity argument for the assessment. This framework allows issues beyond score accuracy to be framed and investigated, such as the meaning of automated scores and the consequences of automated scoring. Consequently, it broadens the scope of validity investigations regarding automated scoring. Although the relationship of automated scoring to the overall validity argument of the assessment is not emphasized, their paper unveils the complexity of validation work related to automated scoring and provides a foundation for the subsequent work that shifts the focus to an overall validity argument.

Based on a critical analysis of empirical validity work on automated scoring, Yang *et al*. (2002) proposed a comprehensive validation approach, drawing largely on Bennett and Bejar (1998). Using this broadened validation framework as the reference point, they noted two gaps in the existing literature. The first one was the dearth of literature that conceptualized

potential threats posed by automated scoring for construct relevance and representation. They also highlighted the role of the consequences of using automated scoring in a validity argument.

Kelly (2001), in investigating the validity of using automated scoring to complement human scoring in large-scale writing tests, adopted Messick's (1995) unitary concept of validity that encompasses six facets: content, structural, substantive, external, generalizability, and consequential. In framing automated scoring validity research in a general validity framework for an assessment, Kelly implicitly supported Bennett & Bejar's view that an automated scoring system is not just a case of replacing the human rater; rather, the system interacts with the other assessment components in complex ways. Drawing on Messick's validity framework that emphasizes test consequences as part of validity, Kelly effectively extended the boundaries to include the impact of automated scoring on stakeholders' interpretation and use of the scores.

Research by Bennett and Bejar, Yang and his colleagues, and Kelly represents an important step toward the development of an overarching, coherent framework for integrating automated scoring into the overall assessment process. Nevertheless, a mechanism is needed to weigh, combine, and synthesize different lines of evidence to support a validity argument.

Clauser *et al*. (2002) responded to the need for such a mechanism. Extending an argument-based validation approach to automated scoring (Kane, 2002; Kane, 2006), their approach subsumed and coherently integrated the various areas of validation reviewed in Yang *et al*. (2002) into a larger validity argument for the whole assessment. It also extended these areas to include decisions based on automated scores and consequences that result from the use of automated scoring. It thus provides a working framework for weaving automated scoring into the validity argument for the whole assessment. They argue that multiple pieces of evidence need to be provided regarding score generalizability, score interpretation, score-based decisions, and consequences of score use to support the use of automated scoring in an assessment. Their framework is useful for guiding the development and synthesis of evidence to support the proposed interpretation and use of automated scores.

With this approach, validation unfolds in two stages: developing an interpretative argument and evaluating a validity argument (Kane, 2006). In the first stage, for each intended use of test scores, an interpretive argument is articulated through a logical analysis of the chain of inferences linking performance on a test to a score-based decision and the assumptions upon which these inferences rest. The second stage involves an evaluation of the plausibility of the interpretive argument within a validity argument given existing evidence.

In applying this framework to automated scoring, Clauser *et al*. (2002) discussed how decisions regarding automated scoring may strengthen the overall validity argument or potentially weaken it. They also discussed unique validity challenges associated with systems developed using different approaches (e.g., expert-articulated rules, rules inferred from expert ratings). Their discussion focused on the potential threats to the strength of each inference in the chain that may be introduced by automated scoring, pointing to the critical areas of research needed to discount or reduce the threats. Although Clauser *et al*. (2002) may not have covered all the potential validity issues introduced by automated scoring, they provided a working model for integrating automated scoring into this network of inferences used to justify the intended interpretation and use of test scores.

Xi (2008) took this work further by discussing how the Clauser *et al*. framework can guide practical validation efforts related to automated systems used for performance-based language tests. Using an actual automated scoring application as an example, she

conducted a thorough analysis of the kinds of threats automated scoring may introduce and the types of evidence that need to be put forward to minimize or reduce these threats. In particular, she proposed that validation priorities may be determined by how automated scoring is used (e.g., alone or in conjunction with human scoring) and the kind of decisions the assessment intends to support (e.g., low-stakes purposes versus medium- or high-stakes decisions). Later in her introduction to a special issue on automated scoring and feedback systems in *Language Testing* (Xi, 2010), she summarized the validity questions that should be asked in validating automated scoring systems for language tests and linked them to specific inferences in the argument-based validation framework (Chapelle *et al*., 2008; Kane, 2006).

Weigle (2013), drawing on Kane's work on validation frameworks, described an approach to validity for using e-rater on the TOEFL iBT by articulating an interpretive argument involving the use of automated scoring. Her work represents another attempt to situate the use of automated scoring in the larger context of assessment use.

Empirical work

Although the earlier empirical validation work on automated scoring for performance-based language tests focused on the agreement between human and automated scores, more recent work has started to show more sophistication and breadth, inspired by advances in conceptual frameworks. Specifically, validation work has been expanded to address the relationships between automated scoring features in relation to the constructs of interest (Enright and Quinlan, 2010; Xi *et al.,* 2008) and the consequences of automated scoring (Kelly, 2001; Xi *et al*., 2016). In addition, the piecemeal approach, in which a single validity aspect is investigated without connections to other evidence, has evolved into a more complex, integrated approach that synthesizes multiple pieces of evidence to support the use of automated scoring for a specific assessment (Bernstein *et al*., 2010; Enright and Quinlan, 2010; Kelly, 2001; Xi *et al*., 2008).

Recommendations for practice

Asking relevant validity questions

Given the many different areas of expertise required to fully understand the complexity of automated scoring systems, it seems a daunting task for practitioners to attempt to validate their use. A useful first step is to ask the following questions:

For what purpose are the scores intended to be used?
How is the automated scoring system used to produce the scores? As the sole rater? As a check for human scores? Or as a contributory score in the final reported score?

It is crucial to ask these initial questions to determine the types of evidence that should be prioritized and the standards against which the results should be evaluated to support the use of the system.

The following is a list of the relevant questions to be asked that address the accuracy, generalizability, meaningfulness, and usefulness of the scores and the consequences of using automated scoring (Xi, 2010). The questions correspond to the specific inferences in

a validity argument that need to be supported (Chapelle *et al*., 2008; Kane, 2006). For each question, a brief discussion of the issues and the relevant evidence needed is included.

- *Does the use of assessment tasks constrained by automated scoring technologies lead to construct underrepresentation or misrepresentation?* (domain definition)

There is a tension between task types included in an assessment and automated scoring capabilities, especially when the computer is intended as the only rater. Automated scoring for simple, constrained tasks may lead to higher score accuracy and reliability, but heavy reliance on these tasks in high-stakes testing may lead to construct underrepresentation and incur negative washback on teaching and learning (see also Yan and Fan, this volume). In contrast, systems designed to score more complex, performance-based tasks lack the level of sophistication and maturity that is required for accurate evaluation of the quality of the performance if used as the only rater. In designing and validating an assessment that uses automated scoring, this tradeoff between test design and automated scoring capabilities needs to be considered.

So far, automated scoring technologies have been used to score a variety of speaking tasks in large-scale high-stakes tests, including read aloud, sentence repeat, oral summarization, picture description, story recall, etc. For more complex tasks such as synthesizing information from multiple sources to present it orally or small group discussion tasks, automated scoring technologies may not be mature enough to support using automated scoring as the only rater in high-stakes testing. Completeness, accuracy and representation of content, coherence, and topic development are some of the higher-order skills that are challenging for automated scoring technologies to assess. Using automated scoring in combination with human scoring may be a way forward to explore the benefits of automated scoring without having fully addressed all these concerns, as the presence of human scoring will negate the importance of some of them. This approach can potentially improve the reliability and efficiency of scoring while not constraining the type of tasks used to measure communicative language abilities.

To evaluate whether computer scoring has resulted in a more limited domain representation, a critical analysis of the test tasks in relation to the target language use domain involving both the characteristics of the tasks and the knowledge, skills, and processes involved will provide the relevant evidence.

- *Do the automated scoring features underrepresent or misrepresent the construct of interest?* (explanation)
- *Is the way the scoring features are combined to generate automated scores consistent with theoretical expectations of the relationships between the scoring features and the construct of interest?* (explanation)
- *Does the use of automated scoring change the meaning and interpretation of scores provided by trained raters*? (explanation)

Although automated scoring may allow some degree of "control" of the constructs we intend to measure, systematic errors or deficiencies may be introduced that result in construct underrepresentation or construct irrelevance. The tendency to extract easily computable, quantifiable aspects of the performance due to the limitations of current technologies would potentially result in construct-irrelevant features or features that do not represent the

full construct. For example, the construct in an oral assessment might include a subconstruct for pronunciation, further defined by several features, not all of which are yet adequately captured by the automated scoring system (e.g., comprehensibility). In addition, although it is a desirable goal to combine automated features in a principled, construct-relevant way, given the complexity of human raters' decision-making processes, it obviously is not an easy task to design a scoring system that adequately reflects those processes.

Evidence required to answer these related questions is conceptual and judgmental. Two essential qualities need to be verified to ensure that automated scoring captures relevant and meaningful aspects of performance: (1) that the construct relevance and coverage of the features is adequate and (2) that the manner in which features are combined can be justified. The evidence largely involves judgments of these qualities by experts who have an intimate understanding of the construct the assessment is designed to measure, the conceptual meaning of each scoring feature used, and the way the scoring features are combined through a statistical model to produce a score that indicates the overall quality of performance.

- *Does automated scoring yield scores that are accurate indicators of the quality of a test performance sample? Would examinees' knowledge of the scoring logic of an automated scoring system impact the way they interact with the test tasks, thus negatively affecting the accuracy of the scores?* (evaluation)

The relevant evidence includes the association between human and automated scores indicated by traditional measures of agreement, such as correlation and kappa. Different scoring methodologies that are used to produce automated scores based on automatically extracted features can be compared based on the strength of the association between the model predicted scores and the human scores. The soundness of the statistical principles underlying each methodology is also an important consideration in employing a scoring methodology.

When evaluating the strength of the relationship between human and automated scores, care must be taken to verify the conditions under which test takers produce the performance samples used for human-automated score comparison (e.g., whether they are aware of computer scoring, whether they have any knowledge about the logic of the computer scoring). Evaluation statistics obtained by running performance samples produced for human graders through an automated scoring engine may be misleading indicators of the quality of the automated scores on performance samples produced for a computer grader.

- *Does automated scoring yield scores that are sufficiently consistent across measurement contexts (e.g., across test forms, across tasks in the same form)?* (generalization)

Two types of evidence can help address this question. One concerns the procedures for developing and evaluating the scoring models, such as the adequacy of the sample size, the representativeness of the sample, and the absence of overlap in candidates between the scoring model training and evaluation data. The second type of evidence includes the reliability of the automated scores across different tasks and test forms. The score reliability estimates could be compared to those obtained for human scores and typical figures acceptable for a context.

- *Does automated scoring yield scores that have expected relationships with other test or non-test indicators of the targeted language ability?* (extrapolation)

The potential evidence for answering this question is the association between automated scores and scores on other relevant test or non-test indicators of language ability. It is helpful to compare the strengths of relationships for human scores and automated scores with scores on these indicators to see if these human and automated scores relate to the other indicators in similar ways. This information will help us decide whether weak relationships between automated scores and scores on these relevant indicators are due to differences in the test measures or the scoring method.

- *Do automated scores lead to appropriate score-based decisions?* (utilization)
- *Does the use of automated scoring have a positive impact on examinees' test preparation, teaching, and learning practices?* (utilization)

Replacing or complementing human scoring with automated scoring may change users' perceptions of the assessment, the way they interact with the assessment tasks, and the way languages are learned and taught. Knowing that scores are produced by automated means or by a combination of automated and human scoring means rather than solely by human raters, users may also interpret and use the scores differently.

Arguments for the usefulness of automated scores are supported by an analysis of the magnitude of the error associated with the intended score-based decision. Arguments about potential consequences of using automated scoring may be supported by investigations of user perceptions: e.g., to what extent the awareness of the scores being produced by a machine impacts the way a user interprets and uses the scores, as well as the impact of using automated scoring on test preparation, teaching, and learning practices.

Determining validation priorities

Both the way automated scoring is implemented and the intended use of the scores determine the validation priorities. For example, the areas of emphasis for validating an automated scoring system intended for a practice environment may differ from those for a system employed in an assessment for high-stakes decisions. If automated scores are intended to support high-stakes decisions, satisfactory answers need to be provided to all the questions raised here. The domain representativeness of the test tasks (or the domain definition inference) and the meaningfulness of the scores (or the explanation inference) are critical yet are most likely to be compromised by automated scoring – if the assessment tasks or automated scoring models underrepresent or misrepresent the construct of interest, the assessment scores may be inadequate indicators of test takers' language ability. Automated scoring may also make the assessment more susceptible to cheating and test-gaming strategies that would negatively impact the trustworthiness of the scores. An automatically scored test that underrepresents or misrepresents the construct of interest either as a result of constrained task designs or deficient scoring models may also incur negative washback effects on teaching and learning and hurt the credibility of the testing program. Such inappropriate use may serve to stigmatize any use of automated scoring by further strengthening the skepticism of its critics.

If an automated scoring system is intended to be used to check the accuracy of human scores or to contribute to the final score along with human scoring, the test design need not be constrained by limitations of automated scoring technologies. Therefore, the use of automated scoring is not expected to introduce additional issues related to the domain definition

inference. However, the evaluation and generalization inferences would require adequate backing. It is less critical for the other three inferences – explanation, extrapolation, and utilization – to be fully supported, given the concurrent use of human scoring.

In evaluating the evidence put forward, different criteria should be used for determining the strength and sufficiency of the evidence for the intended use. Automated scoring used for high-stakes purposes obviously should be subject to higher standards than for low-stakes purposes, with regard to (1) the level of score accuracy, (2) generalizability of scores, (3) the degree to which the scores represent the essential qualities of the performance samples, (4) the degree to which the scores relate to scores on external criterion measures, (5) the degree to which the scores support decision making, and (6) the consequences of using automated scoring.

Because automated scoring may impact more than one aspect of the validity argument, in developing an automated scoring system, we frequently need to reconcile different priorities and determine a design that maximizes the overall quality of the assessment.

Correspondingly, in evaluating a system, we also need to be cognizant that some design decisions may strengthen some aspects of the validity argument while compromising others. For example, automated scoring may reduce task specificity by capturing aspects of performance that are relatively stable across tasks, thus improving the score generalizability. However, it may also reduce task specificity in undesirable ways. It may compromise the explanatory power of the scores in representing the constructs by failing to include some aspects that are construct relevant but are less stable across tasks, thus weakening the explanation inference.

Based on the process discussed earlier, the relevant evidence pertaining to each question or inference can be integrated and evaluated. Then the overall strength of the validity argument can be evaluated in light of the critical inferences that need adequate backing to support the intended claims of automated scoring.

Future directions and developments

Validity research on automated scoring has come a long way from its initial focus on human-machine score comparison. The increasing use of automated scoring in high-stakes contexts necessitates heightened attention to related validity issues, which is accelerating the maturation of research in this area. We expect to see more depth in the conceptual thinking about validation frameworks, more sophistication in the validation methods, and more breadth in the kind of issues investigated.

Establishing standards for evaluating automated scoring

As discussed earlier, there have been a number of approaches to combining and synthesizing multiple types of evidence to develop validity arguments for automated scoring. As our understanding evolves for different contexts and uses of automated scoring, it is imperative to develop standard structures for validity arguments for typical applications of automated scoring. For example, an assessment may use automated scoring alone, as a check, or as a contributing score or use both machine and human raters to score different aspects of the same performance samples to generate scores used for high-, medium-, and low-stakes decisions. Established standards for expected system performance in different use contexts can then facilitate the practical work of evaluating aggregated evidence.

Perceptions by different stakeholders and group-related perception differences

Among other factors, the rigor with which an automated scoring system is developed for a test and its technical and substantive quality may directly impact stakeholders' (e.g., test takers', score users', and teachers') perceived confidence in the scores and their use of them. Each group of stakeholders also brings in their own background and training and their interests and priorities associated with the test. Therefore, their levels of confidence in and acceptance of the scores may also differ. Even within a group of stakeholders (for example, test takers), perceptions of and interactions with automated scoring may depend on their culture, age, discipline, exposure to and attitude toward technology, and their knowledge about the specific scoring system. Research has started to address some of the open questions through surveys, interviews, and focus groups (Kelly, 2001; Xi *et al*., 2016). More research is certainly needed to address many of the remaining questions.

Test-taking strategies adapted to computer scoring

Test takers may adapt to the grading system for the purpose of getting higher scores. They may adapt their test-taking strategies in systematic or idiosyncratic ways based on their understanding of the grading criteria (see Cohen, this volume). Writing or speaking to a computer is likely to trigger more systematic changes in test-taking strategies if the scoring logic is known because computer scoring logic is typically rule based and well defined. This possibility raises intriguing research questions, such as how test takers might respond to the test tasks differently because of computer scoring or computer scoring combined with human scoring in some fashion.

Survey, interview, and focus group methods have been used to gather this kind of information from test takers (Xi *et al*., 2016). More direct methods are needed, such as stimulated recall, which allows insights into the actual test-taking processes engaged by test takers when writing for a human rater, the computer, or a combination of human and computer scoring.

Correspondingly it will be useful for us to understand through discourse analysis how candidates' performance samples may have been altered by test-taking strategies specific to computer scoring and how human raters would interact with these samples. Some alterations may result from construct-relevant strategies (i.e., monitor one's pronunciation more closely) while others may not be consistent with the intended construct (e.g., keep talking even if the speech makes little sense or writing long essays without much concern for the content). For the construct-relevant strategies, it would be interesting to explore how they impact the manifestation of other qualities of the performance samples and whether similar strategies would emerge in response to human scoring. For example, how would a greater focus on pronunciation impact speech accuracy and fluency, given the competition for resources during real-time speech production and thus the overall speech quality?

A look into both the processes and the products in relation to automated scoring will offer insight into the extent to which changes in test-taking strategies lead to changes in score accuracy, meaning, and interpretation.

Impact on language learning and teaching in different contexts

The increasing use of automated scoring for performance-based language tests either in high-stakes environments or for practice and diagnostic feedback is expected to impact the

teaching and learning of language skills in profound ways. Changes in test-taking strategies in response to computer scoring may be prompted by the motivation to "trick" the computer into giving higher scores or may be due to more sustained effects introduced by automated scoring on test preparation, coaching, and language teaching.

As an initial step, it is necessary to understand how language learning and teaching theories and practices might adapt to the use of automated scoring in tests with high-stakes decisions and whether the adaptations are positive or negative from the perspective of learning. Surveys of teachers' classroom teaching techniques geared toward the use of computer scoring may reveal interesting issues to be further explored through direct contact with teachers in interviews, focus groups, and classroom observations.

The reactions to automated scoring and feedback capabilities targeted for classroom use on classroom teaching may be mixed, depending on the contexts of teaching (e.g., ESL versus EFL) and the qualifications of the language teachers (e.g., native vs. non-native speakers of the target language). In studying the impact of automated scoring and feedback capabilities on teaching, factors such as competence in the target language, teaching philosophies and methodologies, level of knowledge, and understanding of the specific automated scoring system need to be considered in the design and interpretation of the findings.

In the last two decades, because of the longer history and more advanced development of automated writing evaluation (AWE) technologies, we have seen increasing numbers of studies on the use of AWE in classroom contexts, including the application of an argument-based approach to validating the use of AWE systems for diagnostic purposes (Chapelle *et al*., 2015), examining the impact of AWE on the feedback teachers provide, students' writing revisions and their writing skill development (Link *et al*., 2020), and comparing the ways in which students engage with teacher feedback versus automated feedback (Zhang and Hyland, 2018). In contrast to the flourishing of research on AWE systems in classroom contexts, investigations of the use of speech technologies are scarce, pointing to future research needs in this area.

Promoting AI literacy among stakeholders

The development of automated scoring and feedback technologies has had the longest history and seen the most popularity among applications of artificial intelligence (AI) technologies in language learning and assessment. AI has also been used for automated test item creation and adaptive learning solutions. Given the increasing use of AI technologies in language assessment, AI literacy should be considered an important part of assessment literacy. However, very little attention has been given to AI literacy in assessment literacy research and practice, given that use of automated scoring technologies in assessments has not become part of the standard training for language testing researchers and practitioners yet.

Developing AI literacy takes the concerted efforts of developers and users. AI scientists and test developers have major responsibilities for communicating relevant information about AI and promoting literacy. Test takers, teachers, and assessment policy makers have responsibilities for developing literacy about AI and rights to demand information about how AI works, how it is used in a specific context, and if acceptable quality has been ensured for their context.

Developing AI literacy requires understanding the inner workings of automated scoring. The more intimate knowledge different stakeholders have about automated scoring, its

strengths, and its potential pitfalls, the more likely we will be able to promote appropriate use of the technology.

In automated scoring, there is the black box versus the white box approach. The black box is sometimes the result of an automated scoring provider's lack of willingness to reveal what they consider trade secrets. Other times, the method chosen to train scoring models is referred to as a black box.

Some black box methods include artificial neural networks (ANNs) and deep learning (DL) (Schmidhuber, 2015), which have dominated recent machine learning applications. The networks are supposed to mimic how the brain works. Unfortunately, such networks are also as opaque as the brain and very hard to decipher and interpret.

In automated scoring work, this black box approach has some legitimate applications. I argue that when users do not interact with AI outputs directly, the black box approach is appropriate. Examples would include using DL in speech recognition or extracting specific errors or other information to derive higher-level automated scoring features, such as grammatical accuracy. In addition, when the machine score is only used as a check on a human score, not as a contributing score or the sole score, the black box approach is appropriate. The black box approach may also work when the scores are used for low-stakes decisions since the test takers are less motivated to game the automated scoring system.

On the other hand, a white box approach that focuses on transparent, construct-driven feature extraction and model building has several advantages. Construct-driven automated features are faithful to the test construct and more interpretable to test users, thus facilitating the development of their AI literacy. Also, the same scoring features could potentially be used to provide diagnostic feedback.

In terms of communication to users, I believe that the following information needs to be communicated to test users in an accessible way:

- Types of automated scoring features used
- How features are combined to predict scores
- Accuracy and consistency of score prediction
- Anticipated impact of the use of automated scoring in specific contexts

Opening the black box would be a key step toward building a virtuous cycle in promoting assessment literacy. It starts with developing cross-disciplinary expertise so that the inner workings of AI can be communicated in language easy to understand by users. Currently, public communication about AI technology in assessment is still opaque and omits or glosses over key information, such as scoring features and scoring models. Greater transparency about AI technology would help promote more literate and savvy users. Users equipped with relevant knowledge about AI would also be in a better position to push for more thoughtful and responsible applications in assessment.

Validity frameworks for using automated scoring in large-scale testing versus classroom assessment

Most current validity frameworks focus on validity issues related to automated scoring used in large-scale testing, which is typically norm referenced, aiming to spread out the test takers. However, as Moss (2003) pointed out, classroom assessments serve different purposes: e.g., whether individual students have mastered certain content, knowledge, and skills. She

further argued that conventional validity frameworks conceptualized for large-scale standardized testing may be ill-fitted to classroom assessments that are formative and criterion based (Moss, 2003). As automated scoring systems are making inroads into classrooms, new conceptions of good practice are much needed to evaluate the use of automated scoring in formative, classroom-based assessments. Chapelle and Chung (2010) and Xi (2010) have raised some validity issues regarding the use of automated feedback systems, but they are framed in Kane's argument-based validation framework, which is best suited to large-scale testing. Automated scoring and feedback systems may be embedded in classroom instruction and assessment activities that serve drastically different purposes than standardized testing. Therefore, any validity frameworks for this context should highlight the prominent characteristics of classroom assessments. For example, for diagnostic assessments that teachers use to understand learners' strengths and weaknesses, validity inquiries should focus on examining the extent to which diagnostic feedback is accurate, reliable, meaningful, actionable, and effective (Xi and Zhang, 2020; see also Jang and Sinclair, this volume). Language teaching professionals have much to offer in helping validity theorists think through the most relevant issues for using automated scoring in classroom contexts.

Concluding remarks

To conclude, we expect to see validation frameworks for automated scoring continue to grow and to mature and empirical research to expand into less explored areas, such as test takers' interactions with the scoring system and the consequences of automated scoring on test preparation, teaching, and learning. In the next decade, we may see some consensus developed for the criteria used for evaluating the use of automated scoring in high-stakes contexts given heightened attention to this controversial area of research. The use of automated scoring and feedback capabilities in classroom assessments may also gather more traction because of the increased participation of language teaching professionals in this debate. The coming decade may see a greater clarification regarding the prominent validity issues and unique challenges in this context. However, debates will continue as to what constitutes acceptable performance for other applications of automated scoring (e.g., automated scoring used for medium-stakes purposes).

The use of automated scoring is expected to have considerable impact on how language abilities are defined, assessed, learned, and taught. As language testers, we have the obligation to actively engage in validation research and practice, to fully participate in cross-disciplinary conversations that inform best practices, and to help promote AI literacy among stakeholders, including test takers, learners, teachers, educators, etc. Ultimately, it should be the human who harnesses the technology to realize its positive potential for users.

Further reading

Bennett, R. E. and Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice* 17: 9–17. https://doi.org/10.1002/j.2333-8504.1997.tb01734.x.
This article provides the first comprehensive treatment of automated scoring as an integral part of an assessment, which is a departure from seeing it as merely a substitute for human scoring. A central argument in this article is that automated scoring should be designed as a dynamic component in the assessment process, interacting with the construct definition, test and task design, test taker interface, and reporting methods.

Clauser, B. E., Kane, M. T. and Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education* 15: 413–32. https://doi.org/10.1207/S15324818AME1504_05.
This is the first attempt to integrate automated scoring into an argument-based validation framework, thus highlighting the complex and expanded effects that automated scoring may have on the overall validity argument for the entire assessment. It discusses potential validity threats to the strength of each inference in the validity argument that may be introduced by automated scoring, pointing to the critical research that is needed to discount or reduce the threats.

Xi, X. (2010). Automated scoring and feedback systems – Where are we and where are we heading? *Language Testing* 27: 291–300. https://doi.org/10.1177/0265532210364643.
This introduction to a special issue offers a critical review of the research and development work in automated scoring and feedback systems for language assessment and learning. It raises a series of relevant validity questions for automated scoring and feedback systems, respectively. These validity questions are linked to the inferences in the argument-based validation framework. It also provides a brief overview of the seven articles featured in the special issue.

Zechner, K. and Evanini, K. (eds.). (2020). Automated speaking assessment: Using language technologies to score spontaneous speech. In J. Norris, S. Ross, S. Weigle and X. Xi (eds.), *Innovations in Language Learning and Assessment at ETS*. New York, NY: Routledge.
This volume provides a state-of-the-art review of research and development of ETS's speech technologies situated in the broader context of developments and advances in the field. It attempts to open the black box of automated speech scoring by offering an under-the-hood analysis of the components of automated speech scoring systems, including the speech recognition system, speech features, and speech scoring models. It also discusses validity issues involved in automated speech scoring and provides an overview of recent developments and future outlook.

References

Attali, Y., Bridgeman, B. and Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment* 10. Retrieved December 21, 2020, from https://files.eric.ed.gov/fulltext/EJ895962.pdf.

Bennett, R. E. (1993). On the meaning of constructed response. In R. E. Bennett and W. C. Ward (eds.), *Construction Versus Choice in Cognitive Measurement: Issues in Constructed Response, Performance Testing, and Portfolio Assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1–27.

Bennett, R. E. and Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice* 17: 9–17. https://doi.org/10.1002/j.2333-8504.1997.tb01734.x.

Bernstein, J., Van Moere, A. and Cheng, J. (2010). Validating automated speaking tests. *Language Testing* 27: 355–77. https://doi.org/10.1177/0265532210364404.

Chalhoub-Deville, M. (2001). Language testing and technology: Past and future. *Language Learning and Technology* 5: 95–97. Retrieved December 21, 2020, from http://llt.msu.edu/vol5num2/deville/default.html.

Chapelle, C. A. and Chung, Y.-R. (2010). The promise of NLP and speech processing technologies in language assessment [Special issue]. *Language Testing* 27: 301–15. https://doi.org/10.1177/0265532210364405.

Chapelle, C. A., Cotos, E. and Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing* 32: 385–405. https://doi.org/10.1177/0265532214565386.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. Mahwah, NJ: Lawrence Erlbaum.

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *English Journal* 93: 47–52.

Clauser, B. E., Kane, M. T. and Swanson, D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education* 15: 413–32. https://doi.org/10.1207/S15324818AME1504_05.

Clauser, B. E., Swanson, D. B. and Clyman, S. G. (2000). A comparison of the generalizability of scores produced by expert raters and automated scoring systems. *Applied Measurement in Education* 12: 281–99. https://doi.org/10.1177/01466210022031796.

Coombs, D. H. (1969). Review of the analysis of essays by computer by Ellis B. Page and Dieter H. Paulus. *Research in the Teaching of English* 3: 222–28.

Cureton, E. F. (1951). Validity. In E. F. Lindquist (ed.), *Educational Measurement*, 1st edn. Washington, DC: American Council on Education, 621–94.

Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring [Special issue]. *Language Testing* 27: 317–34. https://doi.org/10.1177/0265532210363144.

Fitzgerald, K. R. (1994). Computerized scoring? A question of theory and practice. *Journal of Basic Writing* 13: 3–17.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics* 24: 228–42. https://doi.org/10.1017/S0267190505000127.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin* 112: 527–35. https://doi.org/10.1037/0033-2909.112.3.527.

———. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice* 21: 31–35. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x.

———. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurement*, 4th edn. Washington, DC: American Council on Education/Praeger, 18–64.

Kelly, P. A. (2001). *Automated Scoring of Essays: Evaluating Score Validity*. Unpublished dissertation, Florida State University.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes* 25: 259–84. https://doi.org/10.1080/01638539809545028.

Link, S., Mehrzad, M. and Rahimi, M. (2020). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2020.1743323.

Macrorie, K. (1969). Review of the analysis of essays by computer by Ellis B. Page and Dieter H. Paulus. *Research in the Teaching of English* 3: 222–36.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50: 741–49. https://doi.org/10.1037/0003-066X.50.9.741.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practices* 22: 13–25. https://doi.org/10.1111/j.1745-3992.2003.tb00140.x.

Ockey, G. J. (2009). Developments and challenges in the use of computer-based testing for assessing second language ability. *The Modern Language Journal* 93: 836–47. https://doi.org/10.1111/j.1540-4781.2009.00976.x.

Page, E. B. (1966). The imminence of grading essays by computers. *Phi Delta Kappan* 47: 238–43.

———. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education* 62: 127–42. https://doi.org/10.1080/00220973.1994.9943835.

Page, E. B. and Dieter, P. (1995). *The Analysis of Essays by Computer, Final Report of U.S. Office of Education Project No. 6–1318*. Storrs, CT: University of Connecticut. ERIC ED 028 633.

Page, E. B. and Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan* 76: 561–65.

Pearson. (2019). *Pearson Test of English Academic: Automated Scoring*. Retrieved January 2, 2021, from PTE Academic Automated Scoring White Paper (pearsonpte.com).

Petersen, N. S. (1997, March). *Automated Scoring of Writing Essays: Can Such Scores Be Valid?* Paper presented at The Annual Meeting of the National Council on Education, Chicago, IL.

Powers, D. E., Bursterin, J., Chodorow, M. S., Fowles, M. E. and Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Educational Computing Research* 26: 407–25. https://doi.org/10.2190/CX92-7WKV-N7WC-JL0A.

Powers, D. E., Burstein, J. C., Fowles, M. E. and Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computer in Human Behavior* 18: 103–34. https://doi.org/10.1016/S0747-5632(01)00052-8.

Roy, E. L. (1993). Computerized scoring of placement exams: A validation. *Journal of Basic Writing* 12: 41–54. Retrieved December 21, 2020, from www.jstor.org/stable/43443612.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks* 61: 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z. and Harrington, S. (1999, April). *Trait Ratings for Automated Essay Grading*. Paper presented at the Annual Meeting of National Council on Measurement in Education, Montreal, Canada.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing* 27: 335–53. https://doi.org/10.1177/0265532210364406.

———. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing* 18: 85–99. https://doi.org/10.1016/j.asw.2012.10.006.

Xi, X. (2008). What and how much evidence do we need? Critical considerations in validating an automated scoring system. In C. A. Chapelle, Y. R. Chung and J. Xu (eds.), *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment*. Ames, IA: Iowa State University, 102–14.

———. (2010). Automated scoring and feedback systems – Where are we and where are we heading? *Language Testing* 27: 291–300. https://doi.org/10.1177/0265532210364643.

Xi, X., Higgins, D., Zechner, K. and Williamson, D. M. (2008). *Automated Scoring of Spontaneous Speech Using SpeechRater v1.0* (ETS Research Rep. No. RR-08–62). Princeton, NJ: ETS.

Xi, X., Schmidgall, J. and Wang, Y. (2016). Chinese users' perceptions of the use of automated scoring for a speaking practice test. In G. Yu and Y. Jin (eds.), *Assessing Chinese Learners of English*. London: Palgrave Macmillan.

Xi, X. and Zhang, C. Q. (2020). Validity and validation in language assessment: Developments and challenges. *China Exams* 6: 19–26.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J. and Bhola, D. S. (2002). A review of strategies for validating computer automated scoring. *Applied Measurement in Education* 15: 391–412. https://doi.org/10.1207/S15324818AME1504_04.

Zechner, K. and Evanini, K. (eds.). (2020). Automated speaking assessment: Using language technologies to score spontaneous speech. In J. Norris, S. Ross, S. Weigle and X. Xi (eds.), *Innovations in Language Learning and Assessment at ETS*. New York, NY: Routledge.

Zhang, Z. V. and Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing* 36: 90–102. https://doi.org/10.1016/j.asw.2018.02.004.

31
Computer-based testing

Yasuyo Sawaki

Introduction

With the rapid advancement of computer technology, the number of language assessments that utilize such technology in designing, developing, and delivering test content as well as scoring and reporting examinee test performance is growing steadily. With the turn of the twenty-first century, taking a language test on a computer has become a familiar activity for language learners. Various types of technologies that have traditionally been utilized in language assessment, such as audio, video, and telephone, have also been replaced with computer technology, enabling the testing of language learners as long as one has access to a computer. The primary purpose of this chapter is to summarize various substantive and empirical issues surrounding the use of computer-based testing (CBT) in language assessment, with a specific focus on technology use for test administration. (For an extensive discussion of a related topic, automated scoring in language assessment, see Xi, this volume.) This chapter provides a brief overview of the historical development of CBT. This will be followed by a critical discussion of current issues of concern, particularly those that have important implications for the validity of CBT. The chapter will conclude with a discussion of future CBT research in the field of language assessment.

Historical perspectives

The introduction of CBT to language assessment dates back to the early 1980s. Chapelle and Douglas (2006) refer to a 1981 publication that describes a CBT developed as part of a French-as-a-foreign-language curriculum at the University of Illinois. The Language Testing Research Colloquium (LTRC) in 1985 was held on the theme of technology and language testing, yielding a volume of selected papers from the conference edited by Stansfield (1986). Since then, conference presentations and publications on applications of and research into CBT have been increasing steadily, as showcased in multiple volumes and review articles on this topic published in the field over the last two decades (e.g., Brown, 1997; Chalhoub-Deville, 1999; Chalhoub-Deville and Deville, 1999; Chapelle and Voss, 2016; Douglas, 2013; Douglas and Hegelheimer, 2007; Jamieson, 2005). This section

 DOI: 10.4324/9781003220756-41

summarizes how CBT research has unfolded in the field so far and considers our future directions based on the three phases of CBT advancement proposed by Bennett (2000). While Bennett depicted his scenario in the broader context of large-scale educational measurement nearly 20 years ago, it is highly applicable to describing the advancement of CBT research in language assessment so far as well as its future.

Phase 1: Improving efficiency

According to Bennett (2000), the primary goal of the initial phase of the CBT development is to increase the efficiency of test administration and scoring by "automat[ing] an existing process" (Bennett, 2000: 3). This can be achieved in several ways. For instance, existing PBT items may be transferred to computers with little change to the test content. Such projects would typically involve tests based on selected-response items partly because they can more readily be converted to a CBT format than complex constructed-response items. An example is a project reported by Fulcher (1999), in which an existing ESL placement test at a UK university was computerized. For minimizing scoring time, the grammar part of the test comprising 80 multiple-choice items was converted to a CBT format, while the essay questions were administered on paper.

Another approach to increasing efficiency by transitioning to CBT is to employ an adaptive format. Unlike a conventional test in which all examinees receive the same items regardless of ability levels, adaptive test content is tailored to the ability level of each examinee. As can be seen in Chalhoub-Deville and Deville's (1999) review of the number of computer adaptive test (CAT) development projects in language assessment conducted in the 1980s and 1990s, quite a few early forms of CBT were CATs. In a typical CAT, the examinee's rough ability estimate is obtained based on his/her performance on an initial set of items. Then an item with the difficulty level suited to the examinee's ability estimate is chosen from an item bank and presented as the next item for updating the ability estimate. This process is repeated until an accurate enough examinee ability estimate is obtained. An algorithm behind a full-fledged CAT is typically based on item response theory (IRT), a measurement model that allows estimation of examinee ability level and item difficulty on the same scale (see Ockey, this volume).

While the advantages and disadvantages of CAT were discussed in the field from early on (e.g., Stansfield, 1986), the efficiency of CAT in particular was thought to bring language assessment distinct advantages: reduction of testing time, individual testing of examinees, flexibility in scheduling test administrations, availability of precise examinee ability estimates immediately after testing, and a less stressful test experience for the examinee with presentation of test items matched to his/her ability level. Accordingly, various commercial tests with CAT capabilities were launched from the late 1990s including TOEFL CBT (ETS), Business Language Testing Service (BULATS; Cambridge ESOL), and Computerized Assessment System for English Communication (CASEC; Japan Institute of Educational Measurement).

Despite its various potential advantages, however, CAT is not without its limitations. First, developing and maintaining a CAT is extremely expensive because CAT requires continuous item development and large-scale pre-testing efforts to maintain an item bank with new items having item parameter estimates (Nogami and Hayashi, 2010). Second, a CAT format is not necessarily compatible with other assessment task design features. For example, in 2005 the TOEFL CBT was replaced with the TOEFL iBT, which does not employ a CAT format. According to Jamieson *et al*. (2008), among the reasons for this decision was

the conceptual incompatibility between the new task types that integrate modalities and a psychometric assumption of IRT. Another was the logistical challenge of obtaining item parameter estimates for these complex performance-based, human-scored tasks, which did not allow on-the-fly estimation of examinee ability levels during test sessions.

Meanwhile, CBT with self-adaptive capabilities that are less resource intensive than IRT-based CAT emerged in the field as well. One example is the computer-based OPI (COPI) developed by the Center for Applied Linguistics, which is a simulated version of the ACTFL Oral Proficiency Interview (Kenyon and Malabonga, 2001). Because scoring examinee speaking responses during the testing session is infeasible, the COPI examinee chooses the difficulty level of the task based on results of a self-assessment administered at the beginning of a test session. Despite concerns as to whether learner self-assessment can serve as a valid adaptive algorithm (e.g., Norris, 2001), Malabonga *et al*. (2005) supported its validity with high correlations of examinee self-assessment with final examinee ability estimate and faculty rating of examinee speaking ability. Another example is DIALANG, a web-based diagnostic self-assessment of 14 European languages based on the Common European Framework of Reference (Council of Europe, 2001), which combines a self-assessment of overall language ability with a vocabulary test for screening to deliver an item set suited to each examinee's ability level (Alderson and Huhta, 2005; see Read, this volume).

Phase 2: Introducing innovative task design and new CBT delivery modes

As noted earlier, Bennett's (2000) initial phase of CBT focuses primarily on changes in the test format to seek improvements in efficiency. In contrast, the second phase is characterized by taking advantage of functionalities available only on computers, resulting in substantive changes to various aspects of test task design. With the inception of large-scale testing programs that were developed as CBTs from scratch (e.g., PTE Academic by Pearson) and the exploration of computer functionalities that support innovative task designs, CBT development and studies along this line have quickly accumulated in recent years. First, various new test task types that better reflect characteristics of real-life language use tasks have been proposed. For instance, an email writing task can be better emulated in CBT than in PBT because CBT allows examinees to word process their responses (e.g., the TOEIC writing test). Similarly, in a voice message listening task, the CBT can give the examinee a full control over replay of the listening stimuli as done on an answering machine (e.g., the WebLAS-Japanese test; www.weblas.ucla.edu). Functionalities available on computers make innovative response formats possible as well. In one task type in the TOEFL iBT reading section, the examinee drags and drops appropriate options to complete a prose summary of the stimulus text. Moreover, the use of a computer enhances control over the timing for presentation of different parts of a complex task. In the integrated task in the TOEFL iBT writing section, for instance, the examinee reads a written academic text first. After the text disappears, a lecture on the same topic is presented. After the lecture is complete, the written text reappears, along with the space for the examinee to word process a summary of the relationship between the two. In the TOEFL iBT listening section, only the item stem of each multiple-choice question is presented in audio and writing first, followed by the options that appear on the screen some seconds later to encourage the examinee to formulate his/her own answer before reviewing the options.

Bennett (2000) stated that this phase may witness changes in CBT delivery modes, which has indeed happened to CBT in language assessment over the years. In early CBT

applications, test content was distributed to individual computer platforms via CD-ROMs or closed local networks (e.g., TOEFL CBT and CD-ROM versions of the Cambridge ESOL Suite of language tests). Meanwhile, concerns were raised about the compromised security of web-based test delivery for high-stakes purposes (Alderson, 2000b; Roever, 2001). Over the last 15 years or so, however, improvements in computer security have led to the introduction of various high-stakes, large-scale language assessments (e.g., TOEFL iBT, the computer-delivered IELTS, the TOEIC speaking and writing tests, PTE Academic, the internet-based College English Test in China), improving test access for examinees in various geographical locations. Such large-scale, high-stakes CBT programs today employ a variety of measures in combination to ensure test security. As described by Schmidgall and Powers (2017), such measures range from the use of capabilities for secure online transfer and delivery of test data, biometric test-taker identification throughout the test session (e.g., retinal scans), on-site and online monitoring of test takers for the prevention and detection of fraudulent behaviors (e.g., video/audio recording of test sessions, training test center staff), and test data analyses for cheating detection after testing (e.g., voice authentication, plagiarism detection).

In addition to the various security issues for CBT that consume extensive resources of testing agencies described here, the year of 2020 brought an unprecedented challenge to high-stakes language assessments. The outbreak of the COVID-19 pandemic worldwide resulted in the cancellation of numerous test sessions administered at designated test centers, demanding alternative forms of testing that did not rely on test center administrations. Among some solutions that large-scale testing programs have offered are the TOEFL iBT Home Edition (www.ets.org/s/cv/toefl/at-home/) and the IELTS Indicator (www.ieltsindicator.com/). Both tests are designed for at-home testing, along with security measures for test-taker identification and test data transfer and delivery that are equivalent to those used for their respective regular versions. These tests offer secure testing environments by delivering the test content through secure browsers and utilizing AI and human proctoring to monitor test takers during testing. The IELTS Indicator speaking section is conducted live through a video call, which would discourage malpractice. Yet, given that differences in the test administration condition between test-center testing and at-home testing may affect test-taker perception and performance, careful examinations are required into the extent to which scores obtained from the at-home testing versions of these tests could be considered comparable to those obtained from the corresponding test-center-administered versions.

Two other areas that Bennett (2000) noted as features of this phase are the application of multimedia and virtual reality to CBT. As discussed in detail next, language testers have been proceeding rather cautiously to adopt these two capabilities, although relevant studies are indeed emerging. Given their important implications for test validity, these studies will be discussed in more detail later.

Phase 3: Integrating instruction, learning, and assessment

The third and final phase of Bennett's (2000) scenario for the development of CBT is a generation of tests that are radically different from traditional assessments. Here, we are concerned with what truly innovative CBT may look like. A "curriculum-embedded, distance assessment" (p. 14) proposed by Bennett would be administered frequently as an integral part of a curriculum, accommodating the diverse needs of individual learners. Such testing may be available in the form of learning tools with virtual reality simulations for both individual and collaborative work accessed through the internet by learners from various

locations. In other words, this third phase primarily concerns how effectively and meaningfully we use CBT to promote learning while taking advantage of various functionalities available from the first two phases of CBT development.

Although CBT applications illustrative of this phase are scarce in the field so far, an important related area in which applied linguistics research has made great progress in the last decade is dynamic assessment (DA), a method for assessing learning potential based on sociocultural theory (Antón and García, this volume; Poehner, 2008). This line of research is relevant to Bennett's third phase because the computerized version of DA (called C-DA) is portable and can be embedded in a school curriculum and deployed for frequent progress checks for formative decision making. A main goal of DA is to examine the learner's zone of proximal development (ZPD; Vygotsky, 1978), or the difference between the level at which a learner can perform without help and the highest level at which he/she can perform with mediation. In previous C-DA studies conducted by Poehner and Lantolf (2013), Poehner *et al*. (2014), and Teo (2012), learners received pre-programmed feedback based on human teacher behaviors during test administration. Given the difficulty of implementing traditional DA sessions in face-to-face group instruction settings, C-DA would greatly improve the practicality of DA implementation. Also conceptualized within the DA framework are Wolf and her associates' studies (Choi *et al*., 2019; Wolf *et al*., 2016), in which the stimuli for story retell tasks were presented in a series of video segments so that the entire story could be broken into steps in order to provide scaffolding to enhance young English learners' speaking performance.

In summary, early CBT research in the field focused on the increased efficiency and convenience brought by the replacement of existing paper-based testing (PBT) for CBT, while recent research emphasizes the exploration of improvements in assessment design and administration that can be achieved only in CBT. Thus, it is safe to say that the field is now well into Bennet's (2000) second phase of CBT development, although advancements in areas such as using multimedia and virtual reality are relatively slow. Yet many available CBT tests still rely largely on traditional assessment techniques. These could be termed "incremental evolutions" (Chapelle and Douglas, 2006) in that the advancements are still short of what one might call a true innovation that would fundamentally change the ways in which we do testing. However, studies such as those focusing on C-DA applications offer a great opportunity for the field to move toward Bennet's (2000) third phase.

Critical issues and topics

With the advancement of computer technology apace, many new CBT applications incorporating the latest functionalities are likely to continue to be developed in the future. No matter how attractive such CBT applications may look, however, what is critical is the quality of information about examinee language ability that these tests produce. As with any traditional assessment, the validity of CBT should be evaluated systematically in order to ensure justifiable test score interpretation and use. Application examples of test validity frameworks to CBT development and validation are now abundant, ranging from early CBT studies employing Bachman and Palmer's (1996) qualities of test usefulness framework (e.g., Chapelle *et al*., 2003; Chun, 2006, 2008; Downey *et al*., 2008) to more recent studies based on Weir's (2005) socio-cognitive framework in the IELTS context (e.g., Nakatsuhara *et al*., 2017), as well as those conceptualized within argument-based test validation frameworks, such as those of Bachman and Palmer (2010) and Kane (2006; this volume) for the

TOEFL iBT and TOEIC tests (Chapelle *et al*., 2008; Schmidgall, 2017) and specifically for CBT purposes (Chapelle and Douglas, 2006; Chung, 2017).

While CBT quality should be examined from various perspectives, issues that deserve particularly close attention concern the appropriateness of the task design because it directly affects test score interpretation. In Bachman and Palmer's (2010) assessment use argument (AUA) framework, for instance, relevant qualities are (1) meaningfulness (the extent to which the test score reflects the target language ability), (2) generalizability (the degree to which test results can be generalized to language performance in the target language use [TLU] domain), and (3) impartiality (the absence of bias in the test task design). In the context of CBT, Chapelle and Douglas (2006) elaborate specifically on how the meaningfulness and generalizability of test tasks can be enhanced by a thorough domain analysis. Adopting a context-sensitive definition of language ability for communication through technology, Chapelle and Douglas define the abilities required for successful communication in today's digitized society as linguistic competence, technological competence, and strategic competence. Given various technological options in L2 communication now available, the learner must select an appropriate option strategically (e.g., selecting email over social media to send important business documents to a client) and make use of relevant linguistic resources and technological skills to complete a specific language use task. In designing CBTs, such plasticity of the target construct in technology-mediated communication should be considered a key defining feature.

Current contributions and research

Comparability of PBT and CBT

A CBT validity issue that has attracted the attention of researchers, and which is particularly relevant to Bachman and Palmer's (2010) meaningfulness and impartiality, is the comparability of scores obtained across the PBT and CBT modes, a potential test method effect on examinee test performance. When PBT and CBT versions of the same test are used concurrently for decision making, psychometric equivalence of scores across the modes must be established in consideration of test fairness. The same applies to a case in which CAT items are pre-tested in a PBT form and then added to an item bank because the functioning of the items (e.g., IRT difficulty and discrimination parameters) should not change across the modes (Sawaki, 2001).

The CBT versus PBT comparability issue has been examined from multiple perspectives. One is the effects of computer familiarity on examinee test performance in order to ensure that examinees are not disadvantaged by taking CBTs simply because they are not familiar with using or taking tests on computers. Two early studies on this issue conducted in the context of the TOEFL (e.g., Taylor *et al*., 1998; Trites and McGroarty, 2005) showed that the effect of computer familiarity on language test performance is negligible and can be controlled with tutoring as needed. Despite the rapid advancement of computer technology, however, computer familiarity still appears to affect today's examinees who are more used to CBT than those a few decades ago. In a recent study of the writing section of the Internet-based College English Test (IB-CET) conducted in China by Jin and Yan (2017), among study participants who were already familiar with computers, those who were relatively more familiar with computers tended to score higher, produce texts of higher quality with fewer errors, and perceive the composing process more positively than less computer-familiar students.

A second approach is to examine the degree to which test scores obtained in the CBT and PBT conditions are equivalent to each other. Many previous studies addressed this issue by comparing observed group mean scores and by examining the invariance of the psychometric factor structure across the modes, although the latter is more fundamental in ensuring test score equivalence (APA, 1986). Schroeders and Wilhelm's (2011) own confirmatory factor analysis and 11 other previous comparability studies on L2 reading and listening comprehension tests reviewed by them confirmed at least some degree of measurement invariance. This result suggests that, in general, the construct a given test taps into remains unchanged across the modes, despite the statistically significant mean differences across modes found in some studies (e.g., Choi *et al*., 2003; Coniam, 2006).

The issue of test score comparability was also examined with regard to the effects of response mode (handwriting versus word processing) on writing test scores. Although previous findings are not entirely consistent, no difference in holistic essay scores between handwriting and word processing were found when a choice was given to examinees in multiple studies (e.g., Blackhurst, 2005; Breland *et al*., 2004; Wolfe and Manalo, 2004). Similar results were obtained in a recent experimental study that compared analytic scores of essay and integrated writing task scores across the modes (Brunfaut *et al*., 2018), though there was a small negative effect for lower-ability examinees taking an integrated writing task in the computer-based mode. Breland *et al*. (2004) and Wolfe and Manalo (2004) also found that low-ability examinees performed worse when word processing, although no such effect was found for high-ability examinees. Taken together, these findings suggest that the response mode may have differential effects on language ability subgroups.

Chapelle and Voss (2016) point out that CBT versus PBT comparability studies that focus on how well the CBT condition can replicate the "gold standard" (p. 119) represented by the PBT condition are limited because developing a new CBT often aims at designing innovative assessments by incorporating functionalities available only on computers. Indeed, this approach may be most applicable to validation studies in Bennett's (2000) first phase of CBT applications, but such studies would be necessary as long as PBT and CBT co-exist within a single decision-making process (e.g., accepting both PBT and CBT versions of the same test for student admission to an academic program). Even for studies dealing with Bennett's second generation and beyond, this approach is also applicable to conceptualizing studies on comparability across different CBT conditions (e.g., Ling, 2017) and comparability across CBTs and PBTs based on different test specifications in certain contexts (e.g., Kunnan and Carr, 2017). The same logic also applies to the investigation into the comparability of the at-home testing versions of high-stakes tests designed to address the COVID-19 pandemic situation described earlier against their test-center-administered counterparts. Future comparability studies should focus more on areas that previous studies had less emphasis on. One is the task change across the modes, which can result in qualitative changes in the test score interpretation, even if the test score is equivalent. In this regard, more studies comparing the test-taking process across the modes such as those by Jin and Yan (2017) and Yu (2010) would be required. Such studies could also contribute to explaining the differential impacts of presentation mode effects on subgroups (Breland *et al*., 2004; Wolfe and Manalo, 2004) and on individual examinees (Yu, 2010). Another is learner preference. As for writing, for instance, Brunfaut *et al*. (2018) reported examinee's preference of CBT over PBT, and Ling (2017) reported the preference for local keyboards as opposed to the US keyboard reported by about 40 percent of study participants from 24 countries in the TOEFL iBT population. Taking account of such examinee perspectives could lead to improvements in test design.

The role of multimedia and virtual reality in language assessment

The employment of multimedia and virtual reality in CBT task design is relevant to Bachman and Palmer's (2010) meaningfulness and generalizability of test scores. Employing these technologies is likely to lead to innovative task designs in appearance, although the value of doing so should be evaluated carefully in relation to how they impact test validity.

An area of multimedia application that has drawn the attention of language testers is the use of visuals in listening assessment. This is because doing so allows better emulations of many real-life listening tasks involving visual information (e.g., listening to academic lectures in the classroom). Chapters in Ockey and Wagner's (2018) volume provide extensive discussions on the effects of test-taker interaction with visuals presented in listening tasks on the L2 listening construct (see also Wagner, this volume).

Concerning the use of virtual reality in language assessment, two recent studies report its promising applications that contribute to enhancing the generalizability of test results to TLU domain performance (Bachman and Palmer, 2010). First, in an English for specific purposes (ESP) context, Park (2018) designed an assessment of aviation English for air traffic control (ATC) communication. Based on a needs analysis with experienced air traffic controllers, Park designed a series of test tasks simulating conversations between ATC personnel and pilots (e.g., changing flight plan, procedure for departure and arrival) in a virtual environment. A pilot test with ten air traffic controllers showed that these tasks elicited cognitive, metacognitive, and communication strategies involved in the corresponding TLU tasks. Second, Ockey *et al*.'s (2017) study demonstrated how virtual reality could contribute to replicating the context in which interactive oral communication takes place. In Ockey *et al*.'s study, 18 participants in remote locations "met" in groups of 3 in a virtual environment. During the discussion, individual participants interacted with one another's avatars. Ockey *et al*. elaborate on various design considerations that went into the development of the virtual environment, which would inform future virtual reality applications to language assessment.

How best to utilize computer technology for assessing speaking

In early stages of CBT development, assessing speaking on computers was considered particularly problematic (Chalhoub-Deville and Deville, 1999), primarily due to technical challenges in replicating an important feature of real-life human interactions: the unpredictable and reciprocal nature of the discourse. Thus, the application of virtual reality to speaking assessment such as Ockey *et al*.'s (2017) study described earlier is considered a major improvement from the perspective of Bachman and Palmer's (2010) meaningfulness and generalizability of test scores. Forms of CBT for speaking assessment introduced to the field early on have typically been non-reciprocal in nature. In other words, a test taker's response to a test question is not followed by another turn by the interviewer (or computer) to generate an interactive discourse. One such form is the simulated interview, in which the examinee responds to a series of computer-delivered questions that a human interviewer would ask in a live, face-to-face interview. Examples of this type of speaking assessment include simulated versions of the ACTFL Oral Proficiency Interview, the COPI (Kenyon and Malabonga, 2001) described earlier, as well as the Oral Proficiency Interview-computer, or OPIc (www.languagetesting.com/oral-proficiency-interview-by-computer-opic), a computer-based speaking assessment in which an avatar figure plays the role of the interviewer to pose questions to the test taker. Another early form of CBT speaking test is

automated speaking tests such as the Versant speaking test, a fully automated speaking test that delivers a series of short items to test psycholinguistic processes that underly contextualized speaking performance (e.g., reading aloud and repeating sentences, responding to short-answer questions) (Downey *et al*., 2008; Berstein *et al*., 2010).

These existing CBT applications have certainly improved test access and enabled the provision of standardized test administration conditions unaffected by interviewer variations inherent in face-to-face testing (see Davis, this volume). However, a challenge the field has been grappling with is how to more fully assess interactional competence: namely, "the ability to co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event" (Galaczi and Taylor, 2018: 226). A potential avenue in this direction that has been gaining the attention of researchers recently is to employ speech technologies that could emulate reciprocal human communication. For instance, Litman *et al*. (2018) describe how an automatic speech recognizer (ASR) and spoken dialogue systems (SDS) could be used in tandem to decode a test-taker utterance, analyze its form and meaning, and generate an appropriate response automatically. As noted by Litman *et al*., previous research employing this approach focused mainly on simulating relatively short and constrained routine conversations for language learning rather than language assessment purposes. While extensive further research is required to model more complex and open-ended interactions that are typically assessed in L2 speaking tests by using ASR and SDS, advancement of this approach would broaden the construct that automated assessments could tap into considerably.

Another recent development worth noting is the employment of video-conferencing programs for assessing speaking (Nakatsuhara *et al*., 2017; Ockey *et al*., 2019). Recently, Nakatsuhara *et al*. piloted a CBT form of the IELTS speaking section with 32 participants by employing Zoom (www.zoom.us). Their mixed-methods study results showed that the participants' test scores were comparable between the standard face-to-face interview and video-conferencing conditions. However, Nakatsuhara *et al*. also identified some important mode differences, such as the increase of clarification questions asked by the participants, effects of voice quality on interlocutor behaviors (e.g., slower speech, clearer articulation), and the difficulty of rating grammar and pronunciation on the part of the examiners in the video-conferencing condition. Meanwhile, Ockey *et al*. (2019) employed Skype (www.skype.com) to test 72 and 74 candidates in the US and China, respectively, experiencing various technological problems, particularly in China. Despite such challenges, however, the participants generally perceived the video-conferencing condition favorably.

Overall, these results suggest that the technology is now mature enough for us to consider some application options for assessing oral interaction in CBT. Nakatsuhara *et al*. concluded that video-conferencing could be a "parallel alternative" (p. 15) to the face-to-face condition for the IELTS test, while Ockey *et al*. (2019) warn us of the challenge of stable large-scale international test administration with this technology. A point worth noting is that Ockey *et al*. (2017) reported better technological stability of their virtual environment program than video-conferencing used in previous studies, which suggest the need for further explorations of these two technologies for CBT applications. A more fundamental issue, however, is that these studies also remind us of the difficulty of emulating face-to-face, real-time communication in CBT conditions. While many important qualities of interactional competence could be assessed by using these new technologies, these studies point to various validity issues such as the change in examinee and interlocutor behaviors due to compromised voice quality and slow data transmission in the video-conferencing condition (Nakatsuhara *et al*.,

2017) and the challenge in designing avatars that can naturally replicate human nonverbal communication in a virtual environment (Ockey *et al*., 2017). As rightly pointed out by Plough *et al*. (2018), assessing candidates in such CBT conditions may force us to adopt a narrow interpretation of interactional competence. Further validity investigations in related topics are thus in order.

Recommendations for practice

Given the rapidly developing nature of computer technology, it may be tempting for the test designer to focus on the appearance of a test or to constantly try to use new functionalities. As noted earlier, however, what matters is ensuring the quality of information about the learner's L2 ability that the CBT offers. Any language test should go through a rigorous, systematic investigation of its validity, and a test designed for CBT is no exception. Toward this end, a CBT should be conceptualized within a test validation framework, and validity evidence to support the test design and use should be accumulated throughout the test's lifetime (Bachman and Palmer, 2010). It is recommended that the test designers refer to previous CBT validation studies as well as applications of frameworks for evaluating L2 test validity and use, such as those reviewed in this chapter. In particular, Fulcher (2003) argues that often little attention is paid to various interface issues even though they have important implications for test quality. A careful examination of various issues raised would help the test developer alleviate a situation in which intended test score interpretation and use are undermined due to suboptimal test design decisions.

Future directions

This discussion shows that while notable incremental advances in CBT have been made in the field of language assessment, there is still a lot more work to be done to pave the way for developing truly innovative CBT. Various tests for CBT have been designed and developed, and documentation on these efforts as well as validation studies are accumulating. However, there is room for a great deal of more empirical research, including studies based on operational data, as well as long-term investigations into how CBT might impact individuals, society, and language instruction.

In the future, the field should proceed along three lines of investigation. First, more studies are required to enhance construct representation. One approach is to explicate the nature of the L2 construct in today's digitized society clearly, along the lines of Chapelle and Douglas's (2006) three-part definition of the construct. Another is to continue exploring how innovative functionalities available in CBT, such as multimedia and virtual reality, could be utilized for designing test tasks that are reflective of the characteristics of real-life language use tasks. Yet another approach is to find ways to assess the target constructs in a more comprehensive manner. For instance, latency of learner performance in reading was mentioned by Alderson (2000a) as an area that can readily be assessed with computer technology, although CBT applications that are designed to assess broader constructs such as this are relatively scarce.

Second, in order for language assessment research to penetrate into Bennet's (2000) third stage of CBT development, we need to actively seek ways to better link assessment with instruction (Alderson, 2005; Jang and Sinclair, this volume). The advancement made recently in C-DA reviewed earlier is promising, while other possibilities should also be explored. For instance, it is worth examining how approaches to providing learners with

performance feedback for diagnosis and instruction, such as cognitive diagnosis (Lee and Sawaki, 2009) and learning oriented assessment (LOA; Turner and Purpura, 2016; Fox *et al*., this volume), could be integrated into CBT design for continuous learner assessment that is blended into instruction and tailored to individual learners' needs.

Finally, making advancements in these areas requires a considerable amount of substantive work. For instance, as Chapelle and Douglas (2006) suggested, collaboration of language assessment and second language acquisition researchers might help us design tests for CBT that are sensitive to how L2 ability develops in the long term. Furthermore, working with measurement specialists would be beneficial in developing systems for measuring, compiling, and analyzing data to further our understanding of L2 development. Although there are a lot of challenges ahead, focused efforts to enhance our substantive and technical knowledge would help us take a step forward in designing truly innovative tests for CBT that transform the conventions of L2 assessment.

Further reading

Chapelle, C. A. and Douglas, D. (2006). *Assessing Language Through Computer Technology*. Cambridge, UK: Cambridge University Press.

This Cambridge Language Assessment Series volume offers an excellent introduction to various aspects of CBT for students and professionals who are new to the topic. The authors start their discussion by identifying notable features of CBT as well as similarities and differences of CBT against conventional assessments. Various issues raised in chapters on the threat of CBT, evaluating CBT, and the impact of CBT are particularly eye opening. The authors' proposal of a context-sensitive definition of language ability for communication through technology marks an important step forward in enhancing our understanding of L2 abilities required to communicate successfully in technology-mediated environments.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.

Chapelle, Enright, and Jamieson offer an overview of the history of the TOEFL as well as a comprehensive account of how the TOEFL iBT was designed and developed. Discussions on rationales behind various aspects of the test design and research studies conducted as part of the test development process and for initial validation of the test are valuable for the reader to understand the complexity and extensiveness of the work involved in the development of the high-volume, high-stakes CBT over the years. In addition, the TOEFL validity argument framework presented in the final chapter illustrates how an argument-based approach can be applied to CBT validation.

Chapelle, C. A. and Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology* 20: 116–28.

This article is a review of papers and book reviews on CBT for language assessment published in the *Language Learning & Technology* journal between 1997 and 2015. Consistent with Bennet's (2000) three phases of CBT development adopted in this chapter, Chapelle and Voss discuss the studies identified according to two categories, technology for efficiency and technology for innovation. Based on the results, Chapelle and Voss note how deeply technology and language assessment are ingrained into language learning and emphasize the importance of teacher training for effective integration of technology and language assessment into the language classroom.

Litman, D., Strik, H. and Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly* 15: 294–309. https://doi.org/10.1080/15434303.2018.1472265.

This paper is in the *Language Assessment Quarterly* special issue, conceptualizing and operationalizing speaking assessment for a new century (guest edited by G. S. Lim), shows what automated speaking assessment might look like in the near future. The authors provide a non-technical explanation of

how an automatic speech recognizer (ASR), which serves as the basis of automated speech scoring, could be combined with spoken dialogue systems (SDS) to emulate spoken interaction. The authors' discussion of the current state, challenges, and potentials of this new technology for L2 speaking assessment application is insightful, raising various important issues of further consideration concerning task design and construct representation in automated L2 speaking assessment.

References

Alderson, J. C. (2000a). *Assessing Reading*. Cambridge, UK: Cambridge University Press.

———. (2000b). Technology in testing: The present and the future. *System* 28: 593–603. https://doi.org/10.1016/S0346-251X(00)00040-3.

———. (2005). *Diagnosing Foreign Language Proficiency: The Interface Between Learning and Assessment*. New York, NY: Continuum.

Alderson, J. C. and Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the common European framework. *Language Testing* 22: 301–20. https://doi.org/10.1191/0265532205lt310oa.

American Psychological Association. (1986). *Guidelines for Computer-Based Tests and Interpretations*. Washington, DC: Author.

Bachman, L. F. and Palmer, A. (1996). *Language Testing in Practice*. Cambridge, UK: Cambridge University Press.

———. (2010). *Language Assessment in Practice*. Cambridge, UK: Cambridge University Press.

Bennett, R. E. (2000). *Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing*. Princeton, NJ: ETS.

Bernstein, J., van Moere, A. and Cheng, J. (2010). Validating automated speaking tests. *Language Testing* 27: 355–77. https://doi.org/10.1177/0265532210364404.

Blackhurst, A. (2005). Listening, reading, and writing on computer-based and paper-based versions of IELTS. University of Cambridge ESOL Examinations. *Research Notes* 21: 14–17.

Breland, H., Lee, Y.-W. and Muraki, E. (2004). *Comparability of TOEFL CBT Writing Prompts: Response Mode Analyses* (TOEFL Research Report No. RR-75). Princeton, NJ: ETS. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01950.x.

Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology* 1: 44–59. https://doi.org/10125/25003.

Brunfaut, T., Harding, L. and Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing* 3: 3–18. https://doi.org/10.1016/j.asw.2018.02.003.

Chalhoub-Deville, M. (ed.). (1999). *Issues in Computer-adaptive Testing of Reading Proficiency*. Cambridge, UK: Cambridge University Press.

Chalhoub-Deville, M. and Deville, C. (1999). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics* 19: 273–99. https://doi.org/10.1017/S0267190599190147.

Chapelle, C. A. and Douglas, D. (2006). *Assessing Language Through Computer Technology* Cambridge, UK: Cambridge University Press.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A., Jamieson, J. and Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing* 20: 409–39. https://doi.org/10.1191/0265532203lt266oa.

Chapelle, C. A. and Voss, E. (2016). 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology* 20: 116–28. https://doi.org/10125/44464.

Choi, I.-C., Kim, K. S. and Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing* 20: 295–320. https://doi.org/10.1191/0265532203lt258oa.

Choi, I.-C., Wolf, M. K., Pooler, M., Sova, L. and Faulkner-Bond, M. (2019). Investigating the benefits of scaffolding in assessments of young English learners: A case for scaffolded retell tasks. *Language Assessment Quarterly* 16: 161–79. https://doi.org/10.1080/15434303.2019.1619180.

Chun, C. W. (2006). An analysis of a language test for employment: The authenticity of the PhonePass test. *Language Assessment Quarterly* 3: 295–306. https://doi.org/10.1207/s15434311laq0303_4.

———. (2008). Comments on 'Evaluation of the usefulness of the versant for English test: A response': The author responds. *Language Assessment Quarterly* 5: 168–72. https://doi.org/10.1080/15434300801934751.

Chung, Y.-R. (2017). Validation of technology-assisted language tests. In C. A. Chapelle and S. Sauro (eds.), *The Handbook of Technology and Second Language Teaching and Learning*. New York, NY: Wiley, 332–47.

Coniam, D. (2006). Evaluating computer-based and paper-based versions of an English language listening test. *ReCALL* 18: 193–211. https://doi.org/10.1017/S0958344006000425.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning Teaching and Assessment*. Cambridge, UK: Cambridge University Press.

Douglas, D. (2013). Technology and language testing. In C. Chapelle (ed.), *Encyclopedia of Applied Linguistics*. New York, NY: Wiley. https://doi.org/10.1002/9781405198431.wbeal1182.

Douglas, D. and Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics* 27: 115–32. https://doi.org/10.1017/S0267190508070062.

Downey, R., Farhady, H., Present-Thomas, R., Suzuki, M. and van Moere, A. (2008). Evaluation of the usefulness of the versant for English test: A response. *Language Assessment Quarterly* 5: 160–67. https://doi.org/10.1080/15434300801934744.

Fulcher, G. (1999). Computerizing an English language placement test. *ELT Journal* 53: 289–98.

———. (2003). Interface design in computer-based language testing. *Language Testing* 20: 384–408. https://doi.org/10.1191/0265532203lt265oa.

Galaczi, E. and Taylor, L. (2018). Interactional Competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly* 15: 219–36. https://doi.org/10.1080/15434303.2018.1453816.

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics* 25: 228–42. https://doi.org/10.1017/S0267190505000127.

Jamieson, J., Eignor, D., Grabe, W. and Kunnan, A. J. (2008). Frameworks for a new TOEFL. In C. Chapelle, M. K. Enright and J. Jamieson (eds.), *Building a Validity Argument for the Test of English as a Foreign Language™*. New York, NY: Routledge.

Jin, Y. and Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly* 14: 101–19. https://doi.org/10.1080/15434303.2016.1261293.

Kane, M. (2006). Validation. In R. Brennan (ed.), *Educational Measurement*, 2nd edn. Westport, CT: American Council on Education and Praeger Publishers.

Kenyon, D. M. and Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning Technology* 5: 60–83. https://doi.org/10125/25128.

Kunnan, A. and Carr, N. (2017). A comparability study between the general English proficiency test-advanced and the internet-based test of English as a Foreign language. *Language Testing in Asia* 7: 17. https://doi.org/10.1186/s40468-017-0048-x.

Lee, Y.-J. (2002). A Comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing* 8: 135–57. https://doi.org/10.1016/S1075-2935(03)00003-5.

Lee, Y.-W. and Sawaki, Y. (2009). Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly* 6: 239–63. https://doi.org/10.1080/15434300903079562.

Ling, G. (2017). Is writing performance related to keyboard type? An investigation from examinees' perspectives on the TOEFL iBT. *Language Assessment Quarterly* 14: 36–53. https://doi.org/10.1080/15434303.2016.1262376.

Litman, D., Strik, H. and Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly* 13: 294–309. https://doi.org/10.1080/15434303.2018.1472265.

Malabonga, V., Kenyon, D. M. and Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing* 22: 59–92. https://doi.org/10.1191/0265532205lt297oa.

Milton, J. (2013). Second language acquisition via Second Life. In C. Chapelle (ed.), *Encyclopedia of Applied Linguistics*. New York, NY: Wiley. https://doi.org/10.1002/9781405198431.wbeal1318.

Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly* 14: 1–18. https://doi.org/10.1080/15434303.2016.1263637.

Nogami, Y. and Hayashi, N. (2010). A Japanese adaptive test of English as a foreign language: Developmental and operational aspects. In W. van der Linden and C. A. W. Glas (eds.), *Elements of Adaptive Testing*. New York, NY: Springer.

Norris, J. M. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning and Technology* 5: 99–105. https://doi.org/10125/25131.

Ockey, G. J., Gu, L. and Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly* 14: 346–59. https://doi.org/10.1080/15434303.2017.1400036.

Ockey, G. J., Timple-Loughlin, V., Davis, L. and Gu, L. (2019). *Exploring the Potential of a Video-Mediated Interactive Speaking Assessment*. (ETS Research Report No. ETS RR-19–05). Princeton, NJ: ETS. https://doi.org/10.1002/ets2.12240.

Ockey, G. J. and Wagner, E. (2018). *Assessing L2 Listening: Moving Towards Authenticity*. Amsterdam, Netherlands: John Benjamins.

Park, M. (2018). Innovative assessment of aviation English in a virtual world: Windows into cognitive and metacognitive strategies. *ReCALL* 30: 196–213. https://doi.org/10.1017/S0958344017000362.

Plough, I., Banerjee, J. and Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing* 35: 427–45. https://doi.org/10.1177/0265532218772325.

Poehner, M. E. (2008). *Dynamic Assessment: A Vygotskian Approach to Understanding and Promoting Second Language Development*. Berlin: Springer.

Poehner, M. E. and Lantolf, J. P. (2013). Bridging the ZPD into the equation: Capturing L2 development during computerized dynamic assessment (C-DA). *Language Teaching Research* 17: 323–42. https://doi.org/10.1177/1362168813482935.

Poehner, M. E., Zhang, J. and Lu, X. (2014). Computerized dynamic assessment (C-DA): Diagnosing L2 development according to learner responsiveness to mediation. *Language Testing* 32: 337–57. https://doi.org/10.1177/0265532214560390.

Roever, C. (2001). Web-based language testing. *Language Learning and Technology* 5: 84–94. https://doi.org/10125/25129.

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning and Technology* 5: 38–59. https://doi.org/10125/25127.

Schmidgall, J. E. (2017). *Articulating and Evaluating Validity Arguments for the TOEIC® Tests*. (ETS Research Report No. ETS RR-17–51). Princeton, NJ: ETS. https://doi.org/10.1002/ets2.12182.

Schmidgall, J. E. and Powers, D. E. (2017). Technology and high-stakes language testing. In C. A. Chapelle and S. Sauro (eds.), *The Handbook of Technology and Second Language Teaching and Learning*. New York, NY: Wiley, 317–31.

Schroeders, U. and Wilhelm, O. (2011). Equivalence of reading and listening comprehension across test media. *Educational and Psychological Measurement* 71: 849–69. https://doi.org/10.1177/0013164410391468.

Stansfield, C. W. (ed.). (1986). *Technology and Language Testing*. Washington, DC: TESOL.

Taylor, C., Jamieson, J., Eignor, D. and Kirsch, I. (1998). *The Relationship Between Computer Familiarity and Performance on Computer-Based TOEFL Test Tasks* (TOEFL Research Report No. 61). Princeton, NJ: ETS. http://dx.doi.org/10.1002/j.2333-8504.1998.tb01757.x.

Teo, A. (2012). Promoting EFL students' inferential reading skills through computerized dynamic assessment. *Language Learning & Technology* 16: 10–20. https://doi.org/10125/44292.

Trites, L. and McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing* 22: 174–210. https://doi.org/10.1191/0265532205lt299oa.

Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Baneerjee (eds.), *Handbook of Second Language Assessment*. Berlin: De Gruyter, 255–73.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke, UK: Palgrave Macmillan.

Wolf, M. K., Guzman-Orth, D., Lopez, A., Castellano, K., Himelfarb, I. and Tsutagawa, F. S. (2016). Integrating scaffolding strategies into technology enhanced Assessments of English learners: Task types and measurement models. *Educational Assessment* 21: 157–75. https://doi.org/10.1080/10627197.2016.1202107.

Wolfe, E. W. and Manalo, J. R. (2004). Composition medium comparability in a direct writing assessment of non-native English speakers. *Language Learning & Technology* 8: 53–65. http://doi.org/10125/25229.

Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly* 7: 119–36. https://doi.org/10.1080/15434300903452355.

32

Corpus linguistics and language testing

Sara T. Cushing

Introduction/definitions

The index to the 2012 edition of the *Routledge Handbook on Language Testing* had exactly one entry under *corpus*: "corpus-based word frequency analysis" (p. 511), found in Moder and Halleck's (2012) chapter on assessing language for specific purposes. While corpus linguistics has had a much longer history in language testing than this index entry might imply, it is striking to see how much influence the field of corpus linguistics has had on the language testing field and how this influence has expanded in the past twenty years or so. For example, a mere five years following the publication of the first edition of the *Handbook*, a special issue of *Language Testing* (Cushing, 2017) was devoted to corpus linguistics and language testing, with five empirical articles illustrating the use of corpora to address research questions in language testing and two commentaries. In addition, a recent survey of three major relevant journals – *Language Testing, Language Assessment Quarterly*, and *Assessing Writing* – (Plakans, 2018) identifies corpus linguistics as one of two major subthemes in the "deluge" (p. 6) of articles related to the use of technology in language testing published between 2007 and 2017.

Clearly, corpus linguistics has attracted the attention of researchers and practitioners in language testing – but what is corpus linguistics, exactly? A corpus (plural: corpora) is "a text collection which is large, computer-readable, and designed for linguistic analysis" and corpus linguistics can be defined as "the use of computer-assisted methods to study large quantities of real language" (Stubbs and Hale, 2012: 1). Corpus linguistics is thus an approach to the study of language that primarily involves analysis of the frequency of certain linguistic phenomena (morphemes, words, grammatical patterns) in specific settings (Gries, 2009). This definition has clear resonances with language testing/assessment, which has been defined as "the systematic gathering of language-related behavior in order to make inferences about language ability and capacity for language use on other occasions" (Chapelle and Plakans, 2013: 241). The increased availability of large corpora that are publicly available, along with free or low-cost analysis tools over the past two decades, has made this "systematic gathering of language-related behavior" more widely accessible to researchers and test developers alike.

DOI: 10.4324/9781003220756-42

Corpora can consist of either oral or written data and can be categorized into several types. *Reference corpora* are intended to represent a language or language variety as a whole: examples are the Corpus of Contemporary American (COCA; Davies, 2008); the British National Corpus (BNC; Leech *et al*., 1994) and its successor, the BNC2014 (Love *et al*., 2017); and the Corpus del Español Actual (CEA; Subirats and Ortega, 2012). *Specialized corpora* are compiled to represent a specific language use context, such as academic language, the language of medical practice, and so on. Examples of specialized corpora include the Michigan Corpus of Advanced Spoken English (MICASE; Simpson *et al*., 2002), the British Academic Spoken English Corpus developed at the University of Warwick (see Nesi, 2001), and the Call Center corpus, collected by Friginal (2009) to represent typical customer service interactions from a call center in the Philippines. Multilingual corpora contain equal amounts of text from several different languages, typically in the same genre, but not direct translations from one language to the others, while a parallel corpus consists of texts in two or more languages that are exact equivalents, frequently used to study translation strategies (Baker, 2010).

Of particular interest to scholars in language education and assessment are *learner corpora*, which consist of language elicited from language learners, typically, though not always, under standardized conditions such as large-scale tests. Even though data from learner corpora are usually cross-sectional rather than longitudinal, learner corpora are useful for investigating typical language use at different levels of proficiency or by speakers from different first languages. Well-known learner corpora include the International Corpus of Learner English (ICLE; Granger *et al*., 2009) and the Cambridge Learner Corpus (now part of the Cambridge International Corpus), begun by Cambridge University Press and the EFL Division of the University of Cambridge Local Examinations Syndicate (UCLES EFL) in the early 1990s (Barker, 2010).

In the rest of this chapter, I provide a brief introduction to essential concepts in corpus linguistics, followed by a review of how corpus linguistics has been used in the development, validation, and scoring of language tests, along with a survey of methods and tools used to analyze language corpora. The chapter closes with a discussion of future directions in using corpora for test development and validation.

Historical perspectives

Before the advent of the technological innovations that made corpus linguistics possible, lexicographers and other scholars interested in language use (as opposed to theoretical linguists concerned mainly with idealized language competence) had to rely on their own meticulously collected examples of authentic language to draw general conclusions about language. These examples were frequently recorded on small slips of paper and stored in boxes or desk drawers (Svartvik, 2011), clearly creating challenges for large-scale analysis. Barker (2006) notes that, in the 1970s and 1980s, large-scale databases started to be used by second language researchers but were not commonly used to inform large-scale testing projects at that time. One of the earliest large-scale corpora compiled in English was the Standard Corpus of Present-Day Edited American English, called the Brown Corpus as the authors were from Brown University (see Francis and Kucera, 1979). Work on the British National Corpus was begun in 1991 and was made public in 1994 (Leech *et al*., 1994).

Beginning in the 1990s, examination boards, test publishers, and universities started developing electronic collections of learner test data, particularly writing, along with scores and demographic information (Barker, 2006); the Cambridge Learner Corpus, as noted

earlier, was one such corpus. A decade later, across the Atlantic, the Educational Testing Service (ETS) commissioned a corpus of academic language, the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) to inform a major revision of the TOEFL exam (Biber *et al*., 2004).

In the past three decades, more and more corpora have been assembled, both general and specialized, in numerous languages. Equally importantly, an increasing number of computational tools have been developed that allow for more sophisticated parsing, tagging, and analyses. The availability of both large corpora and analysis tools have begun to bring corpus methods to the fore in language testing.

One of the first scholars to predict the usefulness of corpus linguistics to language testing was Alderson (1996), who discussed prospective applications of corpus linguistics to test construction and item development. Around the same time, Biber *et al*. (1998), in an influential volume that served to introduce corpus linguistics to applied linguists, included language testing as an area that could benefit from insights from corpus linguistics. One of the earliest published articles that used corpus data for language testing was Coniam (1997), who proposed a procedure for automatically generating cloze tests using corpus-based word frequency data. Since that time, interest in using corpus data to inform test development and validation has increased substantially, and it would be almost unthinkable now to develop a large-scale language test without referring to corpus data. Recent syntheses of applications of corpus linguistics to language testing include Park (2014), Barker (2006, 2010, 2013), Weigle and Goodwin (2016), and Gablasova (2020). These articles provide good overviews of the ways in which corpora can be used for test development and validation.

Critical issues and topics

In this section, I present a brief overview of how corpora are built and annotated, followed by the main methods of analysis used by corpus linguists.

Corpus development and annotation

A useful discussion of corpus methods in linguistics is provided by Baker (2010). Baker reviews the process of building and annotating (tagging) a corpus and provides an introduction to corpus analysis methods.

While any collection of texts could be called a corpus, McEnery and Wilson (1996: 22) clarify that the collection should be "maximally representative of the variety under examination," be in "machine readable" form, and constitute "a standard reference for the language variety it represents." Thus, as Baker (2010) explains, a corpus should be large enough to reveal patterns of frequency of linguistic phenomena and to ensure that relatively rare phenomena are included. There is no universal rule of thumb for how large a corpus should be; rather, corpus size should be determined in part by the particular phenomenon under investigation. Prosody in speech might be studied with a corpus of 100,000 words (Kennedy, 1998, cited in Baker, 2010); grammatical studies might require a million words (Biber, 1993, cited in Baker, 2010), but a corpus intended to serve as a standard reference for a language variety should be much larger; for example, the BNC contains 100 million words. The Oklahoma aviation corpus produced by Moder and Halleck (2012; see also Moder and Halleck, this volume) consists of 500,000 words and was used as a reference to check the relevance of test items in the development of the Versant Aviation English Test (Van Moere *et al*., 2009).

Three essential considerations in building a corpus are sampling, balance, and representativeness (Baker, 2010). In order to make a corpus representative of the language, language variety, or – to use terminology familiar to language testing specialists – target language use (TLU) situation (Bachman and Palmer, 1996), texts must be selected, sampled, and balanced to avoid over-representation of any one speaker, writer, or topic. For example, in a corpus of academic lectures, to achieve a certain number of words, one might select ten-minute segments from a variety of different lectures by different speakers rather than a few longer lectures. However, these segments should be sampled from different parts of the lectures so that, for instance, the corpus is not skewed toward lecture introductions. Similar care must be taken in compiling a corpus of learner data to ensure that the corpus is representative of the population of learners in terms of such characteristics as first language, proficiency level, and age.

Compiling a corpus of written texts is fairly straightforward, though texts that are not originally in electronic format must be scanned or keyed in, and all corpus files are typically saved as text files. (See Reppen and Simpson, 2002, for further details.) Corpora of spoken language, on the other hand, introduce additional challenges, as spoken language must be transcribed, which involves a series of decisions such as the amount of detail to be included, the inclusion of non- or paralinguistic features such as pauses, laughter, and the like. (See Gablasova *et al*., 2019, for further discussion of these issues.)

Once the corpus is assembled, it is frequently annotated (tagged) for a variety of features, such as parts of speech (POS), grammatical information, and so on. Individual texts may be tagged for author, genre, or information such as proficiency level in the case of a learner corpus so that comparisons may be easily made. Some tagging can be done automatically by various computer programs, but hand checking is always recommended. As an example, Figure 32.1 shows the output of the CLAWS part of speech tagger (Garside, 1987) based on the opening sentence of Jane Austen's novel *Pride and Prejudice* (Austen, 1994). The tags contain specific lexical and grammatical information about words; some of them are fairly transparent (e.g., PRP indicates a preposition and NN1 represents a singular count noun), while others are somewhat more obscure. VVN (as in "acknowledged") refers to the past participle of a lexical verb, while VBZ refers specifically to the verb "be" in third-person singular: i.e., "is." The complete set of tags used here can be found at http://ucrel.lancs.ac.uk/claws/.

POS information can be very useful, for example, in examining word combinations such as preposition + noun combinations. While POS tagging is perhaps the most basic way to tag a corpus, other options for tagging exist as well. For example, the UCREL Semantic Analysis System (USAS; Rayson *et al*., 2004) provides semantic tags, based originally on 21 major discourse categories from the *Longman Lexicon of Contemporary English* (McArthur, 1981). The same sentence from Austen has been tagged in Figure 32.2. The letter in each tag refers to an overall category, and the numbers refer to more specific subcategories. For example, category A refers to general and abstract terms; A3 (is, be) refers to being, A5 (truth, good) refers to evaluation, and A9 (in, possession, of) refers to possession. The complete tag set is found at http://ucrel.lancs.ac.uk/usas/.

```
It_PNP is_VBZ a_AT0 truth_NN1 universally_AV0 acknowledged_VVN ,_PUN that_CJT
a_AT0 single_AJ0 man_NN1 in_PRP possession_PRP of_PRP a_AT0 good_AJ0
fortune_NN1 ,_PUN must_VM0 be_VBI in_PRP want_NN1 of_PRF a_AT0 wife_SENT ._PUN
```

Figure 32.1 Output of CLAWS tagger

```
It_Z8 is_A3+ a_Z5 truth_A5.2+ universally_W3 acknowledged_Q2.2 ,_PUNC that_Z5 a_Z5
single_N5--- man_S2.2m in_A9+[i1.3.1 possession_A9+[i1.3.2 of_A9+[i1.3.3 a_Z5 good_A5.1+
fortune_I1.1+++ ,_PUNC must_S6+ be_A3+ in_Z5 want_X7+ of_Z5 a_Z5 wife_S4f ._PUNC
```

Figure 32.2 Output of USAS tagger

In learner corpora used in SLA research, error tagging (i.e., tagging of incorrect language forms) is frequently desirable since a salient feature of learner language is non-native-like production. However, there is some controversy around both the accuracy of the tagging process and the usefulness of any resulting analysis. As Lüdeling *et al*. (2005: 2) state, "all error annotation implies an interpretation on the part of an annotator"; that is, error annotation can never be completely objective. Furthermore, a persistent issue in making generalizations about errors is that error typologies seem to vary from researcher to researcher (Díaz-Negrillo and Domínguez, 2006). In an early article on computer-aided error analysis of learner corpora, Dagneaux *et al*. (1998: 164) point out following five limitations of error analysis (EA) in general:

1 Limitation 1: EA is based on heterogeneous learner data;
2 Limitation 2: EA categories are fuzzy;
3 Limitation 3: EA cannot cater for phenomena such as avoidance;
4 Limitation 4: EA is restricted to what the learner cannot do;
5 Limitation 5: EA gives a static picture of L2 learning.

Error tagging of a corpus is not yet automatized fully; more commonly, annotators use an editor program that facilitates insertion of error tags according to a predetermined scheme. Error tagging is challenging to accomplish automatically, as even humans do not always agree on what constitutes an error. Furthermore, what may be an error in one variety of a language may be considered acceptable in another variety. Still, tags for errors are particularly valuable for language testing as certain error types may be associated with specific proficiency levels. For example, Thewissen (2013), in an analysis of essays from the ICLE, found 22 error types that appeared with significantly different frequencies in at least one pair of adjacent CEFR levels from B1 to C2, with the majority distinguishing between B1 and higher levels. Only one error type (lexical errors, including prepositions) distinguished between all four levels. Lexical errors were also reported to be a predominant distinguishing feature by Hawkins and Filipović (2012), examining error types in the Cambridge Learner Corpus.

Finally, sentence parsing may be valuable for certain analyses, particularly for investigating syntactic phenomena such as complex sentence formation. One popular sentence parser is the Stanford parser; see Figure 32.3 for a sample, using the same sentence as in the previous examples. The parse can be compared to a syntax tree that indicates hierarchically ordered sentence constituents within parentheses. For example, Figure 32.3 shows that there is a clause (S) embedded within the main clause consisting of a noun phrase (NP) "a single man in possession of a good fortune" and a verb phrase (VP) "must be in want of a wife." Note that the parser uses similar POS tags to the CLAWS tagger discussed earlier.

```
Parse

(ROOT
  (S
    (NP (PRP It))
    (VP (VBZ is)
      (NP
        (NP (DT a) (NN truth))
        (VP
          (ADVP (RB universally))
          (VBN acknowledged) (, ,)
          (SBAR (IN that)
            (S
              (NP
                (NP (DT a) (JJ single) (NN man))
                (PP (In in)
                  (NP
                    (NP (NN possession))
                    (PP (IN of)
                      (NP (DT a) (JJ good) (NN fortune))))))
              (VP (MD must)
                (VP (VB be)
                  (PP (IN in)
                    (S
                      (VP (VB want)
                        (PP (IN of)
                          (NP (DT a) (NN wife)))))))))))))
    (. .)))
```

Figure 32.3 Output from Stanford parser

Basic corpus analysis

Before considering how corpus linguistics can inform language testing, it is important to lay out what information corpora can give us and what information it cannot. Gries (2010) points out that, fundamentally, corpus data consists of distributional frequency data of only two types: specifically, frequencies of occurrence of linguistic elements and frequencies of their co-occurrence. Frequencies are typically reported as normed; for example, frequency per 1,000 or 10,000 words. Reporting frequencies of specific terms can be useful in domain description: for example, in a corpus of conversations between pilots and air traffic controllers, Moder and Halleck (2012; this volume) demonstrated that the most frequent words in the corpus included the numerals one through ten and excluded some words that are most common in general use, such as articles and personal pronouns.

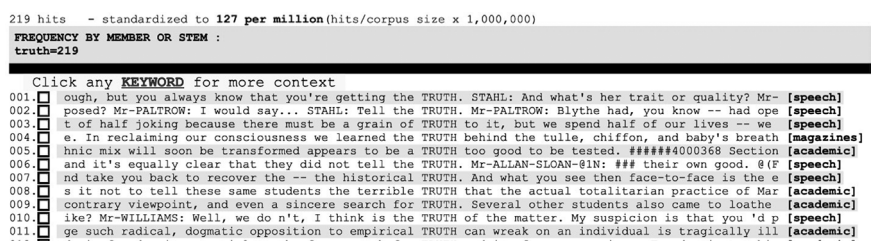
Comparing the frequency in two corpora can be done through calculating the relative frequency ratio, or the quotient of relative frequencies of a word in two corpora, sometimes transformed logarithmically (Gries, 2010). *Keyness* is an important concept in corpus linguistics. A keyword analysis involves listing the words with the highest relative frequency ratios – that is, words that occur significantly more frequently in one corpus than another. For example, Aull (2017) conducted a keyword analysis comparing written responses to different assignments in a first-year college course, such as argumentative essays and annotated bibliographies. Modal verbs and infinitives were more common in the former while

lexical verbs +s (e.g., discusses) were more common in the latter. Similarly, Read (2007) cites Chujo and Utiyama's (2006) corpus-based study of business English as an example of how keyword analysis can inform vocabulary assessment in specific contexts. Beyond domain-specific vocabulary, keyword analysis can also be useful in qualitative discourse analysis. For example, Adolphs *et al*. (2007) combined corpus linguistic keyword analysis with conversation analysis techniques to investigate the communication strategies of health-care workers in consultation phone calls. Statistical measures used to determine keyness are reviewed in Baron *et al*. (2009).

Many corpus analyses go beyond individual words to study words that co-occur with other words or words that appear frequently in specific syntactic constructions (e.g., verb + direct object). Frequency of co-occurrence of words is called *collocation*; some researchers use the terms *collostruction* and/or *colligation* to refer to co-occurrence of constructions/patterns (Gries, 2010). The most frequent statistic reported for collocation is an association measure (such as a pointwise mutual information [MI] score), which is an indication of how strongly two words are predicted to occur together. MI is a measure of the number of times a pair of words is observed together versus the number of times each word is observed separately. With word pairs that occur infrequently, a *t*-statistic is often calculated to estimate the confidence with which an association between two words can be found. For example, Kennedy (2003) examined the use of amplifying adverbs such as *absolutely* and *greatly* in the BNC, extracting general principles of usage of these forms with specific semantic and syntactic forms that can inform language teaching. (See Gries, 2010, for further information on useful statistics for corpus linguistics.)

A more qualitative investigation of collocations can be achieved through a concordancer, which allows researchers to inspect the contexts within which examples of a phenomenon occur. Figure 32.4 shows a screenshot of the term *truth* in the COCA sampler from the concordancer found on the LexTutor website (www.lextutor.ca) (Cobb, 2019). As the figure shows, at least from this example, *truth* appears more frequently with *the* than with *a* (as in "a truth universally acknowledged), a fact that may be of interest to English language learners and their teachers.

Other important concepts in corpus linguistics include the notion of n-grams: that is, strings of *n* adjacent words that appear in a text. An n-gram of two words is called a bigram and one of three words is a trigram. Many corpus researchers are interested in lexical bundles, defined by Biber *et al*. (2004: 376) as "the most frequent recurring lexical sequences in a register." For example, Biber, Conrad, and Cortes found that the most frequent lexical bundles in university teaching and textbooks can be categorized as stance expressions (e.g., "I don't know if"), discourse organizers (e.g., "going to talk about"), or referential expressions (e.g., "that's one of the"). Finally, phrase frames are


219 hits - standardized to **127 per million**(hits/corpus size x 1,000,000)

FREQUENCY BY MEMBER OR STEM :
truth=219

Click any **KEYWORD** for more context

001. ough, but you always know that you're getting the TRUTH. STAHL: And what's her trait or quality? Mr- **[speech]**
002. posed? Mr-PALTROW: I would say... STAHL: Tell the TRUTH. Mr-PALTROW: Blythe had, you know -- had ope **[speech]**
003. t of half joking because there must be a grain of TRUTH to it, but we spend half of our lives -- we **[speech]**
004. e. In reclaiming our consciousness we learned the TRUTH behind the tulle, chiffon, and baby's breath **[magazines]**
005. hnic mix will soon be transformed appears to be a TRUTH too good to be tested. ######4000368 Section **[academic]**
006. and it's equally clear that they did not tell the TRUTH. Mr-ALLAN-SLOAN-@1N: ### their own good. @(F **[speech]**
007. nd take you back to recover the -- the historical TRUTH. And what you see then face-to-face is the e **[speech]**
008. s it not to tell these same students the terrible TRUTH that the actual totalitarian practice of Mar **[academic]**
009. contrary viewpoint, and even a sincere search for TRUTH. Several other students also came to loathe **[academic]**
010. ike? Mr-WILLIAMS: Well, we do n't, I think is the TRUTH of the matter. My suspicion is that you 'd p **[speech]**
011. ge such radical, dogmatic opposition to empirical TRUTH can wreak on an individual is tragically ill **[academic]**

Figure 32.4 Output of concordance on *truth*

non-contiguous word sequences such as "I don't * if," where the asterisk can be replaced by one of several words (Römer, 2017).

Additional analysis methods

Many corpus studies report comparative analyses of specific corpora with reference corpora: for example, comparing language in a workplace setting to a more general reference corpus, investigating changes in word meaning over time, or studying sociolinguistic variation in a language.

While many corpus analyses deal strictly with distributional frequencies and co-occurrences of specific linguistic phenomena, other scholars have used corpus data as a starting point for more complex analyses and for the development of automated measures. These include multi-dimensional analyses closely associated with Biber and his associates (Biber and Conrad, 2009), Coh-Metrix (Graesser *et al*., 2004), and Lu's Syntactic Analyzer (Lu, 2010). Some of these methods and tools go strictly beyond corpus linguistics but have provided valuable insights into textual analyses, particularly with respect to the relationships between features of responses to writing prompts and human scores (Latifi and Gierl, 2020). See Lu (2017) for a review of how syntactic complexity is operationalized in rating scales and automated analysis tools.

How can corpus research inform test development and validation?

As has been pointed out in several recent articles, (Barker, 2010; Weigle and Goodwin, 2016; Cushing, 2017), the capacity for rigorous description and comparative analysis of language can benefit language testing in numerous ways. By comparing a specialized corpus representing a particular TLU situation with a more general corpus, test developers can gain insights into the unique ways in which language is used in specific situations. This information can inform domain description, construct definition, and test task development. For example, the T2K-SWAL was created to develop "a comprehensive linguistic description of the range of university spoken and written registers that are predominant in the university context" in order to "develop procedures for assuring that the language used in assessment tasks is representative of real-life language use" (Biber *et al*., 2004: 2).

Barker (2010) identifies several ways in which learner corpora can also inform item writing. First, analysis of learner corpus data can identify frequent collocational patterns of language use at different levels of proficiency, suggesting which collocations would be appropriate to include in a test. Frequent collocational errors identified in a corpus can also be suitable distractors for multiple-choice items. An example is Voss (2012), which describes the development and validation of a corpus-informed test of academic collocational knowledge. Corpus data can also confirm or refute test writers' intuitions about what learners can do at different proficiency levels. Analyzing learner errors at different levels of proficiency can suggest items that might maximally discriminate between students at those levels.

Corpus data is valuable for developing and validating scoring criteria for performance-based tasks as well. For example, Crosthwaite and Raquel (2019) collected a corpus of student group oral discussion data to validate a group oral discussion test at a tertiary institution in Hong Kong. Analysis of the corpus data was compared with teacher ratings; the researchers found that frequent use of meta-discourse was positively related to student scores, providing validity evidence for the scoring rubric.

Corpus data is also useful for comparing responses to different test tasks or topics, which allows researchers to assess the effects of task variables on performance (Cushing, 2017). For example, using multidimensional analysis, Friginal and Weigle (2014) found that the linguistic features of TOEFL Independent essays differed significantly from those of disciplinary writing, particularly in the natural and health sciences. Furthermore, the linguistic features of the test essays differed by topic as well. Findings such as these have implications for both test development and validation, as test developers want to minimize construct-irrelevant variation due to task variables on the one hand and to demonstrate congruence between test performance and performance beyond the test on the other.

Another application of corpus data in assessment is in supporting the development or revision of rating scales for speaking or writing. For example, Hawkey and Barker (2004) used corpus analysis techniques to derive a common set of writing descriptors for a scale that could be used across several Cambridge English proficiency tests that were targeted at specific proficiency levels. The study involved both manual and computer analyses of responses to a common writing task rated at four levels of proficiency. Such a scale may be useful for helping define the writing construct, interpreting levels of performance in different examinations, and determining how examinations relate to each other in terms of level of difficulty.

Current contributions and research

In this section, I discuss four ways in which corpus linguistics is beginning to make contributions to language testing: (1) construct definition, (2) test validation through an argument-based approach, (3) describing language at different proficiency levels, and (4) automated scoring and feedback tools.

Corpus linguistics and construct definitions

One contribution that corpus linguistics is beginning to make in language testing is in the area of construct definition. Models of language proficiency such as that of Bachman and Palmer (1996) have traditionally kept grammar and vocabulary as distinct subconstructs, and rating scales for speaking and writing frequently, though not always, distinguish between grammar and vocabulary as well. Corpus linguistics studies remind us that it is not individual word use, but words used in combination, that distinguish highly proficient performances from lower ones. For example, Römer (2017), analyzing spoken data from MICASE and the BNC, argues that the prevalence of phraseological elements in spoken texts that cannot be neatly classified into either grammar or lexis suggests that the distinction between these two constructs frequently found in rating scales might usefully be discarded. Along the same lines, Alderson and Kremmel (2013) conclude that "the relationship between syntax and lexis is a continuum" (p. 550) and argue for more tests based on phraseology rather than vocabulary in isolation. Paquot (2019) found that differences in ratings of CEFR proficiency levels on research papers were related to the ability of students to combine words in an effective and idiomatic way, particularly adjective + noun combinations such as "overwhelming majority" and verb + direct object combinations such as "arouse curiosity." Thus, corpus studies are beginning to challenge some of the ways in which we think about language, particularly the relationship between lexis and grammar.

Corpus linguistics and test validation

Over the past two decades, many language testing scholars have focused on an argument-based approach to validity, which sets out specific claims that contribute to an argument for a specific test use (e.g., Kane, 2006, 2012, this volume; Bachman, 2005; Bachman and Palmer, 2010; Chapelle *et al*., 2008; Chapelle and Lee, this volume). Some large-scale tests, such as ACCESS (see Davis *et al*., 2018) and TOEFL (see Chapelle *et al*., 2008), have laid out explicit validity arguments, thus providing a framework for researchers to seek evidence supporting or refuting these claims. Corpus studies are increasingly being used for this purpose; for example, LaFlair and Staples (2017) used corpus-based register analysis (Biber and Conrad, 2009) to examine the extrapolation inference, comparing the linguistic features of spoken responses on the Michigan English Language Assessment Battery (MELAB) with three registers in the target language use domain.

Describing language at different levels of proficiency

One of the most promising applications of corpus linguistics is in describing learner language at different levels of proficiency. An example of such a project is the English Profile project (englishprofile.org; see Kurtes and Saville, 2008), which uses a large corpus of data from Cambridge exams to provide information on the lexical and grammatical characteristics of texts produced by learners at different levels of proficiency as defined by the Common European Framework of Reference for languages (CEFR; Council of Europe, 2001). However, corpus data based on test responses needs to be supplemented by more naturalistic data to be generalizable to language produced in non-test situations.

Automated scoring and feedback tools

One of the most obvious applications of corpus linguistics to language testing concerns the development and validation of automated scoring and feedback engines (see also Xi, this volume). Automated scoring systems require a large corpus of human-scored texts in digital format. These texts are analyzed according to a large number of (ideally) construct-relevant features, and a regression is conducted to determine the features that most accurately predict human scores. The regression model is then applied to a test set of texts that have also been scored to see whether the model can adequately predict the scores. The features that contribute to scores can also be used for automated feedback. Well-known examples of large-scale tests that use automated scoring either fully or in combination with human scores are the Pearson Test of English (PTE) Academic and the TOEFL (Enright and Quinlan, 2010).

Recommendations for practice

Language testing scholars who wish to use corpus data for test development and validation can avail themselves of numerous freely available corpora and analysis tools. In this section, I discuss several resources for locating appropriate corpora and analysis tools, along with some caveats for using them.

What corpora and tools are available?

With every passing year, more and more corpora become available to researchers. English-Corpora.org is maintained by Mark Davies of Brigham Young University and can be found at www.english-corpora.org/corpora.asp. The website contains links to various corpora, some of which are online and others which can be downloaded to a personal computer. In addition to COCA and the BNC, the site includes large web-based corpora such as iWeb (Davies, 2019) and specialized corpora such as the Corpus of Supreme Court Opinions and the Corpus of American Soap Operas. The Corpus Resource Database (www.helsinki.fi/varieng/CoRD/), maintained by the English department at the University of Finland, lists several dozen corpora of historical and current English. A list of learner corpora in a variety of languages is maintained at the Université Catholique de Louvain (Learner Corpora around the World, https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html). Finally, Wikipedia provides a list of corpora in different languages organized by area of the world (https://en.wikipedia.org/wiki/List_of_text_corpora).

Anthony (2013) provides a good introduction to freely available tools for corpus analysis, including WordSmith Tools (Scott, 1996–2016), AntConc (Anthony, 2004–2012), and MonoConc Pro (Barlow, 2012). Earlier tools were limited to processing ASCII characters and had limited functionality. Contemporary tools have several advantages over earlier corpus tools, including common statistics, the ability to process non-ASCII characters, and more user-friendly interfaces, so a knowledge of programming is no longer always critical for corpus researchers.

Anthony (2013) also provides several caveats about the use of these corpus tools. As an illustration, he points out that different corpus tools may count words differently, depending on such factors as whether an apostrophe is counted as part of a word or not. Thus *we'll* can be included in a frequency count of the word *we* by some tools and not by others. It is important for the researcher to be aware of these differences rather than using the tools blindly.

The recent development of very large corpora that are automatically compiled from internet sites and are too large to be stored on personal computers has prompted the development of so-called fourth-generation tools, which store corpora online and are pre-indexed for faster searches. These tools are frequently associated with specific corpora and often require subscriptions to access. The most popular of these tools is corpus.byu.edu (now www.english-corpora.org) according to a survey conducted by Tribble (2006, cited in Anthony, 2013).

Some caveats

Xi (2017) provides two cautions with regard to using corpus linguistics in language testing. First, she cautions against an overreliance on corpus linguistics findings to compare language used on tests with real-world language use. Xi reminds us that assessments are not intended to mimic real-world language use, but rather to "provide an adequate representation of the target language use domain and to elicit a performance that is indicative of a potential performance in corresponding real-world contexts" (2017: 571–72). Furthermore, the choice of features to be analyzed must be motivated by the test construct rather than by the capabilities of a given analytic tool.

Similarly, Xi warns against including fine-grained linguistic features in scoring rubrics simply because they can be extracted from a corpus unless they are "closely motivated by the construct of a particular assessment" (p. 572). While it is tempting to include fine-grained

linguistic features in analytic rubrics, particularly for formative assessment, it is important to ensure that these features are "conceptually well linked to a sub-component of language ability" (p. 572) and can predict human ratings of the same subcomponent.

Future directions

Park (2014) outlines four areas where more work is needed at the interface between corpus linguistics and language testing. One area that needs work is the construction of longitudinal learner corpora to overcome the challenges of comparing learner corpora with comparable reference corpora (Chen, 2013), along with learner corpora that include error annotation and corrections, where feasible. Park also notes a lack of user-friendly corpus analysis tools, though this situation is improving. Two additional areas for more research noted by Park are the use of corpora in dynamic assessment and the development of corpora for regional varieties of English, such as the English used in Singapore. With regard to dynamic assessment, Park suggests that examples of linguistic phenomena from relevant corpora can be used to provide individualized real-time assistance to learners. Corpora of regional varieties of a language can be useful to codify and promote the legitimacy of these varieties.

Xi (2017) introduces several areas in which corpus linguistics may be useful for work in language testing. First, Xi recommends expanding the repertoire of features investigated through corpus analysis, particularly in the area of speaking assessment, while noting the challenges involved in analyzing features beyond syntax and vocabulary. Next, Xi notes that corpus linguistics has been particularly useful in designing test tasks in listening and reading, but less so in speaking and writing. Corpus tools may help in this area by identifying potential tasks that elicit specific kinds of language that prove to be salient in a reference corpus. A third area where corpus linguistics is likely to make inroads in assessment is in the area of formative assessment, since corpus analyses may be beneficial in developing learner progressions across a wide range of linguistic phenomena and how they pattern together rather than focusing on discrete elements such as articles or negation strategies. Xi also recommends closer collaboration between language testers and corpus and computational linguists and between scholars working in L1 and L2 research, who tend to work in parallel without much interaction.

In conclusion, rapid advances in technology have provided numerous opportunities for language testing scholars and practitioners to use corpus linguistics to inform their test development and validation efforts. While some caution in using and interpreting corpus data is always essential, corpus linguistics has the potential to help make language tests more authentic and valid for specific purposes.

Further reading

Cushing, S. T. (2017). Corpus linguistics in language testing research [special issue]. *Language Testing* 34: 441–49.

This issue of *Language Testing* contains five original research articles that illustrate different applications of corpus linguistics data and analysis tools to language assessment, along with an introduction and two commentaries: one by a corpus linguist and the other by an assessment specialist.

Gries, S. T. (2010). Useful statistics for corpus linguistics. *A Mosaic of Corpus Linguistics: Selected Approaches* 66: 269–91.

This article provides an overview of statistics that are useful for analyzing corpus data, which consists essentially of distributional frequency data.

O'Keeffe, A. and McCarthy, M. (eds.). (2010). *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
This edited volume of 45 chapters provides a good introduction to the field of corpus linguistics, with sections devoted to corpus compilation, corpus analysis, and the use of corpora for language analysis, education, and other applications.

References

Adolphs, S., Brown, B., Carter, R., Crawford, P. and Sahota, O. (2007). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics and Professional Practice* 1: 9–28. https://doi.org/10.1558/japl.v1.i1.9.

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas and M. Short (eds.), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London: Longman, 248–59.

Alderson, J. C. and Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing* 30: 535–56. https://doi.org/10.1177/0265532213489568.

Anthony, L. (2012). *AntConc* [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from www.antlab.sci.waseda.ac.jp/.

———. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research* 30: 141–61. https://doi.org/10.17250/khisli.30.2.201308.001.

Aull, L. (2017). Corpus analysis of argumentative versus explanatory discourse in writing task genres. *Journal of Writing Analytics* 1: 1.

Austen, J. (1994). *Pride and Prejudice*. 1813. Retrieved from www.pemberley.com/janeinfo/pridprej.html.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2: 1–34. https://doi.org/10.1207/s15434311laq0201_1.

Bachman, L. F. and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

———. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Baker, P. (2010). Corpus methods in linguistics. In L. Litosseliti (ed.), *Research Methods in Linguistics*. London: Bloomsbury, 93–113.

Barker, F. (2006). Corpora and language assessment: Trends and prospects. *Research Notes* 26: 2–4.

———. (2010). How can corpora be used in language testing? In A. O'Keeffe and M. McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. Abingdon, UK: Routledge, 661–74.

———. (2013). Using corpora to design assessment. In A. J. Kunnan (ed.), *The Companion to Language Assessment*. Hoboken, NJ: Wiley-Blackwell, 1013–28.

Barlow, M. (2012). *MonoConc Pro 2.2* (MP2.2) [Software].

Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik* 20: 41–67.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing* 8: 243–57.

Biber, D. and Conrad, S. (2009). *Register, Genre, and Style*. Cambridge, UK: Cambridge University Press.

Biber, D., Conrad, S. and Cortes, V. (2004). If you look at . . .: Lexical bundles in university teaching and textbooks. *Applied linguistics* 25: 371–405. https://doi.org/10.1093/applin/25.3.371.

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, UK: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V. . . . Urzua, A. (2004). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. TOEFL Monograph Series MS-25. Princeton, NJ: Educational Testing Service.

Callies, M. and Götz, S. (eds.). (2015). *Learner Corpora in Language Testing and Assessment*, vol. 70. Amsterdam, The Netherlands: John Benjamins Publishing Company.

Chapelle, C. A., Enright, M. K. and Jamieson, J. M. (eds.). (2008). *Building a Validity Argument for the Test of English as a Foreign Language™*. Abingdon, UK: Routledge.

Chapelle, C. A. and Plakans, L. (2013). Assessment and testing: Overview. In C. A. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. Hoboken, NJ: Wiley-Blackwell, 241–42.

Chen, M. (2013). Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics* 18: 418–42. https://doi.org/10.1075/ijcl.18.3.07che.

Chujo, K. and Utiyama, M. (2006). Selecting level-specific vocabulary using statistical measures. *System* 34: 255–69. https://doi.org/10.1016/j.system.2005.12.003.

Cobb, T. (2019). *Compleat Lexical Tutor*. Retrieved December 19, 2019, from www.lextutor.ca/conc/eng/.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *Calico Journal*, 14(2–4), 15–33. Retrieved from www.jstor.org/stable/45119223.

Corpus Resource Database (CoRD). Department of English, University of Helsinki. Retrieved December 15, 2019, from www.helsinki.fi/varieng/CoRD/corpora/.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, assessment*. Cambridge: Cambridge University Press.

Cushing, S. T. (2017). Corpus linguistics in language testing research [special issue]. *Language Testing* 34: 441–49.

Crosthwaite, P. R. and Raquel, M. (2019). Validating an L2 academic group oral assessment: Insights from a spoken learner corpus. *Language Assessment Quarterly* 16: 1–25. https://doi.org/10.1080/15434303.2019.1572149.

Dagneaux, E., Denness, S. and Granger, S. (1998). Computer-aided erroranalysis. *System* 26: 163–74.

Davies, M. (2008–). *The Corpus of Contemporary American English (COCA): 520 Million Words, 1990-Present*. Retrieved from http://corpus.byu.edu/coca/.

———. (2018). *The iWeb Corpus: 14 Billion Words*. Retrieved December 28, 2019.

Davis, J. M., Norris, J. M., Malone, M. E., McKay, T. H. and Son, Y. (2018). *Useful Assessment and Evaluation in Language Education*. Washington, DC: Georgetown University Press. Retrieved from muse.jhu.edu/book/59038.

Díaz-Negrillo, A. and Domínguez, J. F. (2006). Error tagging systems for learner corpora. *Revista española de lingüística aplicada*, 83–102.

Enright, M. K. and Quinlan, T. (2010). Complementing human judgment of essays written by English language learners with e-rater® scoring. *Language Testing* 27: 317–34. https://doi.org/10.1177/0265532210363144.

Francis, W. N. and Kucera, H. (1979). *Brown Corpus Manual: Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English for Use with Digital Computers*. Providence, RI: Brown University.

Friginal, E. (2009). *The Language of Outsourced Call Centers: A Corpus-Based Study of Cross-Cultural Interaction*, vol. 34. Amsterdam, The Netherlands: John Benjamins Publishing.

Friginal, E. and Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *Journal of Second Language Writing* 26: 80–95.

Gablasova, D. (2020). Corpora for second language assessments. In P. Winke and T. Brunfaut (eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing*. Abingdon: Routledge, 45–53.

Gablasova, D., Brezina, V. and McEnery, T. (2019). The trinity Lancaster corpus: Development, description and application. *International Journal of Learner Corpus Research* 5: 126–58. https://doi.org/10.1075/ijlcr.19001.gab.

Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside, G. Leech and G. Sampson (eds.), *The Computational Analysis of English: A Corpus-Based Approach*. London: Longman.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers* 36: 193–202. https://doi.org/10.3758/BF03195564.

Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. (2009). *The International Corpus of Learner English. Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gries, S. T. (2009). What is corpus linguistics? *Language and Linguistics Compass* 3: 1225–41. https://doi.org/10.1111/j.1749-818X.2009.00149.x.

———. (2010). Useful statistics for corpus linguistics. In A. Sanchez and M. Almea (eds.), *A Mosaic of Corpus Linguistics: Selected Approaches*. Frankfurt am Main: Peter Lang, 269–91.

Hawkey, R. and Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing* 9: 122–59.

Hawkins, J. A. and Filipović, L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, vol. 1. Cambridge, UK: Cambridge University Press.

Kane, M. (2006). Validation. In R. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: American Council on Education and Praeger, 17–64.

———. (2012). Validating score interpretations and uses. *Language Testing* 29: 3–17. https://doi.org/10.1177/0265532211417210.

Kennedy, G. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

———. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly* 37: 467–87.

Kurtes, S. and Saville, N. (2008). The English profile programme – An overview. *Research Notes* 33: 2–4.

LaFlair, G. T. and Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing* 34: 451–75. https://doi.org/10.1177/0265532217713951.

Latifi, S. and Gierl, M. (2020). Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing* 38: 62–85. https://doi.org/10.1177/0265532220929918.

Learner Corpora Around the World. Retrieved December 15, 2019, from https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Leech, G., Garside, R. and Bryant, M. (1994). Corpus-based research into language: In honour of Jan Aarts. In N. Oostdjik and P. Haan (eds.), *The Large-Scale Grammatical Tagging of Text: Experience with the British National Corpus*. Netherlands: Rodopi Publishers, 47–63.

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22: 319–44. https://doi.org/10.1075/ijcl.22.3.02lov.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15: 474–96. https://doi.org/10.1075/ijcl.15.4.02lu.

———. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing* 34: 493–511. https://doi.org/10.1177/0265532217710675.

Lüdeling, A., Walter, M., Kroymann, E. and Adolphs, P. (2005). Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics* 2005: 14–17.

McArthur, T. (1981). *Longman Lexicon of Contemporary English*. London: Longman.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Michigan Corpus of Academic Spoken English. (2009). Ann Arbor, MI: The Regents of the University of Michigan.

Michigan Corpus of Upper-level Student Papers. (2009). Ann Arbor, MI: The Regents of the University of Michigan.

Moder, C. L. and Halleck, G. B. (2012). Designing language tests for specific social uses. In *Routledge Handbook of Language Testing*. Abingdon: Routledge, 137–89.

Nesi, H. (2001). A corpus-based analysis of academic lectures across disciplines. In J. Cotterill and A. Ife (eds.), *Language Across Boundaries*. London: Continuum, 201–18.

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly* 15: 29–43. https://doi.org/10.1080/15434303.2017.1405421.

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly* 11: 27–44. https://doi.org/10.1080/15434303.2013.872647.

Plakans, L. (2018). Then and now: Themes in language assessment research. *Language Education & Assessment* 1: 3–8. https://doi.org/10.29140/lea.v1n1.35.

Rayson, P., Archer, D., Piao, S. and McEnery, A. M. (2004). *The UCREL Semantic Analysis System*. Lancaster, UK: Lancaster University.

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies* 7: 105–26.

Reppen, R. and Simpson, R. (2002). Corpus linguistics. In N. Schmitt (ed.), *An Introduction to Applied Linguistics*. London: Arnold Publishers, 92–111.

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing* 34: 477–92. https://doi.org/10.1177/0265532217711431.

Scott, M. (2016). *WordSmith Tools Version 7*. Stroud: Lexical Analysis Software.

Simpson, R. C., Lee, D. W. and Leicher, S. (2002). *MICASE Manual. The Michigan Corpus of Academic English*. Ann Arbor, MI: The University of Michigan.

Stubbs, M. and Halbe, D. (2012). Corpus linguistics: Overview. In C. Chapelle (ed.), *The Encyclopedia of Applied Linguistics*. Hoboken, NJ: John Wiley & Sons.

Subirats, C. and Ortega, M. (2012). *Corpus del Español Actual*. Retrieved from http://spanishfn.org/tools/cea/english.

Svartvik, J. (2011). *Corpus Linguistics Comes of Age*. Directions in corpus linguistics: Proceedings of Nobel symposium 82 Stockholm, 4–8 August 1991, 7–13.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97: 77–101. https://doi.org/10.1111/j.1540-4781.2012.01422.x.

Tribble, C. (2012, July 11–14). *Teaching and Language Corpora: Quo Vadis?* 10th Teaching and Language Corpora Conference (TALC), Warsaw.

Van Moere, A., Suzuki, M., Downey, R. and Cheng, J. (2009). Implementing ICAO language proficiency requirements in the versant aviation English test. *Australian Review of Applied Linguistics* 32: 27–1. https://doi.org/10.2104/aral0927.

Voss, E. (2012). *A Validity Argument for Score Meaning of a Computer-Based ESL Academic Collocational Ability Test Based on a Corpus-Driven Approach to Test Design*. Graduate Theses and Dissertations. 12691. https://lib.dr.iastate.edu/etd/12691.

Weigle, S. C. and Goodwin, S. (2016). Applications of corpus linguistics in language assessment. In J. Banerjee and D. Tsagari (eds.), *Contemporary Second Language Assessment*. New York, NY: Bloomsbury, 209–24.

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing* 34: 565–77. https://doi.org/10.1177/0265532217720956.

Section 10
Ethics, fairness, and policy


Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

33
Ethics and fairness

F. Scott Walters

Ethics in language testing

By exploring *ethics* in language testing, one may appreciate its foundational nature. One general definition of ethics given by Yourdictionary.com (2019) is "a moral philosophy or of morals practiced by a person or group of people." An alternate entry states that it is "that branch of philosophy dealing with values relating to human conduct, with respect to the rightness and wrongness of certain actions and to the goodness and badness of the motives and ends of such actions." To some, moral philosophy may at first blush seem somewhat arcane and only vaguely pertinent to the technical "nuts and bolts" of testcraft. Yet in the scholarly literature, language testers have engaged in substantive discussions about what the ethical dimensions of the field might be (e.g., Messick, 1980; Davies, 1997; Schwandt and Jang, 2002; Fulcher and Davidson, 2007), this conversation culminating in a working, planet-wide code of ethical principles (International Language Testers Association [ILTA], 2018). Clearly, ethics (and fairness, discussed later in this chapter) is seen by testers as relevant. Indeed, a practitioner's understanding of what ethics implies for the field could be regarded as being as integral a part of assessment literacy (Fulcher, 2012) as is an understanding of validity, reliability, or test-item design. The ethical imperative for the field might well be summed up in Hamp-Lyons' statement: "No test-taker shall be harmed by [a] test" (1997: 13).

It may be useful to relate language testers' ethical beliefs to more general concepts within moral philosophy. For example, according to Honer *et al*. (2002: 162), human beings tend to express their ethical positions in two stages: first, normative ethics, making assertions about the values they hold. Second, they (sometimes) engage in meta-ethics, explicating the grounds for their value systems. (Honer *et al*. note that while this two-step order is the reverse of what is logical, it is nonetheless psychologically authentic.) Since language testing is an applied science, one may also consider exploring ethics-in-language testing through the lens of the philosophy of science (e.g., Klemke *et al*., 1998; Hatcher, 2004). In general, philosophers describe two broad theoretical approaches to ethics: the deontological, which deals with notions of one's duty to follow "right" versus "wrong" or "good" versus "bad," and the teleological, which deals with the consequences of one's actions (Honer

DOI: 10.4324/9781003220756-44

et al., 2002). Alternatively, various schools of ethics may be conceived as occupying a spectrum (Honer *et al*., 2002). At one end are the more "subjective" approaches: for example, cultural relativism and postmodernism, in which notions of right and wrong are derived from the "social authority" (p. 165) of a particular culture. At the other are "absolutist" or "objective" approaches to ethics, which may be based on reason (e.g., Kant, 1788) and therefore are assumed to be universal. The former, "subjective" set of perspectives might be represented by language testers such as Inbar-Lourie (2008), who describes a sociocultural approach to forming "assessment cultures" in which members "share epistemological suppositions about the dynamic nature of knowledge" and in which "multiple means for assessing learners" are employed and "the need for fair, ethical, valid . . . assessment practices" is recognized (387). Also near the "subjective" end of the spectrum is Hamp-Lyons (2000), who presents a postmodernist, culturally relativistic view, suggesting that language testers "cannot avoid acknowledging the contingent nature of knowledge nor the fact that different stakeholder groups will have competing goals and values. . . . [W]hat is considered ethically acceptable varies from country to country, culture to culture" (581, 589).

These "subjective" ethical perspectives evince a certain perspective on epistemology, the study of the nature of knowledge and of how one determines what is "real or "true." Likewise, the "objective" end of the ethical spectrum also implies a certain "take" on knowledge. One "extreme" objective view is that of logical positivism, which holds that only statements about reality that can be confirmed via direct observation of phenomena or logical arguments are meaningful (e.g., Ayer, 1936). Thus, any subjective statements about reality – including ethical/value judgments – that cannot be objectified by providing evidence and reasoning should be rejected. Some researchers influenced by logical positivism (e.g., Cronbach and Meehl, 1955) have, in turn, had a profound influence on language-testing research: for example, on the notions of construct validity and validation of test scores (Fulcher and Davidson, 2007: 10). However, some philosophers of science point out that such an exclusively objective perspective has problems, since scientific endeavors are grounded in subjective values. For example, Klemke *et al*. (1998) point out that "the search for knowledge expresses a value; and thus distinctions between reliable and unreliable knowledge claims, between good and bad methods . . . are partly normative judgments" (p. 485). In the context of language testing, this may refer to values regarding whom to subject to testing, what skills or knowledge to test, how to interpret test results, or how to use those results. Moreover, as Honer *et al*. (2002) suggest, an absolutist/objectivist view of ethics derived only from reason cannot be the entire picture because even reason (e.g., Kant, 1788) can be informed by the consequences of one's actions – that is, empirical evidence as to what constitutes a good or benefit for the greatest number of people. Concerning an epistemological stance allied with "objective" ethics – namely, materialist reductionism – Hatcher (2004) critiques that paradigm by pointing out that

> objectivity is itself the product of human subjectivity. We can successfully objectify a certain part of our inner model only because we, as human subject, can presume a vast (subjective) context of intuitively understood meanings, feelings, thoughts, intuitions, and desires. [These qualities] derive from our consciousness (self-awareness), which creates the inner, subjective world from which language, logic, and rationality all emerge. . . . The reductionist, then, is looking at objectivity only as a finished *product*, and he is thereby neglecting the whole *process* which has been necessary to produce this product.
>
> *(66; emphasis in original)*

Hence, an exclusively objectivist approach to L2 testing and ethics fails to account for the complexity of human thought and behavior. On the other hand, it may be argued that subjectivism, postmodernism, and cultural relativism are incompatible with an applied science such as language testing, in which the collection, analysis, and interpretation of objective test data are central. In this vein, Fulcher (2012) argues that language testers "should not be tempted to adopt a standpoint epistemology that reduces meaning to individual and group perception" because "if reality or knowledge is what we perceive and may temporarily construct with others, there can be no reality or knowledge at all, and no criteria by which to evaluate the appropriacy of competing validity (or ethical) arguments" (116–17). Similarly, Hatcher (2004) critiques the postmodernist perspective on objectification by pointing out the manifest successes of modern digital technology, noting that the computer "is not a naturally occurring phenomenon. It had to be conceived, constructed, and implemented. If [objective] formal logic and axiomatic systems are simply word games, having no connection with empirical reality, then why do computers work?" (2004: 73). Interestingly, as Hatcher (2004) further points out, objectivist reductionism and relativism/subjectivism do share one point of agreement, namely that "value notions" – and here one may highlight ethical notions –

> cannot be treated objectively and rationally. For the reductionist, objectivity implies viewpoint neutrality which, in turn, implies value neutrality (the suspension of value judgments). . . . For subjectivists, the reality of values is too rich, varied, and fundamentally irrational to admit of rational treatment.
>
> *(2004: 16)*

(To be fair, it is unclear whether Inbar-Lourie's [2008] or Hamp-Lyons' [2000] subjectivist, postmodernist recommendations mentioned here genuinely imply completely nonrationalistic stances.) Regardless, for the purposes of this chapter, it is suggested that mapping out a workable ethical paradigm for the field of language testing seems to require avoiding exclusive adherence to these two philosophical extremes, each of which might undermine or constrain the ethical L2 assessment enterprise.

In fact, L2 assessment researchers over the years seem to have been evolving a conceptual middle ground, which has implications for the exploration of ethics in language testing. Certainly the most well known of these is Messick (1975, 1980, 1989), who highlights the ethical issue of whether or not a given test should be used for a given purpose, warning that the answer to that question "requires an evaluation of the potential consequences of the testing in terms of social values" (1975: 962). He states also that "test use itself is an ethical issue" (1980: 1020) and that testers must be aware of the "consequential validity" of the use of a test (1989: 85–86) – that is, the effect of the use of a test on a given population: the test takers, their families, or the wider community. Messick's work has influenced the thinking of a number of other researchers. For example, Hamp-Lyons (1997) states that "[t]he responsibility of language testers is clear: we must accept responsibility for all those consequences that we are aware of" (1997: 302). Lynch (1997) also lists a number of ethical issues, including the consent, privacy, and confidentiality of the test taker; fairness; and testing decisions. Like Messick (1989), Lynch links test validity with the consequences of its use and further states, "If those consequences entail unethical or immoral outcomes, then the test and those associated with it are accountable" (1997: 317). Similarly, Fulcher and Davidson (2007) argue for what they term effect-driven testing: "The task for the ethical language tester is to look into the future, to picture the effect the test is intended to have,

and to structure the test development to achieve that effect" (p. 144). Also, Davies (1997) emphasizes the professionalization of language testers as a way to develop "a strong ethical foundation" (1997: 328) in the field. Considerations such as these have crystallized in a working code of ethics for the L2 testing profession (International Language Testers Association, 2018).

It should be noted that this so-called philosophical "middle ground" does not appear to dogmatically dismiss those aspects of language testing that relate to legitimate testing issues that happen to be situated toward the ends of the ethical-philosophical spectrum. For example, Hamp-Lyons' (2000) statement that "what is considered ethically acceptable varies from country to country, culture to culture" (2000: 589) may well be accommodated by a pragmatic approach to L2 test development and use, as Fulcher and Davidson (2007) point out, noting that Hamp-Lyons's observation "opens up an awareness and sensitivity to difference and variety, creating a concern for the minority and under-represented in society" (2007: 147). Moreover, the first enumerated principle in the ILTA Code states that "Language testers shall . . . respect all persons' needs, *values* and *cultures* . . . nor [shall they] impose their own *values* . . . to the extent that they are aware of them" (2018: 2, emphasis added). In addition, Principle 8 of the code states that language testers "shall be mindful of their obligations to the society within which they work, while recognizing that those obligations may on occasion conflict with their responsibilities to their test takers and to other stakeholders" (6). Thus, an ethico-philosophical "middle ground" suggests a perspective that encourages a principled engagement with the objective phenomenon of cultural diversity in order to find solutions to any language testing–related issues that may arise within various cultural contexts. Such engagement may imply a dialogic process that resonates with the scientific principle of intersubjective testability – e.g., "Will testing procedure X work in cultural contexts A and B?" (cf. Klemke *et al*., 1998: 32–33).

In any event, these language testers' claims of normative ethics perhaps could be tied to general concepts of moral philosophy mentioned earlier. For example, statements derived from Messick's (1989) concept of consequential validity may be said to collectively occupy a "middle" portion of Honer *et al*.'s (2002) spectrum of approaches to ethics and seem to evince a more pragmatic approach. This resonates with Honer *et al*. (2002), who state, "The utilitarian justifies value judgments by appealing to the pragmatic theory of truth" (172). In addition, it seems that technical concepts of consequential validity, effect-driven testing, and professionalism may be classed as elements – or perhaps tools – supporting a teleological approach to ethics in the field. This focus on the ethical consequences of language testing is, to paraphrase Hatcher (2004), not "a passive, ad hoc compromise" between the extremes of subjectivism and objectivism. Rather, it appears to be "a proactive philosophy of truth-seeking" – perhaps not unlike the process of validation – "which sees philosophy [or language-testing research and development] as a co-operative dialogue with reality [including a variety of researchers, stakeholders, and testing mandates] rather than a polemical dialogue between proponents of various dogmatic points of view" (2004: 15). Indeed, the ILTA code of ethics itself in its current instantiation reflects this dialogic, pragmatic approach: "Both the Code of Ethics and the Guidelines for Practice need to be responsive to the needs and changes within the profession and, in time, these Codes will require revision in response to changes in language testing and in society" (2018: 2).

At any rate, one may assume, for the moment, that the core concept of the ethical behavior of a test developer or test user is regard for the consequences of a test on the lives of test takers. Accordingly, one may find it useful now to move to a consideration of what the

related notion of "test fairness" entails. Indeed, the two concepts seem only to partially overlap, at least from a normative-ethical perspective. For example, Shohamy (2000) states that "language tests employing methods which are not fair to all test takers are unethical" (340), implying that fairness involves behavior and relationships on a more fine-grained level, which may or may not align with an overarching principle of consequential validity as ethics.

Fairness in language testing

Among language testing researchers, the notion of *fairness* has accumulated a number of complex, technical definitions over the years. However, since the word has such widespread everyday usage, consulting a dictionary can be a good starting point for discussion. For example, YourDictionary.com (2019) states that *fairness* is "just and reasonable treatment in accordance with accepted rules or principles" and that to be *fair* is to be "just and honest," "impartial," and "unprejudiced": specifically, "free from discrimination based on race, religion, sex, etc." Merriam-Webster (Mish *et al*., 2003) states similarly that fairness is "marked by impartiality or honesty" and is "free from self-interest, prejudice, or favoritism." As this volume deals with second/foreign/other language testing, it would seem appropriate (or perhaps one should say *fair*) to include a dictionary definition crafted for L2 learners, as is this one from Cambridge (Walter *et al*., 2006): "Treating someone in a way that is right or reasonable, or treating a group of people equally and not allowing personal opinions to influence your judgment."

Such common-sense notions can be contrasted with more technical definitions, which, as mentioned earlier, can actually be somewhat complex (Camilli, 2006). In fact, the various technical definitions of fairness proposed by language testing scholars evince sharp differences as well as overlap. For example, Kunnan (2000) argues that fairness embraces three concerns: validity, accessibility of tests to test takers, and justice (see also Kunnan, this volume). Spaan (2000) first defines fairness as an ideal "in which opportunities are equal" but then argues that "in the natural world, test writers and developers cannot be 'fair' in the ideal sense, but . . . they can try to equitable." This equitability is "the joint responsibility of the test developer, the test user, and the examinee, in a sort of social contract" (Spaan, 2000: 35). Not only do language testing scholars disagree somewhat on the meaning of the word, but L2 test takers can have variable and fluid views of L2 testing fairness as well. For example, Jang reports on a study into "folk fairness" among English as a second language (ESL) learners (2002). She provides evidence that "fairness is not derived [psychometrically] from a test itself, but it is constructed, interpreted and practiced in different ways by various stake holders." Davies (2010) argues that in a language testing context, achieving "fairness" is impossible and even unnecessary. The purpose of the rest of this chapter will be to accommodate and to discuss such differences and similarities of conceptualization in an orderly way.

Fairness can be seen as a *system* or *process* – as distinct from a quality. In contrast, *ethics* is a broader concept than fairness; indeed, language testers can be seen generally as understanding ethics as being broader than fairness (e.g., Xi, 2010; Shohamy, 2000). For the purposes of the remainder of this chapter, *fairness* will be understood as something that may, but not necessarily, be synonymous with justice and may be subsumed under the broader concept of ethics. Nevertheless, it will be seen that although fairness can be so subsumed, formulations of the concept by language testing researchers are not simple.

What is fairness?

What most language testing views of fairness have in common is a desire to avoid the effects of any *construct-irrelevant factors* on the entire testing process, from the test-design stage through post-administration decision-making. The term *construct* may be understood as referring to a piece of knowledge or skill that a test taker may (or may not) possess and that we might be interested in testing for; an example would be ESL oral proficiency. However, certain factors, such as an unconscious bias by a rater against certain non-native accents, may be said to be *irrelevant* to the construct and might distort the test results. Thus, *not* eliminating the effects of these phenomena on the testing process could result in an administration or use of a test that is unfair.

In this context, one dimension of fairness concerns the *roles* of the major stakeholders in achieving language testing fairness. These stakeholders, to use Spaan's (2000) tripartite "social-contract" scheme, are test developers, L2/FL learners, and test users (e.g., teachers or administrators). The role of test developers is manifold. They must try to ensure the validity, reliability, and practicality of their test methods. They must also provide test users with easily understandable guidelines for the use of their tests. Moreover, since test developers are engaged in a social contract with test users and test takers, they must also solicit feedback *from* them to effect improvements to their tests. In this social contract, Spaan points out that test takers have a role to play in test fairness as well: they must become familiar with the testing format and overall test content before taking the test, and they must also try to make sure that the level of the test matches their own skill/knowledge level. Finally, test administrators must give tests to students for whom the tests were designed; to do otherwise would constitute test abuse.

Viewing test developer, test taker, and test user as parties to a social contract highlights a second issue: namely, the phenomenon of *power relations* in testing. Shohamy (2001) points out that language tests have often been used by powerful agencies such as governments, educational bureaucracies, or school staffs, for reasons other than the assessment of language skills. For example, tests have been used to establish discipline, to impose sanctions on schools or teachers, and to raise the prestige of the subject matter being tested. She offers a set of principles organized under the heading of *critical language testing* as a way to engage language testers "in a wider sphere of social dialogue and debate about . . . the roles tests . . . have been assigned to play in society" (Shohamy, 2001: 132–33). One principle is that critical language testing "encourages test takers to develop a critical view of tests as well as to act on it by questioning tests and critiquing the value which is inherent in them." She also "calls for a need to question the purposes and actual uses of tests. Are they intended to measure and assess knowledge or are they used to define and dictate knowledge?" (Shohamy, 2001: 131–32). Such extra-linguistic purposes may even be at variance with purposes stated in official policy documents. A critical language testing perspective asks that all parties to the contract remain vigilant.

Approaches to language testing fairness have evolved over the years. Traditionally, fairness has been conceived relatively narrowly in testing, with a focus being a concern for fairness in the construction of test questions; accordingly, this traditional view is rather technical. Although this view has been somewhat supplanted, it is useful to review it because it is still part of the landscape of language test development.

From this technical perspective, one approach to examining whether or not test questions are fair is through a *test sensitivity review*. Such reviews are performed by trained judges employed by test-development organizations who examine test tasks to determine whether

they contain language or content that may be considered stereotyping, patronizing, inflammatory, or otherwise offensive to test takers belonging to subgroups defined by culture, ethnicity, or gender (AERA *et al*., 1999: 82). An offensive test prompt may adversely affect test-taker performance on such an item. Often such biases in language are subtle (Ramsey, 1993), which is why a panel of experts is needed.

Another concept related to fairness in test questions is *test bias*. This is a technical term indicating a testing situation in which a particular test use results in different interpretations of test scores received by cultural, ethnic, gender, or linguistic subgroups (AERA *et al*., 1999: 81). A frequent synonym for test bias is *differential item functioning*, or DIF. Bias or DIF is considered to be present when a test item is differentially difficult for an ethnic, cultural, or gender-related subgroup that is otherwise equally matched with another subgroup in terms of knowledge or skill (Willingham and Cole, 1997). Among the statistical methods used to uncover DIF are the standardization procedure (Dorans and Kulick, 1986) and the Mantel-Haenszel method (Holland and Thayer, 1988).

More recently, however, researchers have tended to conceptualize fairness rather more broadly and less technically than did earlier testing researchers, crucially with regard to its relationship with validity. Kane (2010) points out that the relationship between fairness and validity depends on how the two concepts are defined. Narrowly defining validity and broadly defining fairness will result in validity being considered a component of fairness. On the other hand, a broad definition of validity and a narrow conceptualization of fairness will result in fairness being understood as a part of validity. Such considerations are not trivial. In fact, arriving at an understanding of fairness in relation to validity is essential as validity is "the most fundamental consideration in developing and evaluating tests" (AERA *et al*., 1999: 9).

Xi (2010) points out that some researchers regard fairness as a test characteristic that is *independent* of validity. One example of this is the *Standards for Educational and Psychological Testing* (AERA *et al*., 1999), which define it as having, at minimum, three components: lack of item bias, the presence of equitable treatment of all test takers in the testing process, and equity of opportunity of examinees to learn the material on a given test. While fairness here is not linked directly with validity, these 1999 *Standards* do mention that fairness "promotes the validity and reliability of inferences made from test performance" (AERA *et al*., 1999: 85). Another example is the *Code of Fair Testing Practices in Education* (2004). This code focuses on the respective responsibilities of test developers and test users to ensure fairness by engaging in a sort of dialogue. In this dialogue, the test developer provides "evidence that the technical quality, including reliability and validity, of the test meets its intended purposes" (*Code*, "Test Developers," sect. A-5). On the other hand, the test users "[e]valuate evidence of the technical quality of the test provided by the test developer and any independent reviewers" (*Code*, "Test Users," sect. A-5). Briefly, test developers and test users must exercise fairness in four major areas of the testing process – creating and selecting tests, administering and scoring them, reporting and interpreting results, and providing certain information to test takers. Such information should include "the nature of the test, test taker rights and responsibilities, the appropriate use of scores, and procedures for resolving challenges to scores" (*Code*, "Test Users," sect. D). While fairness is considered to be "a primary consideration in all aspects of testing," it is "not an isolated concept, but must be considered in all aspects of the testing process" (*Code*, introductory section).

Another approach to conceptualizing language testing fairness, as Xi (2010) further points out, is as an all-embracing concept that *subsumes* validity. Kunnan (2000) articulates a framework in which fairness includes issues of validity, accessibility to test takers, and

justice. Under *validity*, Kunnan includes issues such as construct validity, DIF, insensitive test-item language, and content bias. An example of the latter might be a dialect of English employed in the test prompts that differs in some respects from another English dialect that may constitute the L2 of the test taker. Under *accessibility*, Kunnan indicates issues such as affordability, geographic proximity of test taker to the testing site, accommodations for test takers with disabilities, and opportunity to learn. "Opportunity to learn" is closely connected with the notion of *construct underrepresentation* (Messick, 1989), which indicates the ability of a test to measure some aspects of a construct or skill, but not others. A test may be measuring aspects of a construct, such as knowledge of a particular rule of language pragmatics, that certain test takers will not have had the opportunity to learn; thus, they may score poorly on the test, despite the fact that they may be proficient in other relevant areas of the construct. Finally, Kunnan's facet of *justice* embraces the notion of whether or not a test contributes to social equity. Kunnan (2004) later modified this model to include absence of bias and test-administration conditions.

A third approach, that of Kane (2010), sees fairness and validity as *overlapping* concepts, neither being completely included within the other, but closely related ways of considering the same issue. Kane's definition of test fairness draws on political and legal concepts. One, *procedural due process*, states that the same rules should be applied to everyone in more or less the same way. Kane also bases his definition on *substantive due process*, which states that the procedures applied should be reasonable both in general and in the context in which they are applied. In applying this twin-definition of fairness to assessment, Kane gives two principles: The first is *procedural fairness*, in which test takers are treated "in essentially the same way . . . take the same test or equivalent tests, under the same conditions or equivalent conditions, and their performances [are] evaluated using the same (or essentially the same) rules and procedures." The second is *substantive fairness*, in which the score interpretation and any test-based decision rules are reasonable and appropriate for all test takers (Kane, 2010: 178).

A fourth conceptualization, articulated by Willingham and Cole (1997), views fairness as "an important aspect of test validity" and, more specifically, as *comparable validity* for all test takers and test-taker groups (Willingham and Cole, 1997: 6–7). In this context, a fair test is one for which both the (preferably small) extent of statistical error of measurement and the inferences (hopefully, reasonable ones) from the test results regarding test-taker ability are comparable from individual to individual and from subgroup to subgroup. Willingham and Cole (1997) state that comparable validity must be met at all stages of the testing process – when designing the test, developing the test, administering the test, and using the test results. Comparable validity must be achieved by selecting test material that does not give an advantage to some test takers for reasons that are irrelevant to the construct being measured; by avoiding difficulties in the testing situation that are unrelated to the construct, such as unnecessary inclusion of technical terms in the test prompts; and by employing equivalent scoring procedures. Thus, Willingham and Cole see fairness as having three qualities linked to validity: (1) comparable opportunities for test takers to show their knowledge and skills, (2) comparable test tasks and scores, and (3) comparable treatment of test takers in test interpretation and use.

Recent contributions and research

In the first decade after the turn of the millennium, two approaches to fairness in testing were proposed, which will be summarized next. One is a proposal from Xi (2010), and the other is the Test Context Framework by Kunnan (2009).

Xi both expands the definition of test fairness and offers a new framework for investigating fairness issues. The definition builds on that offered by Willingham and Cole (1997; see earlier in this chapter) – namely, comparable validity for all subgroups – by stating that fairness is "comparable validity for *identifiable* and *relevant* groups across all stages of assessment, from assessment conceptualization to use of assessment results," where "construct-irrelevant factors, construct underrepresentation, inconsistent test administration practices, inappropriate decision-making procedures or use of test results have no *systematic* or *appreciable* effects on test scores, test score interpretations, score-based decisions and consequences for all *relevant* groups of examinees" (Xi, 2010: 154; emphasis in original).

After offering this definition, Xi then outlines a framework for language testing fairness research, consisting of a *fairness argument* embedded within a *validity argument*. A validity argument is a chain of inferences that leads a test user to appropriate interpretations of test results. Xi's validity argument framework consists of six successive sub-arguments: (1) there is evidence that the domain of L2 use that is of interest provides a meaningful basis for our observations of test-taker performance on the test; (2) there is evidence that the observed test scores reflect that domain of L2 use and not construct-irrelevant factors; (3) there is evidence that the observed scores on the test are generalizable over similar language tasks on other, similar tests; (4) there is evidence that the aforementioned generalization of observed scores can be linked to a theoretical interpretation (i.e., the construct, the theoretical skill) of such scores; (5) there is evidence that the theoretical construct can explain the L2 use in actual situations envisioned by the users of the test; and (6) there is evidence that the language test results are "relevant, useful, and sufficient" (Xi, 2010: 157) for determining the level of L2 ability. Each of these sub-arguments is supported by certain *assumptions*. To take one example, for sub-argument (1), which claims an inferential link between observations of test-taker performance and a certain domain of L2 use, there may be one or more assumptions: (a) the assessment tasks on the test represent the intended L2 domain (in this case, academic English skills on the collegiate level), and (b) critical language skills, knowledge, and processes needed for collegiate study are represented on the test.

Embedded in this chain of sub-arguments and underlying assumptions, Xi proposes, can be a *fairness argument*, which consists in part of a series of rebuttals, one or more posed to each of the validity sub-arguments. One can conceive of such rebuttals as research questions into the degree of fairness of a given language test: i.e., each rebuttal serves as a practical check on the claims of each sub-argument. For example, to the first of the given sub-arguments, one can ask whether or not the domain of L2 use actually provides a meaningful basis for observations of test-taker performance. More specifically, a rebuttal may challenge a specific assumption underlying the sub-argument. Are all the test tasks on the test *equally* representative of the intended collegiate-level, academic English L2 domain for *all* test-taker groups? If not, then one could say that the fairness argument has been weakened. As another example, Xi suggests that one might formulate rebuttals to the sub-argument regarding construct-irrelevant factors. Such rebuttals could address issues such as inconsistent test administration, lack of accommodations for test takers with disabilities, and rater bias against certain groups. In short, one or more rebuttals would be used to determine any weaknesses in each sub-argument and hence in the overall validity argument: that is, to see whether or not there are systematic or appreciable effects of construct-irrelevant factors on test scores or score-based decisions.

Another recent proposal for investigating language testing fairness issues is Kunnan's (2009) Test Context Framework. This approach is intended to consider the wider political, educational, cultural, social, economic, legal, and historical aspects of a test. In this, it differs

somewhat from other, more psychometrically focused approaches considered earlier, such as DIF and even Xi's fairness argument framework. It has a certain overlap with Kane's (2010) applications of political and legal concepts to language testing, and it also resonates with the analyses of power relations in language testing offered by Shohamy (2001), both mentioned earlier.

Kunnan's approach is briefly illustrated here by way of an example dealing with versions of the naturalization test, a tool used by the US federal government to screen non-native-English-speaking applicants for US citizenship (see also Kunnan, this volume). The stated purpose of this test is to assess prospective citizens on their "civic nationalism" and potential for "social integration," as well as their English language ability. The test asks the prospective citizen to listen, speak, read, and write short English sentences dealing with facts about US history and government. After giving a breakdown of the components and uses of the naturalization test, Kunnan (2009) asks a set of questions that throw light on the degree of fairness involved: (1) "Test requirement and purpose: Are the requirement and purpose meaningful?" (2) "Test conceptualization, content, and operations: Can the test assess English language ability and knowledge of US history and government?" (3) "Test consequences: Can the test bring about civic nationalism or social integration?"

To the first question, for example, Kunnan takes a politico-historical perspective, pointing out that the naturalization test is a holdover from the anti-communist furor of the 1950s in the United States. It also reflects racist anti-immigration and voting policies in the early twentieth century, through which literacy requirements had been instituted. Kunnan points out that these older policies had become irrelevant by the 1990s, and moreover, recent US court rulings had made the literacy requirement meaningless. Thus, he implies, current use and administration of the naturalization test is unfair. As another example, Kunnan responds to the same question regarding test purpose and meaningfulness by pointing out that the test as currently constituted simply cannot test the constructs of "civic nationalism" or "social integration." This is because the measures of English ability or of knowledge of US history and government are, at best, indirect assessments of those constructs and not direct measures of actual community participation. Hence, one may conclude, use of the naturalization test is unfair to test-taking prospective citizens. Kunnan's approach thus brings wider social factors into consideration when evaluating the fairness of a language test.

Future directions

At this point, one may well ask how language testing fairness might be conceived and addressed over the next decade from the time of this writing. To make a few predictions, it is reasonable to assume that test bias will be a continual concern as long as unequal societal and educational divisions continue, as along gender, ethnic, and racial lines. One might also hope that the theoretical dispute over the relationship between fairness and validity will begin to wind down toward consensus. As can be seen from the discussion in this chapter, researchers and practitioners can indeed conceptualize the two as being different, even to the point of being able to craft research questions into fairness independent of validity. Thus, one may predict that in ten years' time, most scholars will agree that they are not synonymous, but rather are distinct, related aspects of test construction and use.

Yet another prediction involves the evolution of current approaches to fairness research. It would seem that at the present time, we have two general ways of approaching L2 testing fairness. Examination of each general approach reveals an apparent divergence. One, exemplified by Xi's (2010) model, uses a technical-formalistic approach, which

may conceivably be informed by focused test-development tools such as DIF, as well as procedural-fairness concerns outlined by Kane (2010). This is a manifestly *micro-analytic* approach suited to scientific investigations of fairness and so is not, apparently, able to accommodate wider social, historical, or political contextual issues. In contrast, the other general approach to test fairness takes a broad-based, societal perspective, exemplified by Kunnan's (2009) Test Context Framework, Shohamy's (2001) principles of critical language testing, and certain substantive-fairness concerns suggested by Kane (2010). This latter approach offers qualitative, *macro-analyses* of test requirements, purposes, and consequences and thus seems better suited for evaluating social and political issues pertaining to language testing fairness.

There are at least two productive responses to this seeming divergence between the two approaches. One is to consider the framework for validation argued by Messick (1988, 1989). This framework, alluded to in the first section of this chapter, outlines consequential bases for validity that encompass *value implications* and *social consequences*. These are two concerns arguably relevant to a qualitative, interpretive, social justice–oriented approach to language testing fairness. Moreover, Messick's formulation of *consequential bases* for validity posits "test use" as comprising two fairness-related elements (see also Bachman, 1990: 242). One is *construct-related evidence for validity*, which can be an issue amenable to technical micro-analysis. The other element is *social consequences*, which clearly overlaps a socially broad approach to fairness. (In fact, with these considerations, one encounters an intersection between a fine-grained investigation into fairness and the overarching principle of language-testing ethics: consequentialism, or an effect-driven approach to testing, discussed earlier [cf. Fulcher and Davidson, 2007].) With this theoretical perspective, one could modify a given technical-formalistic approach to fairness to accommodate broader social aspects of language-testing contexts. For example, one might craft rebuttals to assumptions in an initial, inferential step involving domain definition. Other rebuttals would be relevant to inferential steps regarding test use (Xi, in her model, calls it *utilization*), challenging claims regarding the relevance and usefulness of test results for making decisions.

With these suggestions in mind, we might make a prediction that language testing scholarship in the area of fairness and ethics will evince a principled synthesis of such "macro" and "micro" approaches to fairness, drawing on Messick's (1988, 1989) unitary model of validity. Thus, a formalistic model could integrally provide meaningful structure for both scientific-technical, micro-analytic research efforts and ethico-social, macro-analytic investigations. Achieving such an integrated practice will no doubt take time, given differing theoretical views of fairness vis-à-vis validity. One may hypothesize, however, that as validity arguments may draw upon multiple and canonical types of validity evidence (e.g., content-related, construct-related, consequence-related), so, too, might a fairness argument integrated with a validity argument draw on different, evolving, theoretical models of test fairness.

Recommendations for practice

In effect, this portion of the chapter, as with the first section on ethics, has been about defining the very thing it seeks to explain: fairness. As can be surmised from the earlier survey of thought and research, this definition is still very much in flux. Accordingly, the following section of the chapter will offer some exercises, which, it is hoped, will help form a definition of fairness of practical use to the reader.

Practical exercise 1

With a group of colleagues, consider an L2 test with which some or most of you are familiar. It may be a test of writing ability used to screen applicants to a university graduate school program or a test of pronunciation given to non-native-speaking graduate students who wish to be university teaching assistants, or perhaps a standardized exit test taken by secondary-level students and promulgated by the testing agency of the national, state, or provincial government. Try to determine the author(s) and stated purpose(s) of the test. If you can obtain a test-user's manual, or some statement as to test purposes and characteristics online, this may be helpful for ascertaining sub-arguments and their underlying assumptions.

Then, using Xi's (2010) model as a guide, first construct a six-step validity argument with sub-arguments and assumptions. Afterward, examine the test items closely and, if possible, interview anyone in your group or class who may have taken the test in question. Then try to construct a series of rebuttals in a fairness argument for as many of the six inferential steps as you can. It may help if you conceive of this as an "anti-fairness" argument. As far as you are able to determine, how fair is the test or test use? You or your group may find a worksheet grid such as the one shown in Figure 33.1 useful in organizing your thoughts.

Keep in mind that the points in the cells are only examples to get you started. You may well find more than one assumption or more than one rebuttal per inferential step.

Practical exercise 2

Regarding the relationship between fairness and validity, Davies (2010: 171) expresses the opinion that everything that can be claimed about fairness is already covered by validity and

Name of test: _____

Test purpose(s): _____

Sub-arguments	Underlying assumptions	Rebuttals
The domain of L2 use (describe here) provides a meaningful basis for our observations of test-taker performance	According to the online test user's manual, test tasks represent the domain (describe)	We feel that the L2 domain is only partly represented by the tasks (for example...)
The observed test scores reflect that domain of L2 use and not construct-irrelevant factors	Reasonable accommodations given to test takers with disabilities (describe)	Rubric is vaguely worded, so construct irrelevancy is possible Rater training uneven across semesters

Figure 33.1 Worksheet grid

Source: based on Xi (2010)

that attempts to integrate fairness into validity are meaningless. Read (or reread) the chapter on validity in this volume (by Chapelle and Lee), paying attention to the descriptions of consequential evidence for validity. Do you agree with Davies's opinion? Why or why not? Is "fairness" just another name for "validity"? Might Kane (2010: 177), mentioned earlier, be correct in saying that they are just "closely related ways of looking at the same basic question"?

Practical exercise 3

In the Future Directions section of this chapter, it was suggested that micro-analytic approaches to fairness, such as that of Xi (2010), and macro-analytic approaches, such as that of Kunnan (2000), could be integrated by virtue of Messick's (1989) unitary model of validity. Do you think this possibility is workable? Why or why not? In a group, discuss the pros and cons of this idea, keeping in mind that Kunnan's view of the relationship between validity and fairness differs from Xi's view.

Further reading

Kunnan, A. J. (2008). Using the test fairness and wider context frameworks. In L. Taylor and C. Weir (eds.), *Multilingualism and Assessment: Achieving Transparency, Assuring Quality, Sustaining Diversity*. Papers from the ALTE Berlin Conference. Cambridge, UK: Cambridge University Press.
This is a presentation of Kunnan's Test Context Framework, giving an example of a "macro-analytic" approach to investigating test fairness.

Lee, Y. (2010). Identifying suspect item bundles for the detection of differential bundle functioning in an ESL reading comprehension test: A preliminary study. In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida: Studies in Language Testing 9*. Cambridge, UK: Cambridge University Press, 105–27.
This paper presents an example of a "micro-analytic" fairness study, taking concepts and techniques of differential item functioning (DIF) and applying them to the "test bundle": i.e., a group of related items on a test.

Toulmin, S., Rieke, R. and Janik, A. (1984). *An Introduction to Reasoning*, 2nd edn. New York, NY: Macmillan.
This is a good resource for understanding the formal antecedents of Xi's (2010) embedded fairness/validity argument structure.

Willingham, W. W. (1999). A systemic view of test fairness. In S. Messick (ed.), *Assessment in Higher Education: Issues in Access, Quality, Student Development, and Public Policy*. Mahwah, NJ: Lawrence Erlbaum Associates, 213–42.
This book chapter is a good, short exploration of fairness issues at all stages of the test development process, from test design and item writing to test administration, interpretation, and use.

Yan, X., Cheng, L. and Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examine L1 background on scores of an ITA speaking test. *Language Testing* 36: 207–34.
This is a good example of a study that implements Xi's (2010) approach to investigating fairness. It first examines, via confirmatory factor analysis (CFA), the factor structure of responses to three different integrated tasks across three different L1 groups and then interprets any bias in the context of the assessment of international teaching assistants (ITAs), shedding light on construct-related evidence of the validity of the tasks, as well as fairness. While the study is clearly "micro-analytic," the authors' wider considerations of "radically different language backgrounds and experiences" of test takers suggest the possibility of a synthesis of "macro" and "micro" approaches to fairness.

References

AERA/APA/NCME (American Educational Research Association, American Psychological Association and National Council on Measurement in Education). (1999). *Standards for Educational and Psychological Testing*. Washington, DC: Author.

Ayer, A. J. (1936). *Language, Truth, and Logic*. London: Penguin Modern Classics. Cited in G. Fulcher and F. Davidson (eds.). (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York, NY: Routledge.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

Camilli, G. (2006). Test fairness. In R. Brennan (ed.), *Educational Measurement*, 4th edn. Westport, CT: American Council on Education and Praeger, 221–56.

Code of Fair Testing Practices in Education. (2004). *The Code of Fair Testing Practices in Education*. Washington, DC: Joint Committee on Testing Practices. Retrieved July 18, 2020, from www.apa.org/science/programs/testing/fair-testing.pdf.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52: 281–302. https://doi.org/10.1037/h0040957.

Davies, A. (1997). Demands of being professional in language testing. *Language Testing* 14: 328–39. https://doi.org/10.1177/026553229701400309.

———. (2010). Test fairness: A response. *Language Testing* 27: 171–76.

Dorans, N. J. and Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the scholastic aptitude test. *Journal of Educational Measurement* 23: 355–68. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly* 9: 113–32. https://doi.org/10.1080/15434303.2011.642041.

Fulcher, G. and Davidson, F. (eds.). (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York, NY: Routledge.

Hamp-Lyons, L. (1989). Language testing and ethics. *Prospect* 5: 7–15.

———. (1997). Washback, impact and validity: Ethical concerns. *Language Testing*, 14: 295–303. https://doi.org/10.1177/026553229701400306.

———. (2000). Social, professional, and individual responsibility in language testing. *System* 28(4): 579–98. https://doi.org/10.1016/S0346-251X(00)00039-7.

Hatcher, W. S. (2004). *Minimalism*. Hong Kong: Juxta Publishing.

Holland, P. W. and Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum Associates, 129–45.

Honer, S. M., Hunt, T. C. and Okholm, D. L. (eds.). (2002). *Invitation to Philosophy: Issues and Opinions*. Belmont, CA: Wadsworth/Thomson Learning.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing* 25: 385–402. https://doi.org/10.1177/0265532208090158.

International Language Testers Association. (2018). *Code of Ethics*. Retrieved July 27, 2020, from www.iltaonline.com/page/CodeofEthics.

Jang, E. (2002). *Folk Fairness in Language Testing*. Paper presented at the Southern California association for language assessment research conference (SCALAR 5), May 15–16.

Kane, M. (2010). Validity and fairness. *Language Testing* 27: 177–82. https://doi.org/10.1177/0265532209349467.

Kant, E. (1999 [1788]). *Critique of Practical Reason* (M. J. Gregor, trans.). Cambridge, UK: Cambridge University Press.

Klemke, E. D., Hollinger, R. and Rudge, D. W. (eds.). (1998). *Introductory Readings in the Philosophy of Scienc*e, 3rd edn. Amherst, NY: Prometheus Books.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida: Studies in Language Testing 9*. Cambridge, UK: Cambridge University Press, 1–14.

———. (2004). Test fairness. In M. Milanovic and C. Weir (eds.), *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference*. Cambridge, UK: Cambridge University Press.

———. (2009). Testing for citizenship: The U.S. Naturalization test. *Language Assessment Quarterly* 6: 89–97. https://doi.org/10.1080/15434300802606630.

Lynch, B. K. (1997). In search of the ethical test. *Language Testing* 14: 315–27. https://doi.org/10.1177/026553229701400308.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 955–66.

———. (1980). Test validity and the ethics of assessment. *American Psychologist*, 1012–27.

———. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and H. I. Braun (eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum.

———. (1989). Validity. In R. L. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: American Council on Education and Macmillan, 13–103.

Mish, F. C. (ed.). (2003). *Merriam-Webster's Collegiate Dictionary*, 11th edn. Springfield, MA: Merriam-Webster, Incorporated.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland and H. Wainer (eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum, 367–88.

Schwandt, T. A. and Jang, E. E. (2002). Linking validity and ethics in language testing: Insights from the hermeneutic turn in social science. *Studies in Educational Evaluation* 30: 265–80. https://doi.org/10.1016/j.stueduc.2004.11.001.

Shohamy, E. (2000). Fairness in language testing. In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment*. Cambridge, UK: Cambridge University Press, 15–19.

———. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. Harlow, UK: Longman.

Spaan, M. (2000). Enhancing fairness through a social contract. In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment: Selected Papers from the 19th Language Testing Research Colloquium, Orlando, Florida: Studies in Language Testing 9*. Cambridge, UK: Cambridge University Press, 35–37.

Walter, E. (ed.). (2006). *Cambridge Advanced Learner's Dictionary*, 3rd edn. Cambridge: Cambridge University Press. Retrieved July 18, 2020, from dictionary.cambridge.org/dictionary/british/fair_1.

Willingham, W. W. and Cole, N. (1997). *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing* 27: 147–70. https://doi.org/10.1177/0265532209349465.

Yan, X., Cheng, L. and Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examine L1 background on scores of an ITA speaking test. *Language Testing*, 36: 207–34. https://doi.org/10.1177/0265532218775764.

YourDictionary.com. (2019). *YourDictionary.com*. Burlingame, CA: LoveToKnow Corporation. Retrieved July 27, 2020, from www.yourdictionary.com/fairness.

34

Standards in language proficiency measurement

Bart Deygers

In testing and in education, standards have been used for over a century, for the purpose of certification, for reasons of accountability, and to keep track of the performance of students and schools. The impact and prominence of standards have been rising since the 1950s (Takala *et al*., 2013), and by now, some educational systems have implemented language proficiency standards from the preschool level up to adult education (Cox *et al*., 2018). This chapter introduces three kinds of language proficiency standards and critically examines their benefits and pitfalls.

Defining a standard

We can define a standard as an agreed-upon way of doing or measuring things in order to stimulate cross-context consistency and comparability. Because of standards, screw threads are compatible worldwide, and freight containers have specific dimensions. Essentially, standards are guidelines or measurable characteristics, but they have a number of specific characteristics that makes them different from regulations, which are legislative in nature (Hatto, 2010b, 2010a). In industry, standards are written by experts, reviewed by peers, and ratified by a recognized body. Industrial standards gain legitimacy by virtue of voluntary consensus and proven utility. If a community of users fails to see the usefulness of a standard, it will receive little uptake and be rendered essentially meaningless. Generally, a standard is considered useful if it increases efficiency, quality, or harmonization or facilitates measurement. Consequently, a good industry standard must be precise, unambiguous, clear, and reliable within a certain margin of error (Hatto, 2010a; International Organization for Standardization, 2016).

Using strict formatting requirements in order to promote consistency and clarity, the International Standardization Organization (ISO) has supported the development of nearly 23,000 internationally recognized industry standards, encompassing a vast array of contexts. The standards managed by ISO need to meet the key characteristics of a robust industrial standard: exactness, transparency, impartiality, effectiveness, relevance, and coherence (International Organization for Standardization, 2018).

Arguably, none of the traditional standards in language testing (e.g., CEFR, ACTFL, etc.) check all six boxes associated with a quality ISO standard. Perhaps the most common

 DOI: 10.4324/9781003220756-45

problem in language standards is vagueness or "descriptional inadequacy" (Fulcher *et al*., 2011: 9). Quite possibly, the context dependency and elusiveness of language proficiency (Hudson, 2012), together with the history and genealogy of the standards in use today, help explain why this is the case. To better understand the standards used in language performance measurement today, this chapter takes a broad and historical perspective.

Types of language proficiency standards

In this chapter, we distinguish between three types of standards: educational performance indicators (e.g., PISA, PIRLS, NAEP, PIAAC), language proficiency frameworks (e.g., ACTFL, CEFR), and institutionalized language tests (e.g., IELTS, TOEFL). These three types of standards have received wide and often voluntary uptake by a community of users, but serve dissimilar goals.

The goal of educational performance indicators is to compare the performance of L2 learners or L1 users by administering the same test in different educational settings. These assessment-driven standards typically focus on a range of domains (e.g., science, mathematics) but often include a linguistic component. Educational performance indicators allow policy makers and researchers to monitor the performance of their own educational system over time but also trace its international ranking in relation to the educational systems in other states or countries. As a result, they aim to impact policy or spur policy change (Singer *et al*., 2018). International educational performance assessments reside under the auspices of organizations such as the International Association for the Evaluation of Educational Achievement (IEA) or the Organization for Economic Cooperation and Development (OECD) and often rely on international research teams to conduct and analyze the assessments.

Centralized tests are at the heart of educational performance indicators; the performance levels and thresholds are inextricably tied to one specific measurement instrument. The same is not the case for language proficiency frameworks, in which the level descriptors take center stage. They serve to offer an external benchmark of language proficiency for examination boards, policy makers, test takers and other stakeholders to refer to (Fulcher, 2016).

Some organizations developing language proficiency frameworks (e.g., the performance descriptors by the American Council on the Teaching of Foreign Languages, or ACTFL) also offer tests, but the frameworks are not meaningless without them, as they would be in the case of a PISA or a PIRLS test. Language proficiency frameworks provide generic descriptions of different levels of L2 proficiency. Often, they are language independent (e.g., the Common European Framework of Reference for Languages, or CEFR; Council of Europe, 2001), but some are dedicated to a specific language (e.g., CEFR-J, an adaptation of the CEFR for Japanese learners of English; see Negishi and Tono, 2016) or a limited number of languages (e.g., Canadian Language Benchmarks, or CLB, for English and French; see Centre for Canadian Language Benchmarks, 2012). Often, the descriptors are context independent, but some language proficiency frameworks are quite context specific indeed (e.g., STANAG 6001; see NATO, 2014). Many, but not all, language proficiency frameworks have ties to national or supra-national politics (e.g., China Standards of English; see Ministry of Education of the People's Republic of China, 2018) while others are private nonprofit organizations (e.g., ACTFL).

With Hatto (2010b), standards issued by national or international bodies but created by commissions of independent experts could be called *formal standards*. *Private standards*, on the other hand, are created by private exam boards such as IELTS or TOEFL. The band

levels developed by these exam boards might not have been intended as standards but have started being used as such due to their widespread use for high-stakes purposes.

The policy perspective: educational performance indicators

The roots of the largest international educational performance indicators, such as the Progress of International Reading Literacy Study (PIRLS) (see Mullis *et al*., 2017), the Programme for International Students Assessment (PISA) (see OECD, 2016), and the Programme for the International Assessment of Adult Competencies (PIAAC) (see OECD, 2019) can be traced back to the 1950s, but arguably, they have never been quite as prominent as they are now.

Educational performance indicators are focused on measuring and comparing the (reading) performances of schools, educational systems, and economies (Takala *et al*., 2013). Their aim is not to certify but to chart trends and to offer data that can impact policy (Singer *et al*., 2018). Most people who use the data or read the reports will thus adopt a norm-referenced approach, scrutinizing the relative position of one country's educational performance over time or in relation to other systems at the same point in time. Additionally, however, international educational performance indicators have a criterion-referenced component, which can be interpreted as minimal performance standards. In PISA, for example, level 2 is also called the baseline level, implying that 15-year-olds should be able to perform tasks at this level in order to be able to participate in society (see Table 34.1).

Table 34.1 PISA reading performance levels

Level 6
Making multiple inferences, comparisons and contrasts that are both detailed and precise.
Level 5
Locating and organizing several pieces of deeply embedded information, inferring which information in the text is relevant.
Level 4
Locating and organizing several pieces of embedded information or interpreting the meaning of nuances of language in a section of text by taking into account the text as a whole.
Level 3
Recognizing the relationship among several pieces of information, or integrating several parts of a text in order to identify a main idea.
Level 2: baseline level
Retrieving one or more pieces of information that may have to be inferred.
Level 1a
Retrieving one or more independent pieces of explicitly stated information, or interpreting the theme or purpose of a text on a familiar topic.
Level 1b
Retrieving a single piece of explicitly stated information in a short, syntactically simple text with a familiar context and text type.

Note: The original – substantially longer and more detailed – descriptors have been edited from OECD (2016: 164–66)

It is important to stress that the organizations behind these initiatives have different agendas and goals, which may impact the construct of the tests. IEA, which develops PIRLS, is a conglomerate of research centers, government-affiliated research institutions, researchers, and analysts. Their stated goal is to "research, understand, and improve education worldwide" (IEA, 2019). In line with this mission, PIRLS aims "to provide the best policy-relevant information about how to improve teaching and learning and to help young students become accomplished and self-sufficient readers" (Mullis *et al*., 2017: 4). The OECD, which funds and – through various local research partners and subcontractors – runs PISA and PIAAC, is an economic partnership between 36 mainly Western member states, which aims "to shape policies that foster prosperity, equality, opportunity and well-being for all" and "to better prepare the world of tomorrow." Correspondingly, the goals of PISA and PIAAC are focused on participation in society and in the labor market (OECD, 2016, 2019).

In spite of – or because of – their impact, educational performance indicators have come under intense scrutiny in recent years. PISA especially has been challenged on a number of fronts. Perhaps the most widely supported condemnation of PISA to date was published in 2014, when over 80 academics and educationalists called for a moratorium on the test (The Guardian, 2014). The main strands of criticism on OECD-based international educational performance indicators focus on their impact and their construct. Critics argue that the relatively short three-year cycles of these assessments encourage policy makers to think in terms of short-term fixes (The Guardian, 2014). Related to this, OECD assessments have been criticized for not being attuned to local educational systems (Gaber *et al*., 2012; Liss, 2013; Sjøberg, 2015), for being restricted to what can easily be measured, and for reducing the aim of education to preparation for the labor market (Sjøberg, 2015; The Guardian, 2014; Y. Zhao, 2020).

The can-do perspective: language proficiency frameworks

The foundational years (1950s–1970s)

A pioneer in the field of language proficiency standards, Kaulfers was one of the first people to observe that that test scores do not carry meaning in and of themselves. Because the same score on different tests might say something very differently about a test taker's ability, Kaulfers (1944) introduced a listening proficiency scale, linked to a listening test. Table 34.2 lists Kaulfers's level descriptors, which – if only for the can-do approach – bring to mind listening descriptors still in use.

Innovations in educational measurement often result from a societal or institutional need to achieve a greater degree of order (Stein, 2016). And so, when, in 1952, the US Civil Service Commission wanted to create an inventory of the foreign language proficiency of government employees and found that there was little or no consistency or system in the dizzying array of incompatible scores and profiles reported, they decided to create a single L2 proficiency scale that all employees could refer to (Sollenberger, 1978; Spolsky, 1995).

In his first-hand account, Sollenberger (1978) explains that the Foreign Service Institute (FSI) scale, as it was to be known, was to comprise six levels, ranging from zero (no competence) to five (native speaker or bilingual competence), but there was no clear theoretical or empirical rationale for choosing a six-level scale. In 1955, based on the outcomes of a pilot of the FSI scale among 200 officers, the descriptors were refined and subdivided into speaking and reading scales. The first FSI definitions of speaking proficiency levels were

Table 34.2 Kaulfers's aural performance scale

Level	*Descriptor*
4	Can understand popular radio talks, talking pictures, ordinary telephone conversations, and minor dialectal variations without obvious difficulty.
3	Can understand ordinary conversation on common topics with the aid of occasional repetition or paraphrastic statements.
2	Can understand the ordinary questions and answers in routine transactions involved in independent travel.
1	Can catch a word here and there and occasionally guess meaning through inference.
0	Cannot understand the spoken language.

Note: Level numbers added by author

Source: Kaulfers (1944: 139)

adopted in 1956, with John B. Carroll consulting on the test construction and scale revision process (Liskin-Gasparro, 1984).

Carroll approved of the scaling efforts made by Kaulfers but resisted the idea of conceptualizing language skills as unitary traits. He supplemented the general speaking scale with more specific descriptors: accent, grammar, vocabulary, fluency, and comprehension (Carroll, 1954). For each of these scales, six proficiency levels were defined. In 1968, these level descriptors were revised, standardized (on educated native speakers), and adopted by the Interagency Language Roundtable (ILR) (Liskin-Gasparro, 1984). The first part of each ILR level descriptor is displayed in Table 34.3. All levels except for level 5 could be modified by a "+" to indicate a performance that exceeds the minimum requirements of one level but falls short of the next.

It is difficult to overstate the importance of the pioneering work done on the FSI scales, which directly influenced many of the standards in use today in form and content. In a modified format, the FSI scales are still in use and are known as ILR scales as they resulted from a 1985 Interagency Language Roundtable that revised the scales and formally included the plus levels.

STANAG 6001: The ILR scale goes transatlantic

In the 1970s, the Bureau for International Language Coordination (BILC) started adapting the ILR descriptors for use among NATO member states. When NATO accepted the standards in 1976, they became known as Standardization Agreement 6001 (STANAG 6001). The purpose of the standards was to facilitate international staff recruitment and to be able to compare national language proficiency standards using one and the same descriptive system (Green and Wall, 2005). The standards are available for English and French – the two working languages of NATO.

Some 50 years after the launch of STANAG 6001, some fundamental problems still prevent it from being a vehicle for full comparability of language proficiency levels across NATO member states. A major issue has to do with interpretability (Brooks and Hoffman, 2013; Dubeau, 2006). Much like the CEFR and the ILR descriptors, the wording in the STANAG 6001 descriptors may be somewhat vague and may leave room for interpretation. Additionally, because they include both general purpose and specifically military descriptions of language proficiency, the STANAG 6001 descriptors may not always be entirely

Table 34.3 FSI/ILR speaking descriptors (1968)

Level	*Code*	*Definition*
Native or bilingual proficiency	S-5	Speaking proficiency equivalent to that of an educated native speaker. Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers . . .
Full professional proficiency	S-4	Able to use the language fluently and accurately on all levels normally pertinent to professional needs. Can understand and participate in any conversation within the range of his experience . . .
Minimum professional proficiency	S-3	Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests . . .
Limited working proficiency	S-2	Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events . . .
Elementary proficiency	S-1	Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on topics very familiar . . . can understand simple questions and statements, allowing for slowed speech . . .
No proficiency	S-0	

Note: Descriptors trimmed to first 30 words. (For full descriptors, see Sollenberger, 1978: 19–22).

suited to the specifics of military communication (Brooks and Hoffman, 2013; Fulcher, 2015). Other problems that may impede the optimal operationalization of the STANAG 6001 descriptors in the field include a lack of theoretical basis in second language acquisition theory and somewhat vague level demarcations (Green and Wall, 2005). In order to achieve a more uniform interpretation of the STANAG 6001 descriptors, BILC members continue to have regular standardization meetings. Additionally, every three years, the STANAG 6001 descriptors are under mandatory revision (NATO, 2014). Currently, the fifth major revision is in use.

ACTFL: An academic language proficiency framework

For the first decades following the introduction of language proficiency standards, academic interest was limited, so the first standardization efforts and language proficiency frameworks resulted from government or military initiatives (Spolsky, 1995). By the mid-to-late 1970s, the interest in proficiency guidelines and curricula started growing in academic circles (Green and Wall, 2005), and in the 1980s and 1990s, a growing concern for accountability helped spur the creation and proliferation of the ACTFL guidelines (Cox *et al*., 2018; Ricardo-Osorio, 2008). In 1982, ETS and ACTFL had begun work on the development of provisional proficiency guidelines that would translate the ILR scale to an academic context (Brooks and Hoffman, 2013; Liskin-Gasparro, 1984; for a detailed overview of ACTFL's history, see Cox *et al*., 2018). Table 34.4 shows how the current ACTFL levels relate to the original ILR levels.

Table 34.4 ACTFL alignment with ILR scales and CEFR levels

ILR-ACTFL concordance[1]				
	ACTFL-CEFR concordance[2]			
	Receptive skills		*Productive skills*	
ILR	*ACTFL*	*CEFR*	*ACTFL*	*CEFR*
4–5	Distinguished	C2		
3–4	Superior	C1.2	Superior	C2
2+	Advanced High	C1.1	Advanced High	C1
2	Advanced Mid	B2	Advanced Mid	B2.2
2	Advanced Low	B1.2	Advanced Low	B2.1
1+	Intermediate High	B1.1	Intermediate High	B1.2
1	Intermediate Mid	A2	Intermediate Mid	B1.1
1	Intermediate Low	A1.2	Intermediate Low	A2
0+	Novice High	A1.1	Novice High	A1
	Novice Mid	0	Novice Mid	0
	Novice Low	0	Novice Low	0

[1] See Brooks and Hoffman (2013); Hudson (2012); Liskin-Gasparro (1984)

[2] See ACTFL (2016)

The ACTFL guidelines describe what learners can do at specific proficiency levels. The scale identifies eleven language proficiency levels that encompass five main levels (from novice to distinguished) and three sublevels (low, mid, high). Table 34.4 also shows these levels and their CEFR counterparts.

What sets ACTFL apart from other language proficiency frameworks is that it not only produces guidelines and descriptors but also controls ACTFL-based assessment through Language Testing International (LTI).[1] Another defining feature of ACTFL is its formal integration into the US educational system (Cox *et al*., 2018: 105). Having been incorporated into the National Standards for Foreign Language Learning in the 21st Century, ACFTL now impacts classroom practice in the US from K–12 to college (Glisan, 2012; Little, 2019).

Like any successful language proficiency framework that has gained global resonance, the ACTFL guidelines have also been subject to criticism. Notably, Fulcher (1996) criticized them for lacking an empirical foundation, for disregarding the qualities and imperfections of real-world utterances, and for an overreliance on native speaker norms. (For a full discussion, see Liskin-Gasparro, 2003.) Subsequent ACTFL revisions (ACTFL, 2012) have tried to answer these points of criticism, but according to some reviewers, they still apply to a certain degree (Little, 2019).

The CEFR: policy meets practice

ACTFL and the CEFR are sometimes seen as two transatlantic sides of the same coin, but there are a number of important differences. The CEFR is embedded in the Council of Europe's vision of multilingualism and lifelong learning (Council of Europe, 2001, 2018; Trim, 2012) and is focused on language learning and teaching. ACTFL's prime focus is on assessment (see Little, 2019). The CEFR is essentially a language policy document and quite possibly the most influential language policy document to date (Figueras, 2012).

Its roots can be traced back to the 1970s, when the Council of Europe's efforts to examine the possibility of creating a unified system of expressing language proficiency levels led to *The Threshold Level* (van Ek, 1975; van Ek and Trim, 1991a). This level would later be known as B1 and described the linguistic context and challenges of L2 users living in a foreign European country. In the wake of *The Threshold Level*'s success, the new language proficiency levels *Waystage* (van Ek and Trim, 1991b) and *Vantage* (van Ek and Trim, 2001) were described and introduced. When the CEFR was published in 2001, it incorporated the work done on the *Threshold*, *Vantage*, and *Waystage* levels, as well as numerous other language proficiency scales, including the ILR levels (Council of Europe, 2001).

Table 34.5 shows the structure underlying the six CEFR levels: three broad levels identify basic, independent, and proficient language users, and within each of these three levels, two sublevels exist. In some scales, these sublevels are further subdivided into a basic level and a so-called "plus-level."

CEFR level descriptors are meant to be general, language-independent, and context-independent descriptions of what learners can do with language (Hudson, 2012; North, 2014a). The impressionistic nature of the CEFR level descriptors means that test developers using the framework are required to adapt it to their own local context and to the needs of the language learners at hand (North, 2014b). It also means that the CEFR allows for – encourages – the development of differentiated profiles or requirements (e.g., uneven requirements, such as writing at B1, speaking at B2).

Of all of the content the CEFR offers, the level descriptors have drawn the most attention. More than likely, the level descriptors, listed in tables, are what have given the CEFR such wide appeal among teachers, testers, publishers, and policy makers (Figueras, 2012; Little, 2007; Trim, 2012). The scholars who built the CEFR's foundation were initially hesitant to compartmentalize language proficiency in levels organized in hierarchical tables, but they decided to use that structure for practical reasons. As John Trim recalled, they had not anticipated the appeal of this approach to curriculum designers and policy makers:

> Practical considerations overrode the theoretical misgivings as to the validity if the concept of "level." We had used the term "level" originally despite deep misgivings about the concept. We could see no reason to break the process of language learning into a series of steps. . . . Over time, it became apparent that our reasoning took too little account of the realities of the social organization of language learning.
>
> *(Trim, 2012: 29)*

Table 34.5 Organization of CEFR levels

Purpose[1]	*Level*		*User categorization*
Survival	A1	Breakthrough	Basic user
Routine transaction & interaction	A2	Waystage	
	B1	Threshold	Independent user
Academic/professional	B2	Vantage	
	C1	Effective operational proficiency	Proficient user
	C2	Mastery	

[1] See Little (2019)

The CEFR was always intended to be more than a collection of illustrative tables (North, 2014a); at its heart are the values of multilingualism, lifelong learning, and the international mobility of people and ideas. However, over recent years, it has become clear that the CEFR is also being used for purposes inimical to these original ideals. Normative use of the illustrative descriptors has been frequently observed (Deygers *et al*., 2018; Fulcher, 2004), and CEFR levels are being used to justify migration policies meant to curb the movement of people and to support monolingual ideologies (Rocca *et al*., 2019). These misuses of the CEFR have engendered a strand of CEFR criticism that is focused on its political use (Barni, 2015). Other authors have criticized the methodological foundations of the framework (Alderson, 2007), the lack of theoretical support from SLA theory (Hulstijn, 2007), and the vagueness of the level descriptors (Galaczi *et al*., 2011). The recently published CEFR companion volume (Council of Europe, 2018) answers some of the critique of the original CEFR: it no longer upholds a native speaker norm (see McNamara, 2014), pays more attention to multilingualism (Krumm, 2007), and focuses on the concept of mediation. Many of the most fundamental critiques of the CEFR have not been addressed in the companion volume, however (Deygers, 2019).

In spite of the criticism, the CEFR has been a remarkably successful document that has received global and voluntary uptake in a wide array of contexts and applications. It has transformed language teaching, language testing, and language policy in Europe and around the world and has inspired the creation of new language proficiency frameworks adapted to local contexts, such as the CEFR-J (Negishi and Tono, 2016), which describes the English language proficiency of Japanese L1 learners.

Canadian language benchmarks: task-based can-dos

The realization that test scores provided by different language schools and exam boards often lack transparency is what started the FSI initiative, what inspired the creation of the CEFR, and what led BILC to start work on STANAG 6001. It is also what led the Canadian government to fund the development of the Canadian Language Benchmarks (CLB) in the 1990s (Peirce and Stewart, 1997).

Rooted in Bachman and Palmer's theory of communicative language ability, the CLB provide a descriptive continuum of language ability that is competency based, task based, and learner centered (Centre for Canadian Language Benchmarks, 2012). The CLB offer descriptive statements of 12 language proficiency levels for the four traditional language skills in specified performance conditions. The benchmarks describe language tasks that learners should be able to perform at these levels. Additionally, the CLB function as a national standard for curriculum planning so as to streamline the ESL tuition for migrants across Canada.

Table 34.6 shows how the 12 benchmarks are structured and how they relate to the CEFR (North and Piccardo, 2018). The collaboration between the CEFR and CLB to align two frameworks reflects the similarities between the two frameworks, which share a can-do approach, a focus on communicative competence, and a learner-oriented philosophy.

China Standards of English: a curriculum-based standard

In an endeavor to establish a common standard of English, various Asian countries have adopted (Taiwan; see Wu, 2014) or adapted (Japan; see Negishi and Tono, 2016) the CEFR. The Chinese government, however, decided to create a language framework specific to the

Table 34.6 Overview of the CLB and their relationship with the CEFR levels

Basic Language Ability		
Communicating in predictable context about basic needs and familiar matters		
CLB1	Initial	Pre-A1
CLB2	Developing	A1
CLB3	Adequate	A1/A2
CLB4	Fluent	A2
Intermediate Language Ability		
Functioning independently in familiar situations and a few less well-known contexts		
CLB5	Initial	B1
CLB6	Developing	B1
CLB7	Adequate	B1
CLB8	Fluent	B2
Advanced Language Ability		
Effective, accurate, and appropriate communication in a wide range of contexts		
CLB9	Initial	B2
CLB10	Developing	C1
CLB11	Adequate	C1
CLB12	Fluent	C2

Chinese context (Jin *et al*., 2017). The China Standards of English (CSE) are meant to streamline the English language curriculum in China from elementary school to university and to create shared performance standards for teaching, learning, and assessment. The CEFR inspired but did not guide the development of the China Standards of English (Jin *et al*., 2017). Nevertheless, even though the CEFR is not mentioned in the CSE, it clearly inspired the can-do approach, the organization of the document, and the use of an overall scale in combination with self-assessment scales and subscales (Ministry of Education of the People's Republic of China, 2018).

The CSE define nine levels of English language proficiency, organized in three stages. In line with the CSE goals, the levels have been linked to the curriculum. Table 34.7 displays the organization of these levels, shows how the CSE vocabulary standards have been calibrated against the CEFR, and links the levels to the Chinese curriculum (W. Zhao *et al*., 2017).

The certification perspective: large-scale language tests

In this section, we focus on test scores that are used as standards. As a clear case in point, we will use the context of university admission, in which IELTS and TOEFL have come to be seen as standards because they have gained wide acceptance (Hyatt and Brooks, 2009).

In the English-speaking world, TOEFL and IELTS have traditionally dominated the competitive market of language assessment for university admission and are still the most widely accepted tests (Green, 2018; Weigle and Malone, 2016). These institutionalized high-stakes international language tests are clearly distinct from international educational performance indicators because they do not aim to impact educational policies but focus on certifying individual performances. In fact, Educational Testing Systems (ETS), the organization behind TOEFL, explicitly disapproves of the practice of ranking countries on the basis of TOEFL scores and regards this as "a misuse of data" (ETS, 2018: 13). Being international language tests, TOEFL and IELTS are also different from language proficiency frameworks

Table 34.7 CSE overview and CEFR link

CSE stage	*CSE level*	*CEFR*[1]	*Curriculum*
Advanced stage	Level 9		
	Level 8	C1	Advanced college English curriculum requirement
	Level 7	B2	Intermediate college English curriculum requirement
Intermediate stage	Level 6	B1+	Basic college English curriculum requirement
	Level 5	B1	End of senior secondary school
	Level 4		
Elementary stage	Level 3	A2	End of junior secondary school
	Level 2		
	Level 1	A1	End of primary school

[1] The CEFR alignment applies specifically to vocabulary; see W. Zhao *et al*. (2017)

since their aim is not to describe but to evaluate language proficiency (IELTS, 2019). The fact that certain scores on these tests are being used as de facto standards is not a policy that is pursued by the test developers.

The University of Cambridge Local Examinations Syndicate (UCLES) started developing and administering English language tests for non-native speakers of English in 1913 (Weir, 2003). TOEFL was first introduced some 50 years later by the National Council on the Testing of English as a Foreign Language (Spolsky, 1995). IELTS was established in the 1980s (Davies, 2008) and has grown to become the largest international English language test worldwide. It was taken by 3.5 million test takers in 2018 (British Council, 2019a).[2]

These tests focus on scoring and certification but also offer test-taker profiles at specific score levels. TOEFL candidates receive separate scores in reading, listening, writing, and speaking. Each skill is scored on a 30-point scale, and performance descriptors have been drawn up for four score ranges on this scale, ranging from "below low-intermediate" to "advanced" (ETS, 2019). An advanced reader, for example, will typically "understand academic passages in English at the introductory university level" (ETS, 2019: 1). IELTS uses a system of ten bands (0–9) to assess reading, listening, writing, and speaking. The four scores on the separate skills are then averaged to yield a final band score. Six profiles have been linked to these overall band scores, ranging from "limited user" to "expert user" (see Table 34.3). A good user (i.e., band score 7) is described as somebody with "an operational command of the language, though with occasional inaccuracies, inappropriate usage and misunderstandings in some situations" (British Council, 2019b).

University admission requirements scores of 6.5 to 7.5 on IELTS and a TOEFL iBT score range of 90 to110 have come to be seen as sufficient proof of a language proficiency level that is minimally required to commence academic studies in English. Table 34.8 shows the minimum IELTS and TOEFL requirements for international students to the ten highest-ranking universities in the Times Higher Education index.

In the admission policy of British and American Universities, language test scores are used more often than language proficiency framework levels (Deygers *et al*., 2018). Since not all institutions make use of language proficiency frameworks, language test providers have drawn up score equivalence guidelines to help institutions with setting level requirements. Table 34.9 shows two such score equivalence tables: one equating TOEFL iBT scores with IELTS bands, published by ETS (2010), and one published by Pearson (2019), equating PTE Academic scores with IELTS and TOEFL scores. As the table shows, the outcomes of the two equivalence

Table 34.8 Minimum English language proficiency requirements for international students in arts and humanities

	TOEFL iBT	*IELTS*
Stanford University (2019)	100	Not accepted
The University of Cambridge (2017)	110	7.5
University of Oxford (2019)	110	7.5
Massachusetts Institute of Technology (2019)	90	7
Harvard University (2019)	English test not required	
University College London (2018)	100	7
Princeton University (2019)	Level requirement not stated	
The University of Chicago (2019)	100	7
Yale College (2019)	100	7
University of California, Berkeley (2019)	80	6.5

Note: Selected universities represent the top ten universities of the Times Higher Education ranking for arts and humanities

Table 34.9 Score equivalence guides compared

ETS score equivalence[1]		*Pearson score equivalence*[2]			*CEFR alignment*		
IELTS	TOEFL iBT	IELTS	PTE A	TOEFL iBT	IELTS[3]	TOEFL iBT[1]	PTE A[4]
0–4	0–31						
4					B1		
4.5	32–24	4.5	0			B1	A2
5	35–45	5	36			(24–35)	(30–42)
			38	40–44			
5.5	46–59	5.5	42	54–56	B2	B2	B1
			46	65–66		(46–93)	(43–58)
		6	50	74–75			
6	60–78		51				
			53	79–80			
6.5	79–93	6.5	58		C1		
			59	87–88			B2
			64	94		C1	(59–75)
7	94–101	7	65			(94–114)	
			8	99–100			
			72	105			
7.5	102–109	7.5	73		C2		
			78	113			C1
8	110–114	8	79			C2	(76)
8.5	115–117	8.5	83			(115–120)	
			84	120			
9	118–120	9	86				

[1] ETS, 2010
[2] Pearson, 2019
[3] Taylor, 2004
[4] Riazi, 2013

tables are not quite the same. For example, the ETS study finds that an overall IELTS score of 5.5 corresponds to a TOEFL iBT 46–49, but the Pearson study equates the same IELTS score with a TOEFL iBT score range of 54–56. The same dissimilarities can be found in the right-hand part of Table 34.9. When plotting the CEFR score alignment outcomes of the three tests (De Jong *et al*., 2014; ETS, 2010; Riazi, 2013; Taylor, 2004) it is easy to see that the threshold levels for the different CEFR levels do not quite match. (For an extensive study on this topic, see Green, 2018.)

Like most language proficiency standards, TOEFL and IELTS have been subjected to criticism. One strand of criticism focuses on the use of tests scores. A wide range of studies have documented consistent uneducated or unsubstantiated use of test scores for university admission (Baker, 2014; Deygers and Malone, 2019; Hyatt, 2013; O'Loughlin, 2008, 2013). Setting score thresholds in a simplistic manner without a careful analysis goes against the recommendations of the test providers, undercuts a usage-based validity argument in the line of Kane (2013; this volume) and Bachman and Palmer (2010), and undermines the credibility of test scores as standards of academic language proficiency. Other researchers have criticized the power that these tests have over the lives of individuals (Hamid and Hoang, 2019) and the lack of accountability they face for the scores they assign (W. S. Pearson, 2019; Sarich, 2012). However, given the rise of international student mobility (Cantwell, 2015; UNESCO, 2018) and the increasing need for documented English language proficiency in the context of education, employment, and migration, it is unlikely that major high-stakes tests of English will soon see a reduction in terms of candidature or that their levels will cease to be used as language proficiency standards in certain contexts.

Conclusion: are language proficiency standards "standards"?

This chapter has provided an overview of three types of language proficiency standards and discussed a range of concrete examples (PIRLS, PISA, PIAAC, the FSI/ILR scales, STANAG 6001, the ACTFL guidelines, the CEFR, the Canadian Language Benchmarks, the China Standards of English, IELTS, and TOEFL). Many of these language proficiency standards are similar but not the same. Many are somehow linked or aligned, but not always robustly so. All standards discussed in this chapter have been widely adopted, but none have been free of fundamental criticism. Commonly recurring critiques include construct irrelevance, theoretical under-specification, and descriptive vagueness. Educational performance indicators such as PISA or PIRLS are intended to be used on a population level and have proven to be highly influential in guiding educational policies. Critics have argued that OECD-based standards focus too heavily on education in function of the economy and do not take into account the variety of educational practices across the globe. The use of large-scale language tests as de facto standards in high-stakes contexts (e.g., university admission, citizenship procedures) has become widespread but is not uncontested. Uneducated use of test scores has been widely documented and has been the subject of criticism.

It is unlikely that any of the language proficiency standards discussed in this chapter meet the defining criteria of a standard: specificity, reliability, and universal recognition (Hatto, 2010b). Of course, standardizing screw threads and freight containers is a more clearly delineated and concrete practice than standardizing levels of language proficiency (see also Fulcher, 2016). Authors have argued that language is changeable across contexts, that each of the world's 7,000 languages has its own unique lexico-grammatical system, that every speaker has their own idiolect, and so on (Hudson, 2012). There is truth to this

argumentation, and it helps to explain the diversity of standards, but it does not adequately explain why language- and context-specific standards (e.g., STANAG 6001) suffer from the same theoretical and descriptive shortcomings as general, language-independent standards (e.g., CEFR).

Variability alone cannot account for the descriptional inadequacy that typifies many language performance standards today. Their shortcomings can probably also be explained – at least in part – by the fact that many build on intuitions and descriptions going back to the 1950s, on reified language proficiency levels (Fulcher, 2004), or on descriptors that may have a rather thin empirical basis. As a result, to an extent, the broad field of language assessment continues to use standards that are supported by shaky foundations (Hulstijn, 2007) and explains the shortcomings of these standards by invoking the elusive variability of language.

To achieve greater robustness and less variability, we may need to fundamentally rethink the standards we currently use. Such a reconceptualization is drastic but not impossible; outside linguistics, too, widely used standards have been changed when the need for improvement was apparent. In 2018, the International Bureau of Weights and Measurement decided to change the definition of the kilogram, the ampere, the kelvin, and the mole by linking them all to the same seven constants. It is not impossible to conceive a similar shift taking place in language measurement, if language proficiency standards are truly to function as veritable standards. A truly empirical, performance-driven description of language could provide a good starting point (Fulcher, 2012). Quite possibly, natural language processing will be instrumental in creating language performance standards that are able to achieve greater levels of precision.

Further reading

Sollenberger, H. E. (1978). Development and current use of the FSI oral interview test. In J. L. D. Clark (ed.), *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service, 1–13.

When tracing the history of language performance standards, it is worth going back to first-hand accounts of the development of the precursors to today's standards.

Stein, Z. (2016). *Social Justice and Educational Measurement*. Oxon and New York, NY: Routledge.

This original and insightful book applies ideas from moral and political philosophy to educational assessment. Stein argues that standardized assessment has not led to increased educational equality and proposes a model that offers a solution for how standardized testing could be used to foster equal educational opportunities.

Trim, J. L. M. (2012). The common European framework of reference for languages and its background: A case study of cultural politics and educational influences. In M. Byram and L. Parmenter (eds.), *The Common European Framework of Reference: The Globalisation of Language Education Policy*. Buffalo, NY: Multilingual Matters, 14–36.

John Trim, one of the people who laid the groundwork for the CEFR, remembers which ideas, intentions, and aspirations inspired the creation of what was to become perhaps the most widely used set of language performance standards.

Notes

1 For STANAG 6001 there is the centralized Supreme Headquarters Allied Powers (SHAPE) test, but individual NATO member states are free to develop and administer their own STANAG 6001–based test alongside SHAPE.

2 Over the years, TOEFL has been administered to over 35 million test takers. The organization does not wish to share or publish annual numbers of test takers, however (private communication, July 22, 2020).

References

ACTFL. (2012). *ACTFL Proficiency Guidelines 2012*. Washington, DC: American Council on the Teaching of Foreign Languages.

ACTFL. (2016). *Assigning CEFR Ratings to ACTFL Assessments*. American Council on the Teaching of Foreign Languages. Retrieved from www.actfl.org/publications/guidelines-and-manuals/assigning-cefr-ratings-actfl-assessments.

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal* 91: 659–63. https://doi.org/10.1111/j.1540-4781.2007.00627_4.x.

Bachman, L. and Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Baker, B. A. (2014). *Investigating Language Assessment Literacy in Canadian University Admissions*. Amsterdam: LTRC.

Barni, M. (2015). In the name of the CEFR: Individuals and standards. In B. Spolsky, O. Inbar-Lourie and M. Tannenbaum (eds.), *Challenges of Language Education and Policy. Making Space for People*. New York: Routledge, 40–52.

British Council. (2019a). IELTS grows to 3.5 million a year. *Take IELTS*. Retrieved from https://takeielts.britishcouncil.org/about/press/ielts-grows-three-half-million-year.

———. (2019b). Understand and explain the IELTS scores. *Take IELTS*. Retrieved from https://takeielts.britishcouncil.org/teach-ielts/test-information/scores-explained.

Brooks, R. L. and Hoffman, M. (2013). Government and military assessment. In *The Companion to Language Assessment*. John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla069/abstract.

Cantwell, B. (2015). Are international students cash cows? Examining the relationship between new international undergraduate enrollments and institutional revenue at public colleges and universities in the US. *Journal of International Students* 5: 512–25.

Carroll, J. B. (1954). *Notes on the Measurement of Achievement in Foreign Languages*.

Centre for Canadian Language Benchmarks. (2012). *Canadian Language Benchmarks. English as a Second Language for Adults*. Ottawa, ON, Canda: Citizenship and Immigration Canada.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg, France: Council of Europe.

———. (2018). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Council of Europe Language Policy Division.

Cox, T. L., Malone, M. E. and Winke, P. (2018). Future directions in assessment: Influences of standards and implications for language learning. *Foreign Language Annals* 51: 104–15. https://doi.org/10.1111/flan.12326.

Davies, A. (2008). *Assessing Academic English: Testing English Proficiency 1950–89: The IELTS Solution*. Cambdridge, UK: Cambridge University Press.

De Jong, J. H. A. L., Becker, K., Bolt, D. and Goodman, J. (2014). *Aligning PTE Academic Test Scores to the Common European Framework of Reference for Languages*. Pearson. Retrieved from https://pearsonpte.com/wp-content/uploads/2014/07/Aligning_PTEA_Scores_CEF.pdf.

Deygers, B. (2019). The CEFR companion volume: Between research-based policy and policy-based research. *Applied Linguistics*. https://doi.org/10.1093/applin/amz024.

Deygers, B. and Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing* 36: 347–68. https://doi.org/10.1177/0265532219826390.

Deygers, B., Zeidler, B., Vilcu, D. and Carlsen, C. H. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly* 15: 3–15. https://doi.org/10.1080/15434303.2016.1261350.

Dubeau, J. (2006). *Are We All on the Same Page? An Exploratory Study of OPI Ratings Across NATO Countries Using the NATO STANAG 6001 Scale*. Ottawa, ON, Canada: Carleton University.

ETS. (2010). *Linking TOEFL iBT TM Scores to IELTS® Scores – A Research Report*. Princeton, NJ: Educational Testing Service.

———. (2018). *Test and Score Data Summary for TOEFL iBT® Tests. January 2017 – December 2017*. Princeton, NJ: Educational Testing Service.

———. (2019). *Performance Descriptors for the TOEFL iBT® Test*. Princeton, NJ: Educational Testing Service.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal* 66: 477–85. https://doi.org/10.1093/elt/ccs037.

Fulcher, G. (1996). Invalidating validity claims for the ACTFL oral rating scale. *System* 24: 163–72.

———. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly* 1: 253–66. https://doi.org/10.1207/s15434311laq0104_4.

———. (2012). Scoring performance tests. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. London and New York: Routledge, 378–92.

———. (2015). *Re-examining Language Testing: A Philosophical and Social Inquiry*. London and New York: Routledge.

———. (2016). Standards and frameworks. In D. Tsagari and J. Banerjee (eds.), *Handbook of Second Language Assessment*. Berlin: De Gruyter Mouton, 29–44.

Fulcher, G., Davidson, F. and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing* 28: 5–29. https://doi.org/10.1177/0265532209359514.

Gaber, S., Cankar, G., Umek, L. M. and Tašner, V. (2012). The danger of inadequate conceptualisation in PISA for education policy. *Compare: A Journal of Comparative and International Education* 42: 647–63. https://doi.org/10.1080/03057925.2012.658275.

Galaczi, E. D., ffrench, A., Hubbard, C. and Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach. *Assessment in Education: Principles, Policy & Practice* 18: 217–37. https://doi.org/10.1080/0969594X.2011.574605.

Glisan, E. W. (2012). National standards: Research into practice. *Language Teaching* 45: 515–26. https://doi.org/10.1017/S0261444812000249.

Green, A. (2018). Linking tests of English for academic purposes to the CEFR: The score user's perspective. *Language Assessment Quarterly* 15: 59–74. https://doi.org/10.1080/15434303.2017.1350685.

Green, R. and Wall, D. (2005). Language testing in the military: Problems, politics and progress. *Language Testing* 22: 379–98. https://doi.org/10.1191/0265532205lt314oa.

The Guardian. (2014). OECD and Pisa tests are damaging education worldwide – Academics. *The Guardian*. Retrieved from www.theguardian.com/education/2014/may/06/oecd-pisa-tests-damaging-education-academics.

Hamid, M. O. and Hoang, N. T. H. (2019). Humanising language testing. *TSL-EJ* 22.

Harvard University. (2019). *International Applicants*. Harvard College. Retrieved from https://college.harvard.edu/admissions/apply/international-applicants.

Hatto, P. (2010a). *Standards and Standardisation. A Practical Guide for Researchers*. Brussels: European Commission.

———. (2010b). *Standards and Standardization Handbook*. Brussels: European Commission.

Hudson, T. (2012). Standards-based testing. In G. Fulcher and F. Davidson (eds.), *The Routledge Handbook of Language Testing*. London and New York: Routledge, 479–95.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal* 91: 663–67. https://doi.org/10.1111/j.1540-4781.2007.00627_5.x.

Hyatt, D. (2013). Stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *Journal of Further and Higher Education* 37: 844–63. https://doi.org/10.1080/03098770X.2012.684043.

Hyatt, D. and Brooks, G. (2009). Investigating stakeholders' perceptions of IELTS as an entry requirement for higher education in the UK. *IELTS Research Reports* 10, 17–68.

IEA. (2019). *About IEA*. Retrieved from www.iea.nl/.

IELTS. (2019). *About IELTS USA*. IELTS. Retrieved from www.ielts.org/usa/about-ielts-usa.

International Organization for Standardization. (2016). *How to Write Standards*. Geneva: International Organization for Standardization.

———. (2018). *International Standards & Trade Agreements*. Geneva: World Standards Cooperation.

Jin, Y., Wu, Z., Alderson, C. and Song, W. (2017). Developing the China standards of English: Challenges at macropolitical and micropolitical levels. *Language Testing in Asia* 7: 1. https://doi.org/10.1186/s40468-017-0032-5.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement* 50: 1–73. https://doi.org/10.1111/jedm.12000.

Kaulfers, W. V. (1944). Wartime development in modern-language achievement testing. *The Modern Language Journal* 28: 136–50. JSTOR. https://doi.org/10.2307/317331.

Krumm, H.-J. (2007). Profiles instead of levels: The CEFR and its (Ab)uses in the context of migration. *The Modern Language Journal* 91: 667–69. https://doi.org/10.1111/j.1540-4781.2007.00627_6.x.

Liskin-Gasparro, J. E. (1984). The ACTFL proficiency guidelines: Gateway to testing and curriculum. *Foreign Language Annals* 17: 475–89. http://dx.doi.org.kuleuven.ezproxy.kuleuven.be/10.1111/j.1944-9720.1984.tb01736.x.

———. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: A brief history and analysis of their survival. *Foreign Language Annals* 36: 483–90. https://doi.org/10.1111/j.1944-9720.2003.tb02137.x.

Liss, J. M. (2013). Creative destruction and globalization: The rise of massive standardized education platforms. *Globalizations* 10: 557–70. https://doi.org/10.1080/14747731.2013.806741.

Little, D. (2007). The common European framework of reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal* 91: 645–55. https://doi.org/10.1111/j.1540-4781.2007.00627_2.x.

———. (2019). Proficiency guidelines and frameworks. In J. Schwieter and A. Benati (eds.), *The Cambridge Handbook of Language Learning*. Cambdrige, UK: Cambridge University Press, 550–74.

Massachusetts Institute of Technology. (2019). IELTS Requirement for International Students. *MIT Media Lab*. Retrieved from www.media.mit.edu/posts/ielts-requirement-for-international-students/.

McNamara, T. (2014). 30 years on – Evolution or revolution? *Language Assessment Quarterly* 11: 226–32. https://doi.org/10.1080/15434303.2014.895830.

Ministry of Education of the People's Republic of China. (2018). *National Language Standard. China's Standards of English Language Ability*. (GF 0018–2018). Ministry of Education of the People's Republic of China and National Language Commission of the People's Republic of China.

Mullis, I. V. S., Martin, M. O., Foy, P. and Hooper, M. (2017). *PIRLS 2016. International Results in Reading*. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement.

NATO. (2014). *Standardization Agreement STANAG 6001 Language Proficiency Levels*. Brussels: Bureau for International Language Coordination.

Negishi, M. and Tono, Y. (2016). An update on the CEFR-J project and its impact on English language education in Japan. In C. Docherty and F. Barker (eds.), *Language Assessment for Multilingualism. Proceedings of the ALTE Paris Conference, April 2014*. Cambridge: Cambridge University Press, 113–34.

North, B. (2014a). *English Profile Studies. The CEFR in Practice*, vol. 4. Cambridge University Press. www.cambridge.org/gb/cambridgeenglish/catalog/teacher-training-development-and-research/cefr-in-practice.

———. (2014b). Putting the common European framework of reference to good use. *Language Teaching* 47: 228–49. https://doi.org/10.1017/S0261444811000206.

North, B. and Piccardo, E. (2018). *Aligning the Canadian Language Benchmarks (CLB) to the Common European Framework of Reference (CEFR). Research report*. Ottawa, ON: Centre for Canadian Language Benchmarks.

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris, France: OECD Publishing.

———. (2019). *The Survey of Adult Skills: Reader's Companion, Third Edition*. Paris, France: OECD Publishing.

O'Loughlin, K. (2008). The use of IELTS for university selection in Australia. *IELTS Research Reports* 8: 145–241.

———. (2013). Developing the assessment literacy of university proficiency test users. *Language Testing* 30: 363–80. https://doi.org/10.1177/0265532213480336.

Pearson (2019). *Score Comparison vs Other Tests for Researchers*. PTE Academic. Retrieved from https://pearsonpte.com/organizations/researchers/score-comparison-vs-competitors/.

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal* 73: 197–206. https://doi.org/10.1093/elt/ccz006.

Peirce, B. N. and Stewart, G. (1997). The development of the Canadian language benchmarks assessment. *TESL Canada Journal/ La Revue TESL Du Canada* 14: 17–31.

Princeton University. (2019). *International Students*. Princeton University Admission. Retrieved from https://admission.princeton.edu/how-apply/international-students.

Riazi, M. (2013). Concurrent and predictive validity of Pearson Test of English Academic (PTE Academic). *Papers in Language Testing and Assessment* 2: 1–27.

Ricardo-Osorio, J. G. (2008). A study of Foreign language learning outcomes assessment in U.S. Undergraduate education. *Foreign Language Annals* 41: 590–610. https://doi.org/10.1111/j.1944-9720.2008.tb03319.x.

Rocca, L., Carlsen, C. H. and Deygers, B. (2019). *Linguistic Integration of Adult Migrants: Requirements and Learning Opportunities*. Report on the 2018 Council of Europe and ALTE survey on language and knowledge of society policies for migrants. Council of Europe.

Sarich, E. (2012). Accountability and external testing agencies. *Language Testing in Asia* 2: 26. https://doi.org/10.1186/2229-0443-2-1-26.

Singer, J. D., Braun, H. I. and Chudowsky, N. (eds.). (2018). *International Education Assessments. Cautions, Conundrums, and Common Sense*. Washington, DC: National Academy of Education.

Sjøberg, S. (2015). PISA and global educational governance – A critique of the project, its uses and implications. *Eurasia Journal of Mathematics, Science and Technology Education* 11: 111–27. https://doi.org/10.12973/eurasia.2015.1310a.

Sollenberger, H. E. (1978). Development and current use of the FSI Oral Interview Test. In J. L. D. Clark (ed.), *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service, 1–13.

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford, UK: Oxford University Press.

Stanford University. (2019). *Exam Requirements for International Applicants*. Stanford Graduate Admissions. Retrieved from https://gradadmissions.stanford.edu/applying/starting-your-application/required-exams/exam-requirements-international-applicants.

Stein, Z. (2016). *Social Justice and Educational Measurement*. London and New York: Routledge.

Takala, S., Erickson, G. and Figueras, N. (2013). International assessments. In *The Companion to Language Assessment*. John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9781118411360.wbcla052/abstract.

Taylor, L. (2004). IELTS, Cambridge ESOL examinations and the common European framework. *Cambridge English: Research Notes* 18: 2–3.

Trim, J. L. M. (2012). The common European framework of reference for languages and its background: A case study of cultural politics and educational influences. In M. Byram and L. Parmenter (eds.), *The Common European Framework of Reference: The Globalisation of Language Education Policy*. Buffalo, NY: Multilingual Matters, 14–36.

UNESCO. (2018). *Global Education Monitoring Report 2019: Migration, Displacement and Education – Building Bridges, Not Walls*. Geneva, Switzerland: UNESCO.

University of California. (2019). *English Language Proficiency (TOEFL/IELTS)*. UC Admissions. Retrieved from https://admission.universityofcalifornia.edu/admission-requirements/international-applicants/english-language-proficiency-toefl-ielts.html.

University of Cambridge, L. (2017). *English Language Requirements* [Text]. Retrieved from www.undergraduate.study.cam.ac.uk/international-students/english-language-requirements.

University of Chicago. (2019). *English Proficiency Testing*. College Admissions. Retrieved from http://collegeadmissions.uchicago.edu/apply/first-year-applicants/international-applicants/english-proficiency-testing.

University College London. (2018). *English Language Entry Requirements*. International Students. Retrieved from www.ucl.ac.uk/prospective-students/international/applying-ucl/english-language-entry-requirements.

University of Oxford. (2019). *English Language Requirements*. University of Oxford. Retrieved from www.ox.ac.uk/admissions/undergraduate/applying-to-oxford/for-international-students/english-language-requirements?wssl=1.

van Ek, J. A. (1975). *Systems Development in Adult Language Learning: The Threshold Level in a European-Unit/Credit System for Modern Language Learning by Adults*. Strasbourg, France: Council of Europe.

van Ek, J. A. and Trim, J. L. M. (1991a). *Threshold Level 1990*. Strasbourg, France: Council of Europe.

———. (1991b). *Waystage 1990*. Strasbourg, France: Council of Europe.

———. (2001). *Vantage*. Cambridge, UK: Cambridge University Press.

Weigle, S. C. and Malone, M. M. (2016). Assessment of English for academic purposes. In K. Hyland and P. Shaw (eds.), *The Routledge Handbook of English for Academic Purposes*. Routledge, 165–77. Retrieved from www.book2look.com/book/95iF77Y6oe.

Weir, C. J. (2003). A survey of the history of the certificate of proficiency in English (CPE) in the twentieth century. In C. J. Weir and M. Milanovic (eds.), *Continuity and Innovation: The History of the CPE 1913–2002*. Cambridge, UK: Cambridge University Press, 1–56.

Wu, R. (2014). *Validating Second Language Reading Examinations: Establishing the Validity of the GEPT Through Alignment with the Common European Framework of Reference*, vol. 41. Cambridge, UK: Cambridge University Press.

Yale College. (2019). *Applying to Yale as an International Student*. Yale College Undergraduate Admissions. Retrieved from https://admissions.yale.edu/applying-yale-international-student.

Zhao, W., Wang, B., Coniam, D. and Xie, B. (2017). Calibrating the CEFR against the China standards of English for college English vocabulary education in China. *Language Testing in Asia* 7: 5. https://doi.org/10.1186/s40468-017-0036-1.

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change* 21: 245–66. https://doi.org/10.1007/s10833-019-09367-x.

35

Quality management in test production and administration

Nick Saville and Sarah McElwee

Introduction

This chapter presents and discusses the core processes required in developing and administering language tests and makes the case for adopting a quality management (QM) approach to improve testing and to ensure that appropriate professional standards are met. In this respect, language test developers need to adopt the kind of managerial practices which enable successful organisations to implement error-free processes.

In making this case, it is important to define what quality means in the context of language assessment and to describe how an approach based on QM can help achieve the required quality goals and at the same time to enhance a validity argument (see Kane, this volume). By adopting that QM approach, it is possible to ensure that processes are continually improved and standards raised; this is in keeping with the concept of validation as a long-term, ongoing activity (Kane, 2006).

A key point is the necessary convergence of the twin concepts of quality and validity. We argue that QM provides an appropriate basis for guaranteeing the consistency of the processes which underpin a validity argument and the tools and techniques for linking theoretical concerns with the practical realities of setting up and administering assessment systems. We reference particular considerations linked to assessment in digital contexts and the implications for QM approaches in this regard.

The principles of QM and the link with validation can be widely applied in many different test development contexts with the aim of improving the quality of assessment systems. A number of explicit stages need to be followed from the production of the test materials to the issuing of the test results. This applies whether the test development team is a group of teachers producing language tests for students within their own classes or specialised staff within a testing agency responsible for developing and administering high-stakes, international language examinations in many countries.

To illustrate the applicability of the approach, these two contexts – classroom assessment and large-scale assessment – will be referred to throughout this chapter. In both cases, the validity of the assessment needs to be ensured and the integrity of the results guaranteed for the intended purposes, even if these may be very different. To achieve the required validity

DOI: 10.4324/9781003220756-46

profile for the intended purpose and use, the quality of the development and administration procedures has to be checked and monitored at each stage of the assessment cycle (see Figure 35.1). What differs between the school context and the large-scale assessment operation is the complexity of the organisational structure within which the activities take place and, most obviously perhaps, the resources available for achieving the intended assessment goals: i.e., the human, material, and time-related factors which are available to meet the intended objectives. Other factors may relate to the intended uses of the test results and the importance attached to decisions made about the test takers. In other words, whether the test is considered to be high or low stakes by the users and whether this may offer opportunities to increase the resources available (e.g., in order to achieve adequate reliability of test scores to support high-stakes decisions about test takers).

Practical considerations are always important. The test developer must attempt to match the ambitions of the assessment sponsors to the "real world" considerations and constraints which exist in their context. This inevitably requires judgments to be made about deploying the available resources in the most effective way at all stages of the cycle. Careful planning is required at the outset, and appropriate management systems need to be set up to allocate and manage resources throughout. Leadership and communication skills are important, and those involved in a test development project need to employ techniques for managing successful projects and to be aware of critical issues related to the management of change within organisations. Furthermore, as assessment organisations move to computer-based or digital assessments and to increasingly sophisticated ways of using learner data for insights into the learning journey, adaptations to traditional ways of working and current assessment processes may be required to accommodate new modes of test design and delivery (see also Sawaki, this volume).

Bachman and Palmer (1996: 35–37) discuss practicality as an important "test quality" and dedicate a chapter to the identification, allocation, and management of resources Bachman and Palmer (2010: 411–27). These points are expanded later in the chapter when considering aspects of organisational management.

Definition of terms

In the recent literature on quality, a distinction is usually made between the detailed procedures for checking and assuring the quality of a product or service (*quality control* and *quality assurance* – QC and QA) and the overarching concept of QM, which is concerned with the management of *processes* within the organisation responsible for developing the product or service. These distinctions are maintained in this chapter, and the history of QM systems will be reviewed to illustrate how the strands have developed and why they are now relevant to language assessment systems. The finer-grained distinction between QC and QA is addressed later.

Historical perspectives

Language assessment

Interest in developing high-quality language tests for practical purposes has increased since the 1980s, and a large number of textbooks have been written to provide guidance to aspiring practitioners (Hughes, 2003; Alderson *et al*., 1995; Bachman and Palmer, 1996, 2010; Weir, 2005). This has also been reflected to some extent in the literature on educational and

psychological measurement; see, for example, Downing and Haladyna (2006) and Lane, Raymond and Haladyna (2016).

A common feature of these textbooks is that the process of test development is conceived of as a series of logical steps, starting with the test design and item writing, progressing through to administration and reporting of results, and ending with some form of evaluation. We refer to this as *the assessment cycle*.

Bachman and Palmer (1996), for example, set out to "enable the reader to become competent in the design, development and use of language tests" (1996: 3), and in their approach to "test usefulness," they outline a three-stage model: *design, operationalisation*, and *test administration*. In their updated edition titled *Language Assessment in Practice* (Bachman and Palmer, 2010), they switch their focus to the development and support of "assessment use arguments (AUA)." but nevertheless, they still recommend a process-based approach to the assessment cycle which starts with a "design statement" for the intended assessment and its uses and concludes with the language assessments being *used responsibly* in practice.

Lane, Raymond, Haladyna and Downing (2016: 3–18) discuss a framework for test development and validation comprising 12 related components. The aim is to provide a

> high-level overview of the test development process . . . [which] can be used to provide a framework for collecting and organizing evidence to support the psychometric quality of the test and the validity of the test score interpretations and uses.
>
> *(op. cit.: 3)*

It is less common in the testing textbooks to find a discussion of the institutional parameters which need to be addressed in order to set up and manage an organisation which is capable not only of designing good quality test materials but also of maintaining an assessment system over an extended period of time.

In rare instances when the operational management of a testing system is covered in the literature, it tends to be discussed in relation to test *administration*, which is taken as only one part of the overall assessment cycle. For example, in their chapter on administration and training, Fulcher and Davidson (2007: 115) state that "[t]he delivery of tests requires an infrastructure, and a set of procedures that are usually followed," and they go on to discuss "getting things done." They make a case for "doing things consistently" and suggest that "[f]or language testing, part of the validity claim is that the test administration and all the processes used by the testing agency are done according to standardized procedures" (Fulcher and Davidson, 2007: 127). In the same section of their chapter, they briefly refer to "quality management systems" as a way of ensuring that consistency and dependability can be achieved within a company by *defining*, *controlling*, and *evaluating* what gets done.

The role of the "testing agency" (whether this is interpreted as the team of school teachers or the large-scale testing company) and "the processes" required to manage a valid assessment system are two key points. Although not widely discussed, the organisational structure and the processes employed by those responsible for developing and administering tests (the staff) are essential elements in a quality model. Both of these aspects are discussed in more detail later in this chapter.

In the literature on educational and psychological measurement, attention has recently been given to the concept of quality. An important contribution is the volume titled *Improving Testing: Applying Process Tools and Techniques to Assure Quality* edited by Wild and Ramaswamy (2008). This volume is structured around the notion of a quality model and the so-called "*quality triangle*," the three sides of which are planning and design, monitoring

and improvement, and implementing standards. The next section deals with these points within the general discussion of QM.

Quality management

In taking an historical perspective on QM, it is informative to start by reviewing developments in the manufacturing sector, especially in production processes and in the development of approaches to organisational management within companies. However, although the quality movement originated in manufacturing, the same approach is now employed across many types of organisations, including those which provide services rather than products. The developments in the service sector have also been extended to educational systems and to educational assessment, as described by Wild and Ramaswamy (op. cit.). A case is made in this chapter for applying the model in language assessment too.

The concept of quality may appear to be a modern development, but practices related to consumer protection, standards, and inspection have been with us for hundreds of years: e.g., through the development of trades and guilds in the Middle Ages. However, the industrial and scientific advancements in the past 200 years shifted the focus from *maintaining* standards to *improving* standards and the central notion of *continual improvement* emerged. This has now become axiomatic within all QM systems.

The founders of this field – for example, Taylor (1911) and Shewhart (1931, 1939) – emphasised the link between effective management of the organisation (e.g., the company producing the products) and the quality of the products themselves (the outputs of the company). They saw management as a science and began to show how systems could be experimented with and improved in order to get better results. Processes were defined in terms of their *inputs* and *outputs*, and this in turn enabled the *links* between different parts of an organisation to be seen more clearly and, as a result, to be managed more effectively. Furthermore, the high cost of relying on inspections to check on quality post hoc, with the subsequent reworking of a product to ensure its quality, became apparent. The recommended solution was for more effort to go into reducing the number of defective items *before* the product reached the final stage of the process.

W. Edwards Deming (1900–1993) emerged in the mid-twentieth century as the leading figure in this movement; he built on these evolving views of quality to create one of the first coherent "quality management systems," known as Total Quality Management (TQM) (Deming, 1986).

Deming's influence is often illustrated by the Japanese experience, in which his ideas were taken up wholeheartedly in the attempt to rebuild Japanese industry in the 1950s in the aftermath of the Second World War. By the 1960s, industrial companies which implemented TQM, such as Sony, Toyota, and Hitachi, had already achieved a global reputation for the consistently high quality of their products. Competitors in other parts of the world were obliged to take note.

Along with the development of quality as a concept within organisational management, the notion of *quality standards* also emerged as a mechanism of accountability and as the guarantee of quality for consumers. This notion also has a long history and dates back over 100 years to the establishment of the first manufacturing standards with external accreditation. One of the earliest examples is the British Standards Institute (BSI), which was founded as the Engineering Standards Committee in London in 1901. The original quality

mark, known as the BSI "kite mark" because the design of the logo resembles a kite, was introduced in 1903.

BSI defines a standard as "an agreed, repeatable way of doing something. It is a published document that contains a technical specification or other precise criteria designed to be used consistently as a rule, guideline or definition" (BSI, 2012).

Most recently, an important international influence has come from the International Organization for Standardization (ISO) and the ISO quality standards, such as the ISO 9000 range (see later in this chapter). It is important to note that this kind of accreditation is not primarily designed to guarantee the quality of the end product or service; the intention is to certify that processes are being applied consistently and effectively to meet the stated objectives of an organisation. We will see next how the approach to QM represented by ISO 9000, combined with the appropriate principles of good practice in language assessment, provide the basis for auditable processes and the achievement of professional standards: for example, as set out by the AERA/APA/NCME Standards for Educational and Psychological Testing (2014) or in the international codes of practice for language assessment (see the references section later in this chapter).

When adopted and implemented in conjunction with one another, these two elements can help ensure that assessment systems are developed and administered to meet the demanding requirements of regulatory authorities and, at the same time, to provide the basis for fair testing practices which meet the needs of the language test users.

In the rest of this chapter, the ISO approach will be used as an example of how standards form part of a QM system. The ISO itself is a non-governmental organisation (NGO) founded in 1947 with headquarters in Switzerland. It is composed of representatives from various national organisations concerned with standards of the kind noted earlier. The standards, such as the 9000 range, are maintained centrally by ISO but are administered by the accreditation and certification bodies at a national or regional level (e.g., American National Standards Institute (ANSI), Deutsches Institut für Normung (DIN), BSI, Canadian Standards Association, Standards Australia, etc.).

The ISO 9000 group of standards is particularly relevant to organisations involved in educational systems, including those responsible for test development and administration. The requirements for accreditation are updated periodically by ISO, so, for example, the requirements for ISO 9001, one of the standards in the ISO 9000 group, were updated in 2008 and again in 2015. ISO 9001: 2015 is now the world's most widely used QM standard, and it is particularly applicable to our discussion because it can be applied to all types of organisations, no matter what the size. Its stated purpose is to provide a framework for any organisation to control its processes in order to achieve its objectives, including customer satisfaction, regulatory compliance, and continual improvement.

Any organisation which aspires to meet the ISO standard can apply to be independently audited and certified in conformance with the standard in question (e.g., ISO 9001: 2015). Once the organisation has been accredited, it can claim to be "ISO 9001 certified" or "ISO 9001 registered."

The ISO approach exemplifies clearly the underlying principles of QM with the aim of continual improvement. While compliance with external standards and external accreditation are benefits that accrue, the main focus is on meeting objectives and getting the job done properly. The principles and the techniques to assure quality can be applied even within a small-scale testing operation such as a school or university department.

Critical issues

Definition of quality management

QM is an overarching term; it includes the planning and management of processes, which, over time, lead to improvements being implemented (c.f. the key concept of continual improvement noted earlier).

In organisational terms, the aim should be for quality to become synonymous with what the organisation does – normally referred to in QM discourse as the organisation's "mission." A prerequisite is a clear understanding and statement of the organisation's aims and objectives. This means that leadership and good management underpin a quality approach. ISO makes this clear in their description of seven quality management principles: customer focus, leadership, engagement of people, process approach, improvement, evidence-based decision making, and relationship management (ISO 9001: 2015, 2019). While good leadership is named explicitly as a crucial element of quality, good management and leadership practice are also inherent in the successful implementation and execution of the other principles.

The Chartered Quality Institute notes:

> First and foremost, senior management must be committed to quality. Once that is the case . . . three things underpin successful quality management: effective governance that defines the organisation's aims and translates them into action, robust systems of assurance to make sure things stay on track, and a culture of improvement to keep getting better.
>
> *(retrieved from CQI website 2019)*

Heyworth (2013) discusses the application of QM in language education (as opposed to language assessment specifically) and outlines a number of concerns and objections regarding the appropriateness of QM approaches in language teaching institutions. He cites

> an over-concentration on the summative, label-granting aspect of QM and much less on the formative ways of helping the institution to change and improve. The approach can lead to a checklist mentality, in which the existence of a feature in the list of indicators takes precedence over its content and quality, with the result that the most important features of education are neglected.
>
> *(2013: 296)*

The "snapshot" nature of quality audits is also criticised, with the implication that procedures and standards may slip in the interim. These concerns highlight a critical point about the motivation for adopting a QM approach – a major driver must be an internal desire, modelled and supported by an organisation's leaders, to embed quality processes in the day-to-day management of both new and business-as-usual tasks. Many of the objections that Heyworth addresses are related to the external element of QM accreditation, rather than the internal examination of the consistency and robustness of processes in a quest for improvement.

The effective achievement of organisational objectives will, of course, vary, depending on the type of organisation. A school which provides language assessment as part of its educational mission will have different objectives than a testing company which specialises

in large-scale assessment. From the assessment perspective, both types of organisation need to be able to guarantee "fitness for purpose" or, in Bachman and Palmer's usage, to develop and administer "useful" tests (1996: 17).

Customer or client satisfaction is another key concept here. Products or services which can be considered "high quality" adequately meet or exceed the expectations of the customers or clients who buy or make use of those products or services. Identification of customer/client needs is therefore an important initial consideration.

In language testing, the main clients are the test users, the *primary users* are the test takers themselves, and the *secondary users* are those who sponsor the test takers (parents, teachers, employers, etc.) or make use of the test results to make decisions (educational institutions, employers, government bodies, etc.). For discussion of the terms *primary* and *secondary user*, see the Association of Language Tester in Europe's (ALTE's) Code of Practice (1994) that was based on the original Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 1988/2005).

Quality control and quality assurance

QM encompasses quality control and quality assurance activities, which have related but discrete functions.

The broad meaning of quality control is synonymous with checking. In other words, it entails carrying out systematic checks to ensure that all aspects of a workflow meet the stipulated standards as defined and described in relevant documentation.

Quality control activities involve the detailed checking of the components which make up a product and service *at each stage* of the overall process. The checks need to be carried out by the people who have the responsibility for doing the tasks. Quality control activities, therefore, form an integral part of a person's job and are not something carried out by a separate "quality control department."

In the assessment cycle, these kind of checks are essential to ensure that basic mistakes are not made which could threaten validity (e.g., introduce construct-irrelevant variance into scores), disrupt the smooth administration of the assessment, or lead to mistakes such as the wrong result being issued to a test taker.

Quality checks may focus on important details such as:

- the quality of the tests materials themselves – in other words, the test items and other features of the assessment procedures;
- the quality of the information and support provided to the clients (the test takers, teachers, and other users);
- the quality of the documentation needed to support the administration of the tests at the testing venues;
- the quality of the data collected and stored for assessment purposes and for producing the test results.

Quality assurance activities, on the other hand, are carried out to monitor, evaluate, and improve processes; the aim is to ensure that all processes are maintained to the required standard.

Quality assurance activities differ in that they are normally carried out in order to monitor and adjust processes when they are operational within a system. Procedures for monitoring need to be put in place to evaluate whether core processes are working adequately (see core

processes in Figure 35.1). It is crucial that relevant evidence (i.e., data and information) can be collected through QA mechanisms, such as periodic audits, inspections, and formal reviews so that the evaluation can take place. The required standard which is quality assured may be set by:

1 an external regulatory body, such as a government agency;
2 a professional organisation responsible for a specific domain of activity (e.g., law, medicine, educational assessment, etc.);
3 an external standards organisation (e.g., ISO);
4 the organisation itself and its internal rules and regulations.

In some cases related to 1 or 2 on this list, the need to demonstrate compliance may be a statutory requirement. Voluntary accreditation is normally associated with 3 and 4. In a QM system, it is important to recognise, however, that quality cannot be imposed from outside but must be established and monitored within an organisation in the first instance. Detailed illustrations of QC and QA activities in language test development and administration are given next.

Quality management within an organisation

Let us return for a moment to the nature of organisations and the two extreme examples introduced earlier: the classroom teachers in a school context and the staff within a testing agency. In both cases, the assessment cycle needs to be effectively managed, and all those involved need to have a clear understanding of the objectives and how they will be achieved.

The people involved in the test development and administration processes (the employees) need to understand their roles and how to operate effectively within their own institutional context. The identification and allocation of the non-human resources available to get the job done also need to be managed effectively, and the time frame for carrying out the processes must be clearly communicated to all concerned.

The importance of employee skills and team development

The lessons to be learnt from QM suggest that employees are a key asset in assuring quality and that senior managers within an organisation must take responsibility for ensuring that each member of their team is qualified to carry out their own responsibilities. Support for the recruitment and training of qualified specialists, as well as effective communication across teams, is central to this.

Given the cost of recruiting and training employees, successful organisations seek to retain those who have acquired relevant skills and to provide regular in-service training for them to ensure that up-to-date knowledge is maintained. For example, it will be important for any organisation involved in language assessment to maintain a good level of assessment literacy throughout the organisation, as well as specialist knowledge for carrying out specific tasks, such as item writing, rating, and test analysis.

However, many organisations involved in developing assessment systems cannot afford to retain *all* the expertise they need on their permanent staff, and so they need to contract temporary employees or specialised consultants to carry out some tasks within their

processes. This might include the following kinds of task: item writing, rating, rater training, professional support coordination for teachers, administration at test venues (as the proctors or invigilators), and so on.

As these tasks play an essential part in core processes, the assessment provider has a responsibility to ensure that these external consultants are also appropriately recruited and trained. Relevant quality control and quality assurance procedures can be used to guarantee that their contributions meet specified quality standards and that they contribute to the overall quality of the assessment system. The monitoring and evaluation of the work of staff and consultants over time can be used to identify threats to validity, as well as to identify where processes can be made more efficient. Where necessary, staff changes can be made or additional training provided. See Saville and Hargreaves (1999: 42–51) for a discussion of rater training and monitoring systems in language assessment; also Fulcher and Davidson (2007: 129–32) on administration and training.

Setting up the structure for managing employees and monitoring their performance is the responsibility of senior managers, but the QM model also requires *all colleagues* within an organisation to take responsibility for their *own* role and the contribution they make to quality. They should seek to gain a sound understanding of how their role fits within the overall organisation and its stated objectives (its mission and goals).

Individuals must be able to connect their own jobs to those of others working in their team or work group, and they should identify professional development and training requirements in order to improve their knowledge and skills over time.

Linking quality management to the assessment cycle

In this section, we will look in more detail at how the concepts of QM can be directly linked to the assessment cycle and at ways that test development projects can be effectively managed by an organisation. Specific quality control and quality assurance techniques are illustrated in relation to the *core processes* within the assessment cycle. In doing so, the core processes are distinguished from *wider processes* within an organisation which support the assessment activities and the *support processes* which are needed to manage the organisation as a whole.

Wider processes include activities such as document control and record keeping, internal audits and liaison with external regulators and auditors, conformance to relevant statutory regulations and standards (e.g., equal opportunities, health and safety, data protection, financial regulation, etc.), risk management and continuity planning (e.g., contingency for when things go wrong), etc.

Support processes within a company include recruitment and training of all staff (including administrators, cleaners, etc.); finance systems, including payment of salaries, purchase payments, issuing of invoices, collection of fees and credit control; technical support services related to IT and communications; project management; and so on.

Although many of these wider support processes are best demonstrated in large-scale assessment providers (such as testing companies), many are also relevant in other contexts in which tests are developed and administered by smaller organisations or groups of individuals (e.g., in schools, university departments, etc.). In order to ensure the quality of core processes, appropriate attention needs to be given to all these areas as they influence important considerations, such as resource allocation and management, which affect the quality of the assessment system.

The importance of defining processes

We return here to the language testing frameworks which we referred to earlier and which address practical test development issues within the assessment cycle. From among the many suggested frameworks and models, five main steps or stages can be identified in a basic assessment cycle:

- Planning and design following a decision to develop a test
- Development, including initial development and systems for test assembly
- Delivery, including routine test assembly and the administration of the assessment
- Processing, including marking and grading and issue of results
- Review and evaluation

These are the *core processes* commonly found in the handbooks on testing (although the names given to them may vary), and they represent the day-to-day reality for any organisation which undertakes to develop and administer assessment systems.

The convergence of QM and the principles of language assessment can be illustrated through the assessment cycle shown in Figure 35.1.

We have already noted that defining processes is a crucial part of a QM approach. It can be argued that it is only when the processes are adequately defined and appropriately documented that the necessary quality control and quality assurance procedures can be carried out. The QC and QA procedures ensure that the defined processes are being followed effectively and efficiently and provide the necessary information to make improvements.

Figure 35.1 shows five *core processes* and four *outputs* in the assessment cycle. In QM terminology, this is the central part of a *process interactions map*, which shows how the core processes fit together and enable an organisation to manage the interaction between them to meet its objectives.

The core processes (design, production, administration, processing, review, and evaluation) are supplemented by the wider support processes noted earlier. These vary considerably depending on the nature of the organisation within which the core processes are being implemented.

The outputs which the core processes generate are:

- Specifications
- Assessment materials and procedures

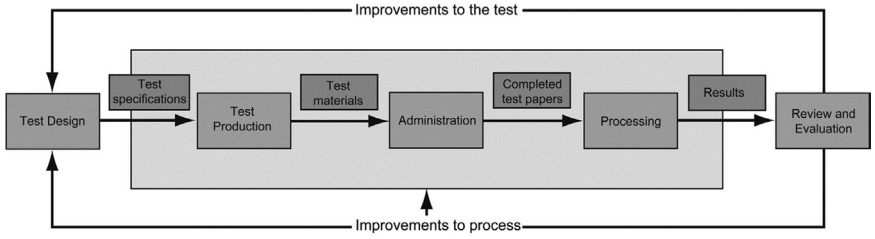


Figure 35.1 Core processes and outputs of the assessment cycle

- Completed assessments (test takers' responses)
- Results and interpretive information

In implementing this model, the relevant processes are defined by senior managers, who can then coordinate the links between those processes as effectively as possible and ensure that the staff are capable of carrying out their tasks effectively.

To achieve this within a QM model, *process owners* are appointed to ensure that the critical interactions on the map are identified and documented appropriately. Each core and support process should be described in a process document, and each sub-process is described in a *standard operating procedure (SOP)*. Ultimately, each task to be carried out should be described in detailed *work instructions*, which must be explicit enough to ensure that the activity can be performed without mistakes being made and to meet the quality standard.

The documents, such as the SOPs and work instructions, need to be stored so that they can be readily accessed by members of staff and kept up to date when improvements are made to the processes. In assessment systems, some of these documents may need to be stored so that only certain people can have access to them to meet requirements for security and confidentiality.

Within the ISO approach, it is recommended that some documents should be stored in a *central document register*; this applies to documents which are needed by groups of staff across the organisation. In addition, *local document registers* can be compiled to store the documents needed for carrying out particular activities by specific groups of staff. These include the detailed work instructions, manuals, standardised forms, guidelines, checklists, etc. which are particularly important for quality control purposes.

Quality assurance techniques are needed to monitor whether the procedures for document production, document control, and records management are being followed effectively. Procedures such as internal audits can be implemented for this purpose.

The continual improvement cycle is also central to QM, as illustrated by the ISO 9001 Standard in which it is characterised as the *Plan, Do, Check, Act* model. The feedback loop in this model requires action to be taken after the checking takes place to ensure that well-specified improvements are made to the operational system.

Figure 35.2 illustrates this iterative process within the assessment cycle; each iteration allows feedback to be used to evaluate effectiveness and efficiency and to enable improvements to be made in subsequent iterations.

The planning stage begins with the perceived need for a new (or revised) test or assessment system. The aim is to define the intended context and use of the prospective assessment, and the key output of this stage is a set of specifications defining the requirements for the validity argument and its operational procedures (see Jin, this volume). This stage occurs when the assessment system is developed for the first time or is being significantly modified.

Each test version or administration of the test needs to be delivered according to the required quality standards. There are typically three stages in this process: the assembly of test forms; the administration of the test at testing venues; and the processing of test takers' responses, including the rating and grading procedures, to produce and issue the results to users.

As part of the operational system, the quality control and quality assurance procedures noted here should be implemented to ensure that the design features are maintained effectively under operational conditions and that unexpected threats to validity are not introduced.

The aims and outputs of each core process are shown in Figure 35.2.

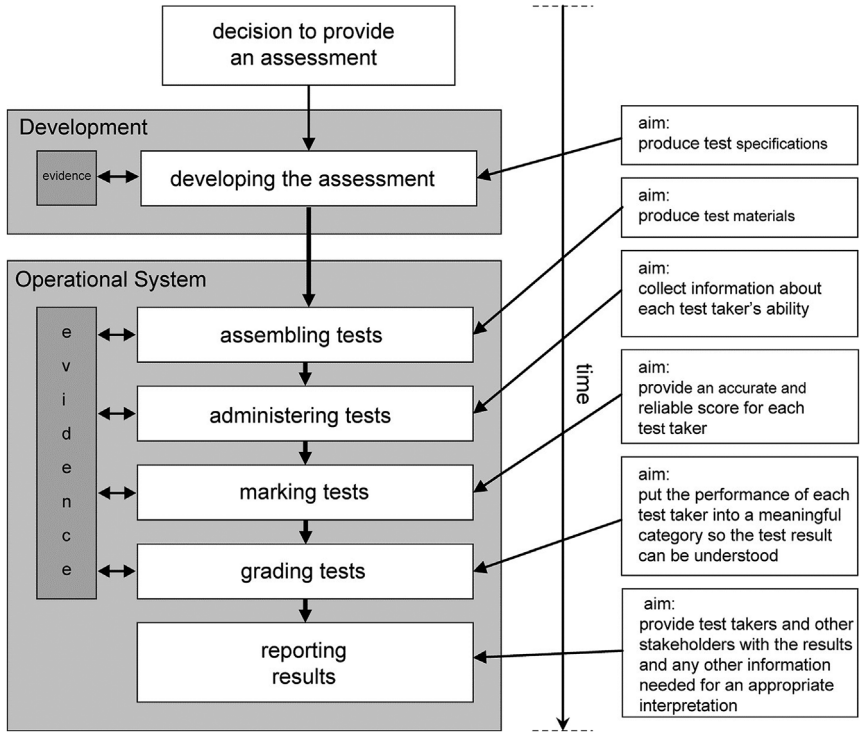


Figure 35.2 The assessment cycle

How does this work in practice?

To illustrate how this can work in practice, in this section, we will look in more detail at QC and QA activities related to two processes in the operational system: assembling tests and test administration.

QC and QA procedures in assembling test materials

At the development stage, the production of test specifications and instructions for item writers which can be used in assembling tests is an essential prerequisite. Checking against the specifications must be done at various stages of the operational cycle before the assessment is finalised and ready to be administered.

To ensure that test materials are controlled for quality, the test provider must set up a system to collect, store, and process the test tasks and associated materials. This is particularly important if a large number of items and tasks are involved.

All test materials should go through standardised quality control processes, which will involve editing and try out (piloting or pre-testing). Technical characteristics such as the level of difficulty of items, any actual or potential bias in the materials, the content coverage of the tasks (topic, genre, etc.), and conformance to the required format of tasks and items (in terms of length, type of item, etc.) must be checked.

It is important to be able to track progress and to know at any time what stage a task or item has reached in this process. The workflow therefore has to be managed efficiently, and, if necessary, appropriate technology should be employed to deal with the complexity. This becomes increasingly important as the number of items and range of people involved increase and will usually entail using item-banking techniques.

A basic system for managing the test materials might include the following:

- an identification number and description of each task or set of materials produced;
- a checklist to track progress and to provide a complete record of the stages completed, the changes made, and other relevant information about decisions taken;
- a system to ensure the stored materials are readily accessible along with relevant management information which is accrued during the editing process.

QC and QA procedures in administering tests

The test developer needs to be confident that assessments are administered in a standardised way. Consistency is important as uncontrolled variation can undermine validity. Standardisation between testing locations, if different venues and groups of people are involved, and over time from one session to the next needs to be guaranteed. Quality control and quality assurance techniques can be used to ensure that the administration of an assessment, whether in a classroom context or as part of a large-scale assessment system, is based on a set of standard operating procedures which are susceptible to monitoring and improvement.

In large-scale operations, the administration of the assessment will be delegated to personnel at the testing venue. This means that the administrative procedures need to be clearly and comprehensively described and produced in a format which can be used under operational conditions.

For example, the following areas should be covered:

- the physical setting – how the venue is selected and how the rooms where the assessments take place are managed, including the safety and security of the premises;
- the storage and handling of secure and confidential materials and information (before, during, and after the event);
- the type and number of personnel needed to oversee the administration on the day of the assessment to ensure safety and guarantee test security;
- the procedures for recruitment, training, and management of invigilators/proctors and examiners, including practical aspects (such as payments for travel and subsistence);
- the management of the interaction between test takers and administrators before, during, and after the assessment takes place – including the arrival at the venue of the test taker, the checking of their identities and eligibility, the seating arrangements, and provision of accommodations to meet special needs or requirements;
- the management of the assessment procedures themselves – providing instructions, handing out and collecting test papers and answer sheets, monitoring for malpractice (including all kinds of cheating), and ensuring that timings are respected;
- the handling of unforeseen eventualities and catering for specific circumstances (illness, fire alarms, power cuts, minor accidents, disruption from outside the venue).

All these procedures need to be carefully planned and fully documented before they are put into practice and must be turned into operational documents, such as instructions, regulations, checklists, etc.

The staff involved must be accountable for carrying out the quality control checks to ensure that each time the assessment is administered at a venue, things run efficiently, and unfair outcomes are avoided.

To ensure that the administrative procedures are implemented effectively at *all venues*, quality assurance techniques also need to be implemented on a wider scale. These might include monitoring visits or inspections carried out on behalf of the assessment provider across a range of venues. This form of monitoring can have a dual function: it can identify and deal with irregularities (non-conformances) where standard procedures are not followed effectively, and it can identify opportunities for improvement.

Large-scale testing agencies can employ data-informed forensic techniques to monitor test taker behaviour and to identify possible cases of malpractice, including cheating. This cannot usually happen in real time as the analysis is based on test-response data (see Foster *et al*., 2008 for a discussion of this).

Quality improvements resulting from QA monitoring might include:

- better conditions for administering listening tests (high-quality digital sound files, improved audio systems, use of headphones, etc.);
- improved security procedures to cut down on fraud, such as identity checks using biometric techniques (test-day photos, etc.);
- better ways to accommodate test takers with special needs (e.g., improved access at venues for wheelchairs).

Quality and validation

Current thinking on validity emphasises the importance of *validation* as a process of accumulating evidence to support inferences made using the results of tests. This accrual of evidence should take place at all stages in the cycle (as shown in Figure 35.2).

Good practice also requires an assessment system to be reviewed regularly based on the monitoring of operational processes. QM methodology suggests that improvements can always be made; they may be small in nature and implemented in an ongoing way, or they may require a major revision project to carry them out. During routine monitoring, the need for a full-scale revision may be identified, and if so, the planning and design stages are repeated. Fulcher and Davidson (2009) use the metaphor of a building to consider those parts of a test which must be changed more regularly and those which may be changed infrequently.

Kane's chapter on validation in the fourth edition of *Educational Measurement* (2006) makes the case for "validity as argument." He suggests that in developing and presenting *a validity argument*, an examination provider must provide an overall evaluation of the intended interpretations and uses of the test or examination which is being validated based on evidence (Kane, 2006: 23). As validation needs to be maintained over time, we see a synergy between test validation and the QM concept of continual improvement.

Kane builds on the seminal paper by Messick (1989) which presented the *unitary concept of validity*, and construct validation remains central to his approach. Validating inferences about a construct requires attention to be paid to all aspects of the assessment cycles. Failure to maintain the quality of any aspect of the system may threaten validity and lead to the

assumptions of the validity argument being undermined. For example, undetected problems within the assessment system may lead to construct-irrelevant variance being introduced which may affect all test takers, groups of test takers, or individuals and thus lead to unfair outcomes (see Kunnan, 2000, 2004).

Staff whose work contributes directly to the development of a validity argument for a particular assessment system have a responsibility to ensure that they understand the concepts involved and that they are able to supply the required evidence from their own area of work. (See earlier in this chapter for discussion of the role of staff in a quality model.)

By adopting the QM approach, test developers are in a better position to build an appropriate validity argument, and the rigour imposed means that they can be held to account by their stakeholders and other interested parties.

Quality, validation, and standards in digital assessment and learning contexts

Early in this chapter, we outlined the growing adoption of computer-based delivery models and digital assessments, as well as the increasingly sophisticated use of data to understand learners and their learning behaviours. It is pertinent to explore how these changes impact QM approaches for assessment providers. Often, the considerations for QM and its link to validation are similar – many of the general principles that are considered good practice in the design and administration of traditional paper-based assessments hold true for digital assessment formats. However, working in digital contexts presents some additional challenges for a testing agency that are worth describing more explicitly. The following sections outline some key considerations for the reader as examples, although these are by no means exhaustive.

Administration of computer-based tests and impact on task types and construct

Since the late 1990s, many research overviews of computer-based language testing have been published, with a range of different focuses, including their potential for innovative assessment formats, the design and evolution of computer-based tests, and the effects of technology on language test reliability and validity (e.g., Chalhoub-Deville, 2001; Dooey, 2008; Jamieson, 2005; Ockey, 2009; cited in Brown, 2016). See Brown (2016) for a summary of drawbacks and benefits related to both the physical (administrative) and performance (broadly, validity) aspects of digital assessment delivery.

A key measure of quality in test administration is replicability – the provision of a standardised experience for all candidates, regardless of where or when they sit their test. For assessment organisations of all sizes, ensuring that the assessment body itself or third-party distributors of their language tests have suitable equipment, internet connection, and staff training to run the assessment to a high standard is important. The impact of variables such as screen size, resolution, scrolling, internet speed, and graphics and their potential to affect candidate performance must be considered (Brown, 2016). On the other hand, the use of automated scoring technologies (see Carr, 2014; Yu and Xu, forthcoming, for useful overviews; Xi, this volume) for productive responses ensures the consistent application of assessment criteria to all candidates, free from the variability inherent in even the most rigorously trained teams of raters. The quality management of computer-based test administration, and therefore its consistency and replicability for all candidates, is a linchpin of validity.

As computer-based delivery becomes more sophisticated, the interface between quality and validity is highlighted in the types of tasks that can be deployed, beyond a straightforward translation of a paper test to on-screen delivery. Innovations such as using videoconferencing to examine speaking (e.g., Nakatsuhara *et al*., 2017) may allow the testing process to be streamlined but require consideration of whether different aspects of the construct need to be emphasised and specified accordingly. As automated marking technology improves, computer-based tests are moving away from a reliance on multiple-choice or cloze items, which mostly test knowledge skills and receptive skills. The development of machine learning approaches to automated marking, for example, allows for longer and more sophisticated productive responses to be assessed quickly and for feedback to be given immediately if appropriate. This enhanced technology can then be applied to innovative formats and integrated task types (e.g., listening into speaking or listening into writing), allowing candidates to respond more appropriately and potentially offering benefits in terms of pedagogy and ecological validity. Development and deployment of new features such as these are supported by the QM processes associated with the assessment cycle outlined in Figures 35.2 and 35.3, which help underpin the validity claims of the assessment.

Diversification of employee teams and knowledge transfer

An important consideration for assessment organisations in transitions to digital assessment is the degree and diversity of expertise needed to sustain a robust, fit-for-purpose testing solution. A move to computer-based delivery ultimately means an increasing reliance on software code, which in turn requires diversification of the resource skillset in the organisation; those completing the coding, for example, are less likely to be subject matter experts. Developing adaptive tests requires interdisciplinary teams, with not only experience of language testing principles and constructs but also extensive expertise in software programming, user experience (UX) and design, database and item bank management, and advanced knowledge of psychometric techniques such as item response theory. New advances in automated marking, particularly for writing and speaking tests, rely on experts in areas including machine learning, natural language processing, and corpus development. The increasing numbers of areas of interface and potential handovers between teams adds to the complexity of the test development system and the importance of established mechanisms for knowledge transfer between teams. Where aspects of this work are outsourced or off-the-shelf solutions integrated into in-house assessment products, rigorous QM procedures must account for the additional management of those ancillary technologies and external partners.

Timing and scope of quality control checks

In 2011, Jon Jenkins, then engineering lead at online retailer Amazon, revealed that on weekdays in a single month that year, the company made a software code deployment on average every 11.6 seconds (O'Reilly, 2011). This statistic highlights both the possible pace of change in digital technology and the complexity of ensuring that quality is maintained, even for organisations working on a much smaller scale.

Paper-based test administrations have a window for results processing, and even where technical solutions such as optical mark readers and advanced data processing code are used to increase the speed of results release, a buffer of time is available for QC checks on aspects

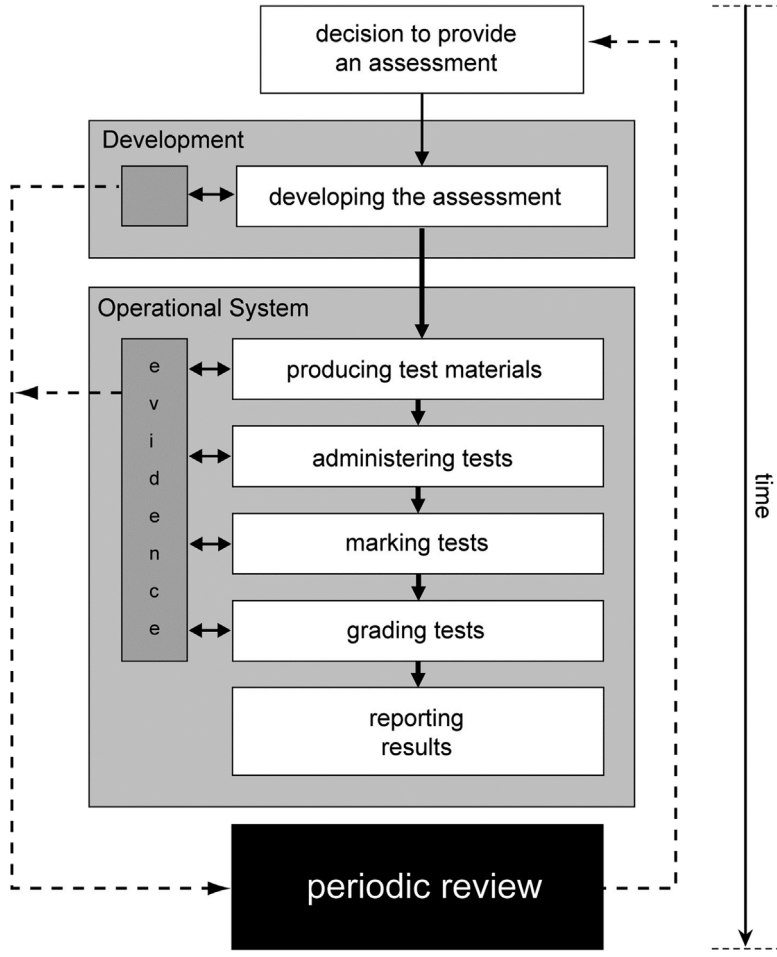


Figure 35.3 The assessment cycle showing periodic test review

including answer keys, application of grade boundaries, and equating standards across test sessions. Digital delivery of assessment and automated marking of productive skills tests support faster (or even instant) reporting of results for customers. In turn, QA process documentation and QC checks earlier in the assessment cycle are imperative as they may replace processes previously scheduled for the period before results release. Precise management of item banks (particularly for adaptive tests), enhanced tagging of test items with metadata, quality assurance of psychometric calibrations, and careful user-acceptance testing (UAT) of code deployments to identify unanticipated issues are critical. As discussed earlier, investigations into malpractice typically rely on aggregated response data from many candidates and so are not usually conducted in real time. Procedures for identifying malpractice require consideration, and institutional policies for handling suspected cheating retrospectively (while still satisfying regulatory requirements where necessary) may need to be created or updated to supplement the desire for faster results release (see Jones, forthcoming).

Management of enhanced reporting

The availability of computer-based testing and fast or instant reporting creates new dynamics in the relationship between the assessment organisation and the learner or teacher, which must be planned and managed. Learners may assume that computer-based testing affords easier and more rapid access to granular data on their performance while the assessment organisation may have pedagogical and validity concerns about sharing this data – for example, concerns about over-interpretation of aspects of performance based on a small number of items. Anticipation of stakeholders' needs and wishes and communication plans designed to address these needs are key parts of the quality of service offered to customers that can be incorporated as part of the QM approach while at the same time adding to the validity argument for the assessment.

Use of data to provide insights on learners and learning

The expansion of computer-based testing has led to an associated exponential growth in the available data related to assessment. Assessment providers have access to wide-ranging candidate data that may be demographic; be representative of behavioural interaction with, or performance on, an assessment (e.g., time spent on task); or represent interaction with administrative systems. The development of auto-marking technology presents one opportunity for increased sophistication in data collation and use. (See Link, forthcoming, for a more detailed treatment of this topic.) New innovations in language testing, such as the deployment of technologies including keystroke logging to examine deletions or revisions of productive writing or eye tracking to record physiological responses to reading, offer further possibilities. (See Chapters 2–8 in Yu and Xu, forthcoming, for extensive exploration of developments in this area.) The opportunities created by these sources of "big data" span insights into assessment development and deployment, research, and operational validation and provision of personalised learning experiences for candidates, as well as business and marketing opportunities. One definition of big data is "things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value in ways that change markets, organisations, the relationship between citizens and governments, and more" (Mayer-Schönberger and Cukier, 2013: 6–7).

QM approaches are crucial to supporting testing providers to derive data insights, as developing appropriate data architecture and ensuring consistent management of data are the fundamental enablers of the knowledge discovery and innovation promised by large-scale data analysis. Challenges of managing assessment data in order to support its use in analytics include:

- dealing with diverse and distributed data sources which may be archived in multiple systems;
- tracking and validating data;
- heterogeneity in data formats and structures;
- ensuring data integrity and security; and
- the development of specialised analytical tools.

(National Research Council, 2013, cited in McNeely and Hahm, 2014)

Legal and ethical implications of technology and data use in assessment

The possibilities offered by big data are attractive to large assessment organisations, but in a field as young as big data analytics, standards that govern the appropriate legal and ethical use of data and the associated insights are important to consider.

Legal requirements such as the EU General Data Protection Regulation (GDPR) represent one such way that the use of data and associated analysis may be governed by national and international regulators, and incorporating appropriate organisational guidelines into QM systems to ensure compliance with these regulations is vital for test providers of any size. The penalties for non-compliance are sizeable – 4 percent of annual global turnover or €20 million in the case of the most serious breaches of GDPR.

Other relevant bodies have started to derive their own principles for ethical management of data to which organisations can informally ascribe. One example is the FAIR Guiding Principles for scientific data management and stewardship in research (Wilkinson *et al*., 2016), jointly derived by stakeholders representing academia, industry, funding agencies, and publishers, which stipulate that data should be findable, accessible, interoperable, and reusable – ensuring that data and metadata are described, indexed, authenticated, referenced, and accessed according to standard guidelines.

Broader consideration of the impact and influence of data use and the systems and technology we develop to drive it is also crucial. Richards and King (2014: 426) state:

> We need to ensure that we think ethically about big data and other new information technologies. . . . [T]hey are the product of human choices and they will affect human values. We need to be sure that these human technologies shape the kind of society we want to have, for these technologies will shape the societies we will live in and the humans we will become.

There is a clear need for test providers to have transparent policies on what data is collected and how it is used for assessment purposes, in line with the maxim of *positive impact by design* (Saville, 2016). While larger assessment organisations may have greater scope for harvesting and harnessing big data, there is a responsibility for all test providers to determine how these considerations are incorporated into their QM documentation and processes.

Recommendations for practice

In this chapter, we have seen that an effective QM system within an organisation involves a clear statement (in the form of a mission statement or charter) of the aims and objectives to be achieved. The adoption of an internationally recognised code of practice can provide this function for many language test providers, especially smaller organisations (such as schools) or those whose main function is not language assessment (university departments, cultural bodies, etc.). It is recommended that all those involved in providing language tests should become familiar with one of the recognised codes of practice listed in the references and should review their aims and objectives in light of the underlying principles. If appropriate, those principles should be integrated into working practices.

Doing so is often easier said than done as well-established procedures might need to be changed, with implications for resources and how they are deployed. The *management of*

change therefore is an important consideration here, and the timescale for bringing about necessary improvements must be taken into account. Introductory textbooks on *organisation theory* and *innovation* are recommended for readers who are interested in this aspect of organisational management (e.g., Hatch and Cunliffe, 2006).

QM systems require a commitment to *change processes*, and so the adoption of a QM approach can be helpful in bringing about change and continual improvement in a well-managed and principled way. It is axiomatic that improvement is always possible, even where cases of good practice may already exist. It is recommended, therefore, that test developers should become familiar with QM principles and, where appropriate, seek to implement QM systems.

It should be remembered that quality standards cannot be imposed from outside but must be established and monitored within an organisation in the first instance. This means that awareness raising and self-evaluation should precede any form of external monitoring, which can be introduced at a later stage to check and confirm that the required standards are being met and to achieve external accreditation. This is consistent with the notion of validation based on the accrual of evidence over time.

To address issues of accountability, the need to develop employees' knowledge and skills to ensure understanding of language testing principles is highlighted as teams become more diverse and specialised to deliver computer-based and digital assessments. In this regard, a QM approach has the potential to act as a driver for the development of improved assessment literacy (Saville, 2014).

Any test provider can introduce a typical QM sequence into their working practices, as exemplified by the following bullet points.

- Define your assessment objectives (mission), including the role of your institution, its goals, and future ambitions.
- Assess what you currently do.
- Identify areas in most need of improvement.
- Decide on measures of improvement and an action plan to bring about the changes.
- Carry out the planned action to implement the improvements.
- Review progress and revise the action plan.
- Carry out new action to bring about further improvements.
- Repeat the cycle.

As QM systems provide the basis for wider accountability, organisations employing this approach can be confident of meeting requirements for accountability and should consider accreditation by an external agency.

An example of this practice is the ALTE Quality Management System, which provides a range of materials which can be accessed from the ALTE website. ALTE adopted a quality model and adapted its code of practice (ALTE, 1994) to function as a QM tool (Kuijper, 2003; Saville, 2005, 2010; van Avermaet, 2003; van Avermaet *et al*., 2004).

This redesigned format (as Microsoft Excel–based checklists) reflected the assessment cycle which test developers are very familiar with, as discussed earlier (development, administration, processing, analysis, and review). Taking both the code of practice and QM checklists into account, parameters were established for setting *minimum standards*. ALTE members (which include small test providers in university departments as well as international examinations boards) are now required to make a formal claim that a particular test or suite of tests has a quality profile appropriate to the context and use of the test.

Members build a reasoned argument explaining how the minimum standards are met and provide relevant evidence to back up this claim. The formal scrutiny of these parameters takes place during an auditing process, which is intended to ensure that adequate standards are achieved. Auditing has both a consultancy and a quality control role; if quality standards are not being met, the audited organisation is required to implement an action plan aimed at working toward and ultimately reaching the required standards.

Future developments

Language test providers are expected to account for their assessment practices with reference to theory and to provide evidence to back up the claims they make about the quality of their tests. Debate exists over the nature of regulation and who is best qualified to set and monitor professional standards; does the responsibility lie within the language testing profession itself (e.g., through self-regulation), or is there a need for an independent regulator of some kind? If so, how might regulation operate on an international level, and who would "guard the guards themselves"?

This chapter has suggested that whatever the jurisdiction for regulation, the basis for accountability needs to derive from a clear understanding of the principles of assessment coupled with a capacity to ensure that tests are developed and maintained effectively over time within organisations. First and foremost, the responsibility lies with the test providers themselves. This will mean that the larger test providers (testing agencies and examinations boards) will increasingly need to employ QM techniques to ensure that they can meet the demands of the regulatory environment.

To keep up with these demands, access to specialist knowledge and skills will become increasingly relevant, and more attention will need to be given to the professional development of language testers working within their own organisational contexts (whether teachers in schools or staff in testing agencies). Emphasis on digital approaches to assessment and the use of data for insights into learners and learning necessitates adaptation of existing QM approaches to ensure that quality and validity are maintained in transitions to new ways of working. In terms of staff training, a greater focus on management skills (e.g., covering the administrative aspects of the assessment cycle) is likely to be necessary.

The need for regulation and accountability in language testing has become more widely acknowledged over the last decade; it is now expanding to consider assessment practices in both established and emerging technologies. As before, a pragmatic approach is to ensure that best-practice models which employ QM techniques are identified and shared, in order to help test providers develop workable systems and for tangible improvements to be made incrementally. The sharing of expertise is essential, and in this regard, the professional language testing associations and international discussion forums have a central role to play in enhancing awareness of quality within the wider concept of *assessment literacy*.

Conclusions

In this chapter, we have illustrated how a QM approach can help improve language assessment and how the monitoring of quality through auditing procedures can enhance professional standards. We have noted that a QM approach can be helpful in fulfilling obligations for accountability to key stakeholders (especially the test takers and users of the test results) and to regulatory bodies. Digital approaches to assessment may rely on many of the same principles of good practice as traditional paper-based formats, but some adaptations to how

expert teams interact and additional consideration of both validation processes and ethical concerns may be necessary. A validity argument and the supporting evidence which has been constructed using QM principles can be readily scrutinised during an audit or inspection in order to verify whether appropriate industry standards are being met and/or statutory regulations are being adhered to (e.g., the code of practice of a regulatory body).

In recent years, codes of ethics and codes of practice have been available for language testers to guide their practices (see Walters, this volume), but these are not designed to assist practitioners in carrying out the day-to-day work of writing and administering tests or for establishing minimum standards for their work. While an understanding of the ethical concerns and the underlying principles of assessment provides necessary conditions for developing fair and valid assessments, a QM approach can help link these concerns to practicable ways of doing things effectively. By adopting a QM approach, it is possible to connect the principles of good practice in language assessment with operational procedures which are transparent and accountable. In other words, it provides a practical approach to putting *principles into practice*.

Acknowledgements

We are grateful to our colleagues Dr. Kevin Cheung, for his input to this chapter, and Dr. Michael Corrigan, for his help in designing the figures.

Further reading

Quality in educational assessment

Wild, C. L. and Ramaswamy, R. (2008). *Improving Testing. Applying Process Tools and Techniques to Assure Quality*. London, UK: Routledge.

This is an edited volume of 18 chapters focusing on a range of process tools and techniques to assure quality and to improve testing. It is wide ranging in coverage, and, although not specifically aimed at language testing, it deals with many of the issues raised in this chapter in an accessible way. The editors themselves set the scene by discussing the risks and the costs of poor quality in testing, and they round off the volume with thoughts on the future of quality in the testing industry.

Background to quality management systems

There are numerous textbooks available which explain the principles of quality management. One readily available text which is aimed at non-specialist readers is *Quality Management Demystified* by Sid Kemp (2006, New York, NY: McGraw-Hill).

It defines the basic concepts and describes the tools for defining and measuring quality. In its historical review it covers Total Quality Management, as well as more recent schemes such as Six Sigma. Anderson (2008) in Wild and Ramaswamy also describes the use of Six Sigma in a large-scale testing company. Rose (2010) describes the ISO-based approach to quality management, which has been adopted by Cambridge Assessment English in the context of English language assessment in "Setting the standard: Quality management for language test providers" (*Research Notes 39 [2–7]*). Beresford-Knox (2015) also discusses QM in the Cambridge Assessment English context in "The role of Quality Management in ensuring accurate and meaningful test scores" (*Research Notes 59 [40–44]*). To download these articles, please visit www.cambridgeenglish.org/research-notes/. Useful information

on quality management is available on many QM websites: e.g., Chartered Quality Institute (CQI): www.thecqi.org. The ISO standards can be accessed from the ISO website (www.iso.org), which also provides many useful documents related to quality. National sites – e.g., British Standards Institute (www.bsigroup.com) – also provide helpful advice and accessible explanations. See, for example, documents such as ISO 9001: 2015 *Quality Management System Requirements*; ISO 9001: 2015 *Quality Management Systems – Fundamentals and Vocabulary*.

Codes of practice

Codes of practice in language testing emerged relatively recently, starting in the 1990s.

A more general *Code of Fair Testing Practices in Education* was published in 1988 by the Joint Committee on Testing Practices (JCTP) in cooperation with the American Psychological Association and was particularly influential as it deliberately provided an approach which could be adapted to a wide range of testing contexts. See www.apa.org/science/programs/testing/fair-testing.pdf. For example, the Association of Language Testers in Europe (ALTE) produced its code of practice (1994) based on the JCTP model. ALTE also provides a range of other resources based on their code of practice and have developed an approach to quality management which employs techniques such as auditing to improve quality. See the ALTE website: www.alte.org. Reference materials which can be downloaded include "Minimum standards for establishing quality profiles in ALTE examinations" (2007), "The ALTE Quality Management and Code of Practice Checklists" (2008), and "Manual for test development and examining" (2010). Produced by members of ALTE on behalf of the Council of Europe to accompany the CEFR.

Other associations have developed codes of practice and guidance for test providers in regional contexts, including the Japanese Language Testing Association (JLTA) and the European Association for Language Testing and Assessment (EALTA). See EALTA Guidelines for Good Practice in Language Testing and Assessment (2006; expanded in 2014, www.ealta.eu.org). The International Language Testing Association (ILTA) promotes ethical standards in language testing through its code of ethics (adopted in 2000) and its guidelines for practice (2007). These resources are available from the ILTA website: www.iltaonline.com.

References

AERA, APA and NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA Publishing.

Alderson, J. C., Clapham, C. and Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge, UK: Cambridge University Press.

ALTE. (1994). *ALTE Code of Practice*. Retrieved from www.alte.org/resources/Documents/code_practice_en.pdf.

Bachman, L. F. and Palmer, A. (1996). *Language Testing in Practice*. Oxford, UK: Oxford University Press.

———. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

British Standards Institute (BSI). (2012). *Pocket Guide to Standards Development*. London, UK: British Standards Institute.

Brown, J. D. (2016). Language testing and technology. In F. Farr and L. Murray (eds.), *The Routledge Handbook of Language Learning and Technology*. Abingdon, UK: Routledge, 141–59.

Carr, N. T. (2014). Computer-automated scoring of written responses. In A. J. Kunnan (ed.), *The companion to language assessment*. Vol. 2, 1063–1078. Chichester, UK: Wiley.

Deming, W. E. (1986). *Out of the Crisis*. Cambridge, UK*:* Cambridge University Press.

Downing, S. M. and Haladyna, T. M. (2006). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum.

Foster, D., Maynes, D. and Hunt, B. (2008). Using data forensic methods to detect cheating. In C. L. Wild and R. Ramaswamy (eds.), *Improving Testing. Applying Process Tools and Techniques to Assure Quality*. London, UK: Routledge.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment – An Advanced Resource Book*. Abingdon, UK: Routledge.

———. (2009). Test architecture, test retrofit. *Language Testing* 26: 123–44. https://doi.org/10.1177/0265532208097339.

Hatch, M. J. and Cunliffe, A. L. (2006). *Organization Theory: Modern, Symbolic and Post-Modern Perspectives*, 2nd edn. Oxford, UK: Oxford University Press.

Heyworth, F. (2013). Applications of quality management in language education. *Language Teaching* 46: 281–315. https://doi.org/10.1017/S0261444813000025.

Hughes, A. (2003). *Testing for Language Teachers*, 2nd edn. Cambridge, UK: Cambridge University Press.

ISO (International Organization for Standardization) (2015). *Quality Management Systems - Requirements (ISO standard no. 9001:2015)*. Retrieved July 18, 2021, from https://www.iso.org/standard/62085.html

ISO (International Organization for Standardization) (2019). *ISO 9001:2015 How to use it*. Geneva, Switzerland: ISO. Retrieved July 18, 2021, from https://www.iso.org/files/live/sites/isoorg/files/store/en/PUB100373.pdf

Joint Committee on Testing Practices. (1988/2005). Code of fair testing practices in education (revised). *Educational Measurement: Issues and Practice* 24: 23–6. https://doi.org/10.1111/j.1745-3992.2005.00004.x.

Jones, E. (forthcoming). Technology and cheating on tests. In G. Yu and J. Xu (eds.), *Language Test Validation in a Digital Age. Studies in Language Testing Volume 52*. Cambridge, UK: UCLES/Cambridge University Press.

Kane, M. (2006). Validation. In R. L. Brennan (ed.), *Educational Measurement*, 4th edn. Washington, DC: American Council on Education/Praeger, 17–64..

Kemp, S. (2006). *Quality Management Demystified*. New York, NY: McGraw Hill.

Kuijper, H. (2003). QMS as a continuous process of self-evaluation and quality improvement for testing bodies. Retrieved August 2, 2011, from www.alte.org/qa/index.php.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (ed.), *Fairness and Validation in Language Assessment*. Cambridge, UK: Cambridge University Press, 1–13.

———. (2004). Test fairness. In M. Milanovic and C. Weir (eds.), *European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001*. Cambridge, UK: UCLES/Cambridge University Press, 27–50.

Lane, S., Raymond, M. R. and Haladyna, T. M. (2016). *Handbook of Test Development*, 2nd edn. Abingdon, UK: Routledge.

Lane, S., Raymond, M. R., Haladyna, T. M. and Downing, S. M. (2016). Test development process. In S. Lane, M. R. Raymond and T. M. Haladyna (eds.), *Handbook of Test Development*, 2nd edn. Abingdon, UK: Routledge, 3–18.

Link, S. (forthcoming). The interface between automated essay scoring and socio-cognitive research. In G. Yu and J. Xu (eds.), *Language Test Validation in a Digital Age. Studies in Language Testing Volume 52*. Cambridge, UK: UCLES/Cambridge University Press.

Mayer-Schönberger, V. and Cuckier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Boston, MA: Houghton Mifflin Harcourt.

McNeely, C. L. and Hahm, J. (2014). The big (data) bang: Policy, prospects and challenges. *Review of Policy Research* 31: 304–10. https://doi.org/10.1111/ropr.12082.

Messick, S. (1989). Validity. In R. Linn (ed.), *Educational Measurement*, 3rd edn. New York, NY: Macmillan, 13–103.

Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly* 14: 1–18.

National Research Council (NRC). (2013). *Frontiers in Massive Data Analysis*. Washington, DC: The National Academies Press.

O'Reilly. (2011). *Velocity 2011: Jon Jenkins, "Velocity Culture"* [Video File]. Retrieved from www.youtube.com/watch?v=dxk8b9rSKOo.

Richards, N. M. and King, J. H. (2014). Big data ethics. *Wake Forest Law Review* 49: 393–432.

Saville, N. (2005). Setting and monitoring professional standards: A QMS approach. *Research Notes* 22: 2–5. Cambridge, UK: Cambridge ESOL.

———. (2010). Auditing the quality profile: From code of practice to standards. *Research Notes* 39: 24–28. Cambridge, UK: Cambridge ESOL.

———. (2014). Using Standards and Guidelines. In A. J. Kunnan (ed.), *The Companion to Language Assessment: Approaches and Development Volume II*. Oxford, UK: Wiley Blackwell.

———. (2016). *Principles of Good Practice*. Cambridge, UK: Cambridge English Language Assessment.

Saville, N. and Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal* 53: 42–51. https://doi.org/10.1093/elt/53.1.42.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. New York, NY: D. Van Nostrand Company.

———. (1939). *Statistical Method from the Viewpoint of Quality Control*. Washington, DC: The Graduate School, the Department of Agriculture.

Taylor, F. W. (1911). *The Principles of Scientific Management*. New York, NY: Harper and Brothers Publishers.

van Avermaet, P. (2003). *QMS and The Setting of Minimum Standards: Issues of Contextualisation Variation Between the Testing Bodies*. Retrieved August 2, 2011, from www.alte.org.

van Avermaet, P., Kuijper, H. and Saville, N. (2004). A code of practice and quality management system for international language examinations. *Language Assessment Quarterly* 1: 137–50. https://doi.org/10.1080/15434303.2004.9671781.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke, UK: Palgrave Macmillan.

Wild, C. L. and Ramaswamy, R. (2008). *Improving Testing: Applying Process Tools and Techniques to Assure Quality*. London, UK: Routledge.

Wilkinson, M. D. *et al*. (2016). *The FAIR Guiding Principles for Scientific Data Management and Stewardship*. Retrieved from www.nature.com/articles/sdata201618.

Yu, G. and Xu, J. (eds.). (forthcoming). *Language Test Validation in a Digital Age. Studies in Language Testing Volume 52*. Cambridge, UK: UCLES/Cambridge University Press.

36

Epilogue

Language testing: where are we heading?

Luke Harding and Glenn Fulcher

Introduction

A common lament in our field is a lack of longitudinal research: studies tracking change in test-taker performance over time, charting the development of rating quality as raters' experience develops, or documenting the impact and long-term consequences of a test following its implementation. This situation is gradually changing (for example, see Green and Van Moere, 2020; Knoch *et al*., 2020; Lamprianou *et al*., 2021), and this is a welcome development. There will always be value in tracing patterns of change, not least because such patterns can help predict future trends based on past trajectories. The publication of this second edition of the *Routledge Handbook of Language Testing* provides us with a unique opportunity to observe patterns of change, too, but with a broad focus on the development of language testing and assessment as a field of study. A consideration of the revisions and updates to chapters in the first edition and the content of newly commissioned chapters illuminates questions concerning such changes in the field. What new themes are common across chapters? What future problems are identified? What now constitutes the scope of language testing scholarship? In what direction(s) is language testing and assessment progressing (or regressing)? It may sound grandiose, but a collection of this kind – written by experts on their respective topics – provides direct insight into the state of the art in language testing and some glimpses of what the next decade may hold.

In our close reading of the chapters, working with authors on their drafts and revisions, we have identified four salient themes that run across the collection and connect many contributions:

1 The increasing role of technology
2 Connecting assessment with learning
3 Grappling with complexity
4 Theorising the socio-political nature of language testing

DOI: 10.4324/9781003220756-47

In this epilogue, we will expand on these four themes, discussing what we have learned from the chapters about the current state of the field and how it is evolving. We will highlight some commonalities across the chapters but also point to areas where there is still room for the field to grow and explore over the next ten years (perhaps predicting some salient topics for the third edition). In the final part of this epilogue, we will make a call, as Fulcher and Davidson (2007) have done before, for effect-driven language testing as a way to mitigate potential future challenges we see on the horizon, and we will discuss the important role of a broad, multidimensional language assessment literacy in bringing the effect-driven testing approach to its full realisation.

The increasing role of technology

During the editing of this book, the world was thrown into turmoil as the COVID-19 pandemic swept the globe. The country we both live in – the United Kingdom – went into its first of several lockdowns in March 2020, and daily working life shifted online. Meetings were held on Zoom or Microsoft Teams; university teaching was performed using a variety of digital platforms. Wi-Fi became a lifeline to the outside world (and access to Wi-Fi and digital devices a source of compounding inequality as schools also moved to online provision). Language testing, too, was affected in different and sometimes unanticipated ways. In large-scale testing, face-to-face operations were shut down in many countries. Testing organisations responded by rapidly launching a range of "take-at-home" tests, which would compete with existing take-at-home options such as the Duolingo English Test (see Isbell and Kremmel, 2020). Meanwhile, for classroom teachers, assessment practices became intertwined with digital literacy as educators and learners grappled with different ways of assessing language online.

The authors of the chapters in this collection – who mostly wrote or updated their pieces prior to the pandemic or in its early stages – were prescient in understanding that digital technology would be a major emerging issue in language testing. Whether viewed as an exciting innovation or a looming threat to good practice, technology features heavily in the final parts of many chapters. We see Saville and McElwee (Chapter 35), for example, extolling the positive influence technology can have on quality management processes, and Cushing (Chapter 32) pointing to the wide range of applications of corpus linguistics across various aspects of language testing from domain description to the development of scoring systems through to validation – benefits also reinforced by Moder and Halleck (Chapter 5) focusing on the use of corpora in aviation English language tests. Jang and Sinclair (Chapter 11) describe the great potential of machine learning in dealing with the vast range of data that might be collected through various behavioural analytics for the purposes of diagnosis, and Nissan and Park (Chapter 23) highlight the advantages of technology in prototyping procedures. Brunfaut (Chapter 15) discusses the increasing of eye-tracking technology for validating reading tasks, and Xi (Chapter 30) describes new approaches to automated scoring, the use of which has spread exponentially and become more normalised since the first edition of the *Handbook*.

It is clear from the chapters in this collection that digital technology had already woven its way into essential elements of language testing practice even before the pandemic struck. However, there is one area where we anticipate there will need to be further discussion

and innovation in the field: understanding the changing nature of communicative practices across digital domains and what this means for the test constructs we work with and the inferences we make based on test scores. Isbell and Kremmel (2020: 616) summed this up well in their review of take-at-home tests:

> [T]he full potential of innovation in computer-delivered testing has not yet been fully taken advantage of. . . . Online tests could potentially allow for more interactivity, multimodality, and more authentic representations of today's communication and target language use domains than has been possible in traditional paper-based tests administered in secure test centers.

A challenge for the future, then, is to address the question: How can we harness technology to explore richer, more authentic constructs of language use in digital domains?

Researchers have been exploring these issues over the past decade (and more) in areas such as the use of video-conferencing in speaking tests (e.g., Nakatsuhara *et al*., 2017), the use of virtual environments as settings for performance assessment (Ockey *et al*., 2017), and the extensively researched area of multi-modal listening comprehension (Batty, 2020; Gruba, 2020). However, the coronavirus pandemic has made the situation more urgent. At the time of writing (March 2021), professional and educational communicative practices take place *primarily* online for many of us. There may be a tendency to treat this situation as temporary – that we will return to regular patterns in time. But it is likely that aspects of our online communication practices will stay with us into the future. Education will become more blended; workplaces will rely more on virtual communication; collaboration will occur in virtual spaces in addition to (or instead of) face to face. We only need to look at the growing prevalence of telehealth to understand how digital technology is changing not only the mode but also the nature of communication (see Marra *et al*., 2020). Yet we are far from capturing these new types of communicative behaviours in current language assessments.

We therefore see two future challenges for the field in responding to changing communicative contexts. The first is developing innovative, prototype tasks that can be trialled and, ultimately, used alongside or in place of existing assessment tasks. These might include tasks that target emerging, salient communicative activities: interactional competence in group chat, interactive lecture listening with a concurrent chat channel, collaborative writing in digital environments (see Knoch, Chapter 14, this volume), video creation (e.g., narrating online gameplay; video portfolios of spoken performance, captured with mobile devices), or speaking with an AI assistant. Such tasks would arguably tap into the kind of communicative ability that would better reflect the digital domains of the future. They are also likely to challenge the orthodox division of language behaviour into discrete constructs. This shift has been happening for some time (see Plakans, Chapter 21), but for a new generation, such divisions may make even less sense than they currently do. As Butler points out in her chapter on young learner assessments: "Some digital technology (e.g., social networking services [SNS] and chats) obscures the traditional boundaries between oral and written language that are used in language assessment" (Chapter 9). We would predict that conceptualising digitally mediated language ability is likely to move the field further toward more authentic, integrated tasks.

The second challenge is that *existing* tests will need to provide validity evidence of the inferential link between current tasks and rapidly shifting domains of language use (see Ross, Chapter 3). For many existing tests, the extrapolation inference (see Chapelle and Lee, Chapter 1, and Kane, Chapter 2) – the connection between the assessment and the

target domain – has been disrupted by the rapid shift to online communicative practices. As a result, new validity arguments will be required in the coming years to connect testing practices with real-world communicative settings. Computer-based delivery of a test, on its own, does not guarantee that tasks will assess construct-relevant skills and abilities. What will be required is a much deeper and more radical reappraisal of communicative constructs.

Connecting assessment with learning

Across many of the chapters in this volume, we see a blurring of boundaries between traditional paradigms of large-scale, standardised tests and learning-oriented, classroom-based assessment (see Gebril, 2021, for further discussion of paradigm "blurring" versus paradigm "clarity"). We also see a growing interest in connecting assessment with learning and *embedding* assessment within authentic classroom activities. This general topic was not only observed in those chapters that deal more explicitly with assessment in classroom contexts (chapters in Section 3) but also in chapters in which we might have anticipated a concentration on issues related to large-scale testing. For example, Jin (Chapter 16) provides an extended – and very important – discussion of the role of test specifications in classroom language assessment. Yan and Fan (Chapter 28) note that classroom teachers' ability to make multiple, up-close observations of their own students' performance provides an important rebuttal to the blanket misconception that teacher judgments lack reliability. And Knoch (Chapter 14) places formalised, high-stakes writing tests along a continuum with written corrective feedback (WCF), offering a critique of the typical divide between classroom-based approaches and the formal, summative approaches that, at times, seem to exist as separate entities in our field.

This shift toward what we would describe as a "learning-oriented turn" is not surprising given general trends in the field over the past ten years (see Turner and Purpura, 2016). Various paradigms – outlined in the chapters of this book – have conceptualised the connection between learning and assessment in different ways: diagnostic assessment (Jang and Sinclair, Chapter 11), dynamic assessment (Antón and García, Chapter 10), and classroom-based assessment (Fox *et al*., Chapter 7). These chapters outline research which demonstrates the *pedagogical* effectiveness of different learning-oriented approaches such as the provision of individualised feedback, self-assessment, and portfolio assessments. In addition to such benefits for learning, however, the pandemic and its subsequent impact on the ability to conduct in-person testing in communal test centres and school halls has only further highlighted the *logistical* benefits of a shift away from "one-shot" high-stakes exams. A recent think piece by Taylor and Harding (2020) on the issue of assessment during the COVID-19 pandemic made the following point:

> [T]he Covid 19 situation highlighted the reality that high-stakes, end-of-school exams are high-risk endeavours if something goes wrong. This strengthens the case for greater emphasis on continuous assessment/assessment for learning/learning oriented assessment, not only as a better approach pedagogically, but also a more risk-averse approach.

Amidst these developments, digital technology is also playing an important role in opening up a range of new affordances for connecting assessment with learning. For example, Meurers *et al*. (2019) described a study in which computational methods were used to track learning "in the wild" (p. 18). The researchers introduced a range of computer-based learning tasks and then, over the course of one year, logged interactions with these tasks and with

feedback; monitored uptake of particular language features; and combined this data with other measures of learners' working memory capacity, aptitude, and proficiency to arrive at a holistic understanding of the learners' development over time. This all occurred within the classroom context; the fact that tasks were delivered via computer allowed for this range of metrics to be gathered and utilised. The researchers argued, therefore, that such approaches provide a way of analysing learning in a way that is "ecologically valid" (p. 24): the learner does not have to move to the artificial environment of the examination hall to be measured. Assessment happens *in situ* as learning takes place.

This approach overlaps with the sorts of future possibilities predicted by some authors within this volume. Sawaki (Chapter 31) describes the integration of computer-based testing (CBT) with learning as the "third and final phase" of Bennett's (2000) predicted stages for the development of CBT. Testing at this stage would be "curriculum embedded" and would potentially take the form of "learning tools with virtual reality simulations for both individual and collaborative work." The data that could be drawn on would be diverse. Jang and Sinclair (Chapter 11) discuss the range of data sources that might be useful for making diagnostic decisions about learners' ability and the types of difficulties they experience. These include "process-oriented multichannel data involving physiological, neural, behavioral, and metacognitive traces."

The closer connection of assessment with learning is a positive development, and technology plays an important role in realising this ambition of ecological validity: assessing learners within a classroom environment and providing rich data over time. However, these developments also raise critical questions about the ethics of such approaches (see Walters, Chapter 33; Harding and Kremmel, forthcoming) and their potential impact on relational aspects of learning and teaching. What does it mean when we remove from the teacher the decision of what to assess, when to assess, and how to assess? What impact do constant surveillance and data gathering have on learning and the learning process (e.g., the ability to take risks)? What will happen to the data that is generated? (Who will store it? Who will have access to it? Who will have authority and expertise to process it?) There is a risk that in this new approach to assessment, power is shifted upward to those who manage such systems or have the technical skills to generate or analyse data and away from learners and teachers themselves. Such systems will need to be monitored carefully, and their impact on both the learner and the teacher will need to be understood.

Grappling with complexity

The need to deal with complexity arises across the chapters in this handbook in different ways. First, we can understand complexity in terms of language itself. There is an increasing recognition of complexity in the constructs of language ability that we seek to measure and in the varieties of language used in real-world settings (see Young, Chapter 4). In the context of listening assessment, Wagner (Chapter 13) advocates enhanced text authenticity and the need to introduce features into listening texts which more accurately reflect the complex nature of real-world spoken communication. And Moder and Halleck (Chapter 5) highlight the need to understand and model the complexities of English as a Lingua Franca (ELF) communication in the context of aviation English language tests. It is therefore clear that the field is acknowledging in a very serious way the nature of language use as it occurs in real-world communicative contexts. Relatedly, several chapters focus on the challenges of judging multifaceted constructs of spoken interaction. Interactional competence and pragmatic

competence are burgeoning areas of research interest (see Nakatsuhara *et al*., Chapter 12), but their complexity creates challenges in ensuring fair judgments. Performance assessment requires scales that capture these rich constructs (Galaczi and Lim, Chapter 29). Training for raters and interlocutors may help to reduce complexity, though as Davis (Chapter 19) points out, uniformity in judging something as intricate as language performance may be a chimera. Rating is itself a complex and dynamic process (see Eberharter, 2021), and so in the case of speaking assessment – for example – we see layer upon layer of variables interacting to produce a single score.

Second, we see complexity in the interaction between test-taker characteristics and the specific demands of assessment tasks. A number of chapters highlight, specifically, the rich vein of current research into test takers' cognitive process, strategies, and individual difference variables that influence the ways in which they engage with assessment tasks. For example, Brunfaut (Chapter 15) shows that reading is made up of numerous processes which interact and overlap in unique ways to achieve text comprehension in reading assessment. Cohen (Chapter 22) surveys literature identifying the wide range of strategies that have relevance for test takers. These issues are also highly salient in young learner assessment. Research on young learners in assessment contexts is only starting to fully understand the challenges of assessing primary and secondary school children and the research points to enormous complexity in developing appropriate tests for this test taker population. As Butler explains (Chapter 9):

> Unlike adult additional/foreign (L2) language learners, YLLs [young language learners] are still developing their first language (L1) in addition to developing an L2. They are also in the midst of cognitive, social-cognitive, and affective development. Importantly, there are substantial individual variabilities in these linguistic- and age-related developments both within and across chronological age groups.

For all learners, awareness of specific learning difficulties and test accommodations is also a growing area of study (see Abedi, Chapter 18), and we would anticipate that as many institutions shift their focus toward diversity and inclusion, this topic will be very much at the forefront of the field over the next decade.

The final area in which we see increasing complexity is in the processes of language test development, administration, and scoring: the professional practice of language testers themselves. The decisions and judgments language test professionals make in constructing the content of assessments (e.g., what to include in specifications, which items to select for a test form, what to allow/disallow in a marking scheme) have long been an under-researched element in our field, despite strong theoretical frameworks (e.g., evidence-centred design, Yin and Mislevy, Chapter 17). One reason for this might be that decision making in such contexts can be very challenging to reveal in operational settings. Shin (Chapter 20), for example, demonstrates the vast range of issues that can (and need to) be discussed regarding the practices of item writers. Other chapters in this volume demonstrate the rich lines of research that could be performed around item prototyping (Nissan and Park, Chapter 23), piloting (Read, Chapter 25) and pre-operational testing more generally (Kremmel *et al*., Chapter 24). On a broader scale, quality control mechanisms are also becoming increasingly challenging given the wide range of professionals from adjacent fields who are now involved in test development (as explained by Saville and McElwee, Chapter 35). And then there is the increasing complexity of validation itself, with respect not only to the frameworks and arguments that form the conceptual basis for validation efforts but also to the

statistical literacy required to conduct validation research and interpret its quality (Brown, Chapter 26; Ockey, Chapter 27).

We might argue that complexity has always been a feature of language assessment, though once again, technology brings new affordances and constraints. New digital technology allows us to capture this complexity in useful ways. Corpus approaches, for example, can provide key insights into the nature of spoken and written communication, making plain the true nature of discourse within specific domains of language use. We see this in the corpus used by Moder and Halleck to provide evidence of the English as a Lingua Franca (ELF)–like quality of aviation communication. And technology also provides ways of capturing complexity in validation research through the use of new technologies such as eye tracking. Still, there is an ever-present risk that digital technology might reduce complexity in unhealthy ways. An obvious example is the narrowing of constructs in computer-based, computer-scored speaking tests (what Isaacs [2018] calls "fully automated" speaking assessments). Constraints on what technology can do in such contexts – e.g., the difficulty automated scoring systems have in dealing with extemporaneous speech and the subsequent use of less complex tasks to produce more predictable language – have not stopped such tests from being widely adopted even for high-stakes purposes. We are currently at a crossroads in this sense as we recognise the need to capture increased complexity, but with the attractiveness of efficient testing systems harking back to constructs and tasks that are the antithesis of this. This tension between conceptual advances and technological-based efficiency is likely to be at the forefront of discussions into the coming decade.

Theorising the socio-political nature of language testing

The final theme concerns the socio-political nature of language testing. This is by no means a new observation; the socio-political character of language testing and assessment has been well explained in the work of various scholars (e.g., Fulcher, 2015; McNamara and Roever, 2006; Mirhosseini and De Costa, 2020; Shohamy, 2001). In the current volume, Kunnan's detailed critique of the United States Naturalization Test (Chapter 6) provides a classic account of the way in which language proficiency/literacy assessment can be implemented for ideological motives. Deygers (Chapter 34) provides an erudite insight into language proficiency standards, exploring their policy function and touching on the political character of the use and misuse of influential frameworks such as the CEFR.

Several chapters in this volume point to the need for a much broader theorisation of social and political factors in language testing. Two examples will illustrate this point. First, Cheng and Sultana (Chapter 8) raise numerous points regarding washback as a socio-political phenomenon. One of the most interesting passages, however, relates to Sultana's doctoral research conducted in the context of an exam reform in Bangladesh (2019). Sultana describes a situation where "political instability in Bangladesh caused delays in delivering instructional materials." In the absence of these materials, an unintended form of washback occurred in the classroom. Here, washback was influenced by wider socio-political forces – political instability – in a very tangible way. Yet Cheng and Sultana note the lack of theory which would help explain this relationship, observing that "existing literature contains very few studies systematically documenting the impact of political unrest on washback."

On a separate point, in his chapter on item writing and item writers, Shin (Chapter 20) raises thought-provoking questions about the conditions of labour markets and the ultimate effect on the quality of items and tests produced as a result. Specifically, he writes:

> When item writers are underpaid, when parts of item writing jobs (e.g., searching for relevant passages on the internet or editing/proofreading item drafts) are unethically foisted upon part-time or freelance writers, when some cultures of communication do not allow group dynamics to be horizontal, and when job coordinators or peers take too much power in decision making, then even good test specifications or well-qualified item writers cannot be guaranteed to generate good items for the testing topics.

A broader, related concern here is the under-theorising of economics in language testing. At the time of writing, test providers have been faced with economic challenges resulting from losses incurred by the suspension of in-person testing. This has led to new products coming onto the market and existing tests gaining in reach and recognition as they meet a market need for efficient and accessible online testing. Yet there is little discussion of the profound impact of market forces on testing practices in our extant research literature. The economic aspect of language testing is always the elephant in the room. The business model of language testing is in many was unique. What financial incentives are there for test providers to spend money on validation research or to follow best-practice approaches? In an unregulated space, we rely to a considerable degree on the good will and professionalism of those who create and research tests. Yet decisions about test design, construct operationalisation, administration, scoring, use, and impact are often dictated by financial and market-based considerations, and these may be in opposition to best-practice, research-based approaches. Such considerations are seldom accounted for or explicitly acknowledged in validation research. Chapters in this volume suggest that language testing scholars require a deep, critical understanding of how language testing – as an activity and a practice – is situated within, and subject to, broader political, social, and economic forces.

Toward effect-driven language testing

In their 2007 book, Fulcher and Davidson advocated effect-driven language testing thus:

> The task for the ethical language tester is to look into the future, to picture the effect the test is intended to have, and to structure the test development to achieve that effect.
> *(Fulcher and Davidson, 2007: 144)*

Effect-driven testing remains a laudable goal: an approach that provides a clear orientation for language testers who strive for ethicality. But how do we get there? Given that the field is constantly in flux, an effect-driven approach rests on a broad understanding of data available *now* to guide decisions about what the future may hold. But what is the scope of the language testers' enquiry into future forces? What are the parameters? The previous sections of this epilogue – and indeed all the chapters in this book – have demonstrated that it is not enough for language testers to be guided only by traditional best-practice methods of writing good quality items, pre-testing, administering, scoring, and reporting, as crucial as each of those steps is. Rather, language testers now require a solid understanding of language tests not only as designed objects but also as social objects, subject to external forces. They

need to understand the increasing role of technology and to be conversant with professionals from other fields increasingly involved in language testing practices. They need to understand how the field of linguistics is evolving, with respect to both psycholinguistic views of language learning and sociolinguistic understandings of language use and language change. And they need to have a deep understanding of how language assessment is connected with pedagogy and how it can take different forms as language learning increasingly shifts to a digital mode.

In essence, an effect-driven approach to testing requires a multi-dimensional type of language assessment literacy (see Fulcher, 2012; Kremmel and Harding, 2020; Taylor, 2013) and a strategy for its acquisition among stakeholders (Fulcher, 2020; Yan and Fan, 2021). Discussions of what constitutes LAL are, ultimately, discussions about the scope of the field. For this reason, it should come as no surprise that language assessment literacy has become a major source of interest in language testing over the past decade. We see LAL as an ever-widening set of concerns that shift and grow over time as the field takes on new directions and leaves others behind. Our discussions about the nature of this field of knowledge are important not only insofar as they inform good training but also because they act as a mirror for our own understanding of ourselves as scholars and practitioners.

The connection between LAL and an effect-driven approach is therefore a vital one; LAL is the foundation of the effect-driven approach. Test developers, validation researchers, academics, teachers, and others involved in the process of conceptualising, designing, administering, and using tests require knowledge that will allow them to pre-empt unintended consequences, to plan for problems, and to mitigate their effects. By extension, effect-driven testing – in its purest form – is not simply a process of identifying positive washback and designing tasks that will work well to promote learning. It involves looking deeply and comprehensively at the structures around language test construction. It means broadening our understanding of socio-political forces which impact our practices – not only policy literacy (as many have argued for [Pill and Harding, 2013; Deygers and Malone, 2019]) but also a critical understanding of the ideologies that surround language testing, the market-based incentives that drive design decisions, the social and political forces that might distort or undermine intended effects, the profound role of technology in the affordances of language testing and the personnel involved, the complexity of language itself and its refusal to be easily shaped into measurable units. In other words, the future of language testing will require a radical sort of introspection to consider the skills and knowledge that will be required for the next generation of language testers to grapple with these problems and make tests which are fair, just, ethical, and of benefit to wider society. The scope of the field is changing, and the type of knowledge and research that we consider to be relevant to language testing is moving from the periphery to the centre of our practices. We anticipate that in another ten to fifteen years, a new volume of the *Handbook* will reflect these changes.

References

Batty, A. O. (2020). An eye-tracking study of attention to visual cues in L2 listening tests. *Language Testing*. https://doi.org/10.1177/0265532220951504.

Bennett, R. E. (2000). *Reinventing Assessment: Speculations on the Future of Large-Scale Educational Testing*. Princeton, NJ: ETS.

Deygers, B. and Malone, M. E. (2019). Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing* 36: 347–68. https://doi.org/10.1177/0265532219826390.

Eberharter, K. (2021). *An Investigation into the Rater Cognition of Novice Raters and the Impact of Cognitive Attributes When Assessing Speaking*. Unpublished doctoral dissertation, Lancaster University, United Kingdom.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly* 9: 113–32. https://doi.org/10.1080/15434303.2011.642041.

———. (2015). *Re-Examining Language Testing: A Philosophical and Social Inquiry*. Oxon, UK: Routledge.

———. (2020). Operationalizing assessment literacy. In D. Tsagari (ed.), *Language Assessment Literacy: From Theory to Practice*. Newcastle upon Tyne, UK: Cambridge Scholars, 8–28.

Fulcher, G. and Davidson, F. (2007). *Language Testing and Assessment: An Advanced Resource Book*. London and New York: Routledge.

Gebril, A. (ed.). (2021). *Learning-Oriented Language Assessment: Putting Theory into Practice*. New York, NY: Routledge.

Green, A. and Van Moere, A. (2020). Repeated test-taking and longitudinal test score analysis. *Language Testing* 37: 475–81. https://doi.org/10.1177/0265532220934202.

Gruba, P. (2020). What does language testing have to offer to multimodal listening? In G. J. Ockey and B. A. Green (eds.), *Another Generation of Fundamental Considerations in Language Assessment*. Singapore: Springer, 43–57.

Harding, L. and Kremmel, B. (forthcoming). Technology, values, ethics and consequences: From innovation to impact in language assessment.

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson and J. Murphy (eds.), *The Routledge Handbook of Contemporary English Pronunciation*. London, UK: Routledge, 570–84.

Isbell, D. and Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high stakes decisions. *Language Testing* 37: 600–19. https://doi.org/10.1177/0265532220943483.

Knoch, U., Huisman, A., Elder, C., Kong, X. and McKenna, A. (2020). Drawing on repeat test takers to study test preparation practices and their links to score gains. *Language Testing* 37: 550–72. https://doi.org/10.1177/0265532220927407.

Kremmel, B. and Harding, L. (2020). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly* 17: 100–20. https://doi.org/10.1080/15434303.2019.1674855.

Lamprianou, I., Tsagari, D. and Kyriakou, N. (2021). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. *Language Testing*. https://doi.org/10.1177/0265532220940960.

Marra, A., Buonanno, P., Vargas, M. *et al*. (2020). How COVID-19 pandemic changed our communication with families: Losing nonverbal cues. *Critical Care* 24: 297. https://doi.org/10.1186/s13054-020-03035-w.

McNamara, T. and Roever, C. (2006). *Language Testing: The Social Dimension*. Malden, MA and Oxford, UK: Blackwell Publishing Ltd.

Meurers, D., De Kuthy, K., Nuxoll, F., Rudzewitz, B. and Ziai, R. (2019). Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics* 39: 161–88. https://doi.org/10.1017/S0267190519000126.

Mirhosseini, S. A. and De Costa, P. (eds.). (2020). *The Sociopolitics of English Language Testing*. London, UK and New York, NY: Bloomsbury.

Nakatsuhara, F., Inoue, C., Berry, V. and Galaczi, E. (2017). Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly* 14: 1–18. https://doi.org/10.1080/15434303.2016.1263637.

Ockey, G. J., Gu, L. and Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly* 14: 346–59. https://doi.org/10.1080/15434303.2017.1400036.

Pill, J. and Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing* 30: 381–402. https://doi.org/10.1177/0265532213480337.

Shohamy, E. G. (2001). *The Power of Tests: A Critical Perspective on the Uses of Language Tests*. London: Longman/Pearson Education.

Sultana, N. (2019). *Exploring the Alignment of the Secondary School Certificate English Examination with Curriculum and Classroom Instruction: A Washback Study in Bangladesh*. Unpublished doctoral dissertation, Queen's University, Kingston, Ontario, Canada). Retrieved from https://qspace.library.queensu.ca/handle/1974/26482.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing* 30: 403–12. https://doi.org/10.1177/0265532213480338.

Taylor, L. and Harding, L. (2020). A testing time for testing: Assessment literacy as a force for social good in the time of coronavirus. *Campaign for Social Science*. Retrieved January 19, 2021, from https://campaignforsocialscience.org.uk/news/a-testing-time-for-testing-assessment-literacy-as-a-force-for-social-good-in-the-time-of-coronavirus/.

Turner, C. E. and Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari and J. Baneerjee (eds.), *Handbook of Second Language Assessment*. Boston and Berlin: De Gruyter, Inc, 255–72.

Yan, X. and Fan, J. (2021). "Am I qualified to be a language tester?": Understanding the development of language assessment literacy across three stakeholder groups. *Language Testing* 38. https://doi.org/10.1177/0265532220929924.

Index

Note: **Boldface** page references indicate tables. *Italic* references indicate figures.

Abdulhamid, N. 4, 124–125, 140, 142, 144
Abedi, J. 8, 309, 313, 315
Abeywickrama, P. 431
academic language 161
academic language proficiency framework 583–584, **584**
accents: listening assessment and 223, 228–229; non-standard 85; raters and 328, 332, 568; second language and 228, 484; USNT and 104
accessibility of assessment 307
accessibility and fairness 570
ACCESS test 554
accommodations: bilingual dictionary/glossary 313; computer-based testing 315–316; customized dictionary 313–314; defining 311; effectiveness of 308–309; English dictionary 312–313; English glossary 313; English Language Learners and 306–307, 310–317; extended time 312; fairness and 8; further reading 317–318; future directions 316–317; historical perspectives 306–307; linguistically modified test 314–315; methodology of research 307–308; native language testing 314; overview 8; performance difference and 310; purpose of 306; read-aloud 309; research 311–316; universally designed 8, 307, 312; USNT 107–108; validity of 309–311; for young language learners 160
ACTFL Oral Proficiency Interview (ATCFL OPI) 331, 537
administration of test procedures 599, 609–610
Adolphs, S. 551
adult literacy 38–39
African Americans' great migration from South 107
age-appropriate assessment development/practice 155–157
Ahmed, A. A. M. 144
Alderson, C. J. 278–279, 281, 284, 450
Alderson, J. C. 20, 83, 136–139, 188, 194, 417, 539, 547, 553
aliens, illegal 101
alignment 5, 123, 139–141
Aljuaafreh, A. 178
Almond, R. 296
alpha testing 415
ALTE Quality Management System 616
alternative assessment 277–278, 346, 491
alternative explanations 290
alternative task contexts 408–411
alternative tasks measuring a defined construct, evaluating 408–411
America with Disabilities Act (1990) 114n11
American Council on the Teaching of Foreign Languages (ACTFL) 25, 143, 210, 331, 537, 579, 583–584, **584**
American Educational Research Association (AERA) 17
American Psychological Association (APA) 17
American Sign Language 108
analytic scoring 496–497, 500
analytics, learning/data 127
Anand, P. 143–144
Anderson, C. 361
Anderson, J. C. 261
Andujar, A. 179–180
Anthony, L. 555
anti-psychometric approaches 343
Antón, M. 5, 178
appraisal stage of validation 35
Aptist speaking test 210–211, 216
aptitude 53–57, *54*, *55*, *56*, 292, 332
argument-based validity: appraisal stage of validation and 35; audiences for 23–24; claims in 21, **22**; developing validity argument and 35–37; development stage of validation and 34–35; evaluation of test and 610; expansion of use of concepts 27; framework 32–35, 50; further reading 45–46; goal of 44; inference from test scores to end-of-course grade 43; inference from test scores to level of achievement and 33, 42–43; interpretation/use

argument and 32–33, 35–36, 44; interpretive inference and 50–51; Kane's perspective 22; observable attributes and 37–40; overview 2–3; performance assessments and 39; placement decisions and 43–44; placement testing and 41–42; praxis orientation 21, 26; standardized assessments and 39–40; target domains and 37–38, 40–41; TOEFL iBT and 26; universe of generalization for language proficiency and 39–41; validity argument and 32–36; validity theory and 44–45
Arieli-Attali, M. 298
Arizona's Proposition 203 99
artificial intelligence (AI) assistants 351
artificial intelligence (AI) literacy 11, 524–525
artificial neural networks (ANNs) 525
Aryadoust, V. 472
Ascención, Y. 360, 362
Asher, H. B. 51
assembly model 295
assessment: accessibility of 307; alternative 277–278, 346, 491; as argument 289–290; bottom-up approaches to 156–157; bottom-up approach to 156–157; constructs 239; criteria 496, 501; delivery layer **291**, 296; design 290; digital 611–615; discrete 431; embedded 431; embedding 625; engineering 290–291; formative 120, 122, 157, 188; implementation layer 295–296; integration of teaching and learning and 178; interconnectedness with teaching and learning 157–159; learning and, connecting 625–626; peer 177–178; portfolio 129; program 180–181; rubrics and 302; standardized 39–40; summative 119–120; symptomatic 171–172; tasks 192–193, 495; top-down approach to 156; Toulmin diagram for 292–293, *292*; vertical discrepancy 189; *see also specific type and test*
assessment-centred language teaching 125
assessment cycle 599, 605–607, *608*, *613*
assessment as learning (AaL) 121
assessment for learning (AfL) 120, 122, 147, 157–158, 279
assessment of learning (AoL) 120
Assessment of Performance toward Proficiency in Languages (AAPPL) 143
Assessment Reform Group 120, 122
assessment use argument (AUA) 300, 418, 535, 599
Association of Language Testers in Europe (ALTE) 416, 603, 619
asylum seekers 74, 98
at-risk learners 174–175
Attali, Y. 332
audio options for items 402
audio-visual texts 226–227
Aull, L. 550
aural performance assessment 582–583, **583**
authentic approach 343
authenticity: defining test 20, 231, 276; in diagnostic assessment 197; elicited imitation tasks and 211–212; independent tasks and 212; of items 92, 351; of language in items 92; needs analysis and 363; reading assessment and 259; test usefulness and 276; textual 347, 350, 626; TLU and 276, 344; virtual environments and enhancing 214; writing assessment and 359
automated essay scoring (AES) 242
automated item generation (AIG) 347–348, 352
automated scoring 11, 242, 504, 554; *see also* automated scoring of performance tests, validity of
automated scoring of performance tests, validity of: artificial literacy among stakeholders and, promoting 524–525; conceptual work 516–518; critical issues 515–516; empirical work 518; frameworks for using in large-scale testing versus classroom assessment 525–526; further reading 526–527; future directions 522–526; historical perspectives 515; impact of 526; language learning/teaching in different contexts and, impact of 523–524; overview 11–12; perceptions 513–514, 523; practice recommendations 518–522; relevant validity questions and, asking 518–521; research 513, 516–518; standards for 522; test-taking strategies adapted to 523; use of tests and 513, 526; validation priorities and, determining 521–522
automated speech evaluation (ASE) systems 211
automated speech recognizer (ASR) 538
automated writing evaluation (AWE) technology 244, 524
aviation English 82, 84–92; *see also* language for specific purposes (LSP) testing
Aviation English Language Test Service (AELTS) 83
Aviation Safety Net website 89

Baba, K. 362
Bachman, L. F. 18, 20–22, 24, 27, 32, 48, 66, 138–139, 210, 272–274, 276, 279–280, 282, 290, 344, 349, 417, 450, 534–535, 537, 553, 586, 590, 598–599, 603
backing for warrant 289
backward design 125
Bailey, K. M. 137, 450
Baker, B. A. 504
Baker, P. 547
band descriptors 496
Banerjee, J. 503
Bangalore Project 122
Bangladesh 141, 628
Barkaoui, K. 504
Barker, F. 546–547, 552

Baron, A. 551
Baron, P. 160
basic interpersonal communication skills (BICS) 161
basic understanding (BU) items 402–403
Batty, A. 230, 472
Baum and Titone 55
Bax, S. 198, 257, 261
Beglar, D. 431
Behrens, J. T. 295–296
Bejar, I. 498
Bejar, I. I. 516–517
benchmark movements 122–123
Bennett, R. E. 516–517, 531–534, 536, 539, 626
beta prototyping 415
Bhattarai, Y. B. 140, 142
bias: construct-irrelevant variance and 41; cultural 25; fairness and absence of 105, 107; item 260, 472, 484–485, 569; item writers/writing and 348; MFRM analyses 472; rater 8, 11, 242, 327–328, 331, 342, 471, 474, 502, 568, 571; in research 457; self-report 506; systematic test 416; task 471; test 416, 569–570, 572
Biber, D. 547, 551
Big Data 614
"bilingual advantage" 55
bilingual dictionary/glossary accommodation 313
Bi, N. Z. 376
Birenbaum, M. 123
Black, P. 122
Black, P. J. 199
Bloom, B. S. 119, 290
Boals, T. 161
Bormuth, J. R. 290
bottom-up approaches 156–157
bottom-up processing 255
Bowles, M. 422
Breen, M. 121–122
Breland, H. 536
Brennan, R. L. 479, 483
Bridgeman, B. 49
British Academic Spoken English Corpus 546
British Standards Institute (BSI) 600–601
Brown, A. 213, 323, 326, 332
Brown Corpus 546
Brown, H. D. 431
Brown, J. D. 10, 422–423, 448, 450, 486
Brunfaut, T. 7, 138, 198, 229–230, 257, 261, 348, 536, 623, 627
Bucholtz, M. 70
Buck, G. 223–224, 228
Bureau of Immigration and Naturalization 102
Bureau for International Language Coordination (BILC) 582
Burke, A. M. 162
Burton, S. J. 262
Butler, Y. G. 5, 627
Bygate, M. 122

cake format 172–173
Call Center corpus 546
Cambridge Assessment 420
Cambridge Certificate of Proficiency in English (CPE) 210, 323
Cambridge English proficiency tests 8, 553
Cambridge ESOL suite of tests 227, 257, 331, 350
Cambridge General English examinations 216
Cambridge Learner Corpus 546–547
Cambridge Local Examination Syndicate 210, 449, 498
Cambridge Main Suite examinations 210
Cambridge Michigan Language Assessment 159
Canada 140, 190–191, 382, 586, **587**
Canadian English Language Proficiency Index Program-General (CELPIP-General) 382
Canadian Language Benchmarks (CLB) 586, **587**
Canagarajah, S. 73
can-do inventories 121
Carroll, J. B. 582
Carroll, J. M. 359
Carter, R. 227
Carver, R. P. 255
case studies *see* research
category response curves (CRCs) 470–471, *471*
CEFR Self-Assessment Grid 496
Census Bureau 99
Center Listening Test (CLT) 227
Central Board of Secondary Education (CBSE) 141
central document register 607
central processing core 256
Certificate of Proficiency in English 498
Chalhoub-Deville, M. 51, 65–66, 161, 531
Chang, A. C.-S. 385
change, managing 615–616
change processes 616
Chan, S. 257
Chapelle, C. A. 2, 19, 23, 27, 48, 50–51, 66, 139, 300, 325–326, 402, 417–418, 420–421, 526, 530, 535–536, 539–540
Chapman, J. W. 255
Chartered Quality Institute 602
checklists 497, 500, 502
Cheng, L. 3, 5, 26, 121, 137–138, 140, 143, 145, 147, 628
Chen, L. 140
Cheville, J. 514
China 64, 236, 586–587, **588**
China Standards of English (CSE) 586–587, **588**
Chinese EFL students 388
Chinese Exclusion Act (1882) 100–101, 114n3
Choi, I. 158, 329, 360
Chomsky, N. 65

Cho, Y. 49, 360–361
Chujo, K. 551
Chung, Y.-R. 526
Cisco Networking Academics 296
citizenship *see* immigration and citizenship
civic nationalism 110, 572
Civil Rights Act (1964) 99
Civil Service Commission 581
Cizek, G. J. 189
claims 21, **22**, 187, 289, 298, 345
CLAN online program 365
Clarke, D. 196
classical test theory (CTT): advantages of 457–458; caveats 458; critical issues 449–450; descriptive statistics in 457; development of 447; disadvantages of 458; in education 448; errors in measurement and 447–448; further reading 459–460; future directions 458–459; historical perspectives 448–449; item analyses and 451–452; item variance and 449; in language testing 448–449; methodology of research 451–457; normal distribution in 457; overview 10; parallel tests and 449–450; practice recommendations 457–458; in psychology 448; reliability estimates and 452–455; research 450–451; standard error of measurement and 455–456; statistical performance and 423; validity analyses and 456–457
classification of test takers 487, **487**, **488**
classroom-based assessment (CBA): alignment and 123; assessment-centred language teaching and 125; assessment of learning and 120; assessment for learning and 120, 122; assessment as learning and 121; critical issues 123–125; defining 119–120; diagnostic 125–126; diagnostic assessment and 125–126; evidence of learning and, reaffirming value of 127–128; external testing and 123–124; further reading 130–131; future directions 128–130; historical perspectives 121–123; language teaching 125; large-scale standards-driven tests and 123; outcomes-based approach and 125; overview 4–6; practice recommendations 128; quality management in test production and administration 597–598; reliability and 491; research 125–127; technologically mediated 126–127; technology and 128–129; terminology, newer 120–121; transdisciplinary research and 129–130
classroom-based diagnostic assessment (CBDA) 125–126
classroom-based language assessment (CBLA) 277–280
Clauser, B. E. 517
CLAWS part of speech tagger 548, *548*
Cleveland, Stephen Grover 103
cloze format 257, 343, 376–377, 388, 390
clustering of test-taking strategies 389–390
COCA sampler 551
Code of Fair Testing Practices in Education 569, 603
co-evaluation hypothesis 57
cognitive academic language proficiency (CALP) 161
cognitive aptitude 53–57, *54*, *55*, *56*
cognitive development 171
cognitive diagnostic assessment (CDA) 158, 193–194
cognitive processing model for reading comprehension 256
cognitive validity 372–373
Cohen, A. D. 9, 627
Cohen's kappa 482
Coh-Metrix online program 365, 552
Cole, N. 570–571
collaborative writing 246–247
College Board 275
College English Test (CET) 388
College English Test Band 4 (CET4) 142–144
College English Test-Spoken English Test (CET-SAT) 275, **276**
College Scholastic Aptitude Test (CSAT) 386
Collins, A. 195
collocations 442, 551
Common European Framework of Reference (CEFR): bottom-up approach and 156–157; descriptors in 501; different proficiency levels and 554; discriminant analysis and 52; impact of 75; language proficiency measurement standards and 584–586, **585**; mapping tasks onto 106; political context of language tests and 74–75; reference levels 189; score tests and 25; vertical discrepancy assessment and 189; writing assessment and 243
Common European Framework of Reference for Languages 274
communication strategies 383
communicative competence model 81, 343
communicative language ability (CLA) 273–274, 586
communicative language teaching (CLT) approaches 121–122, 224
comparable validity 570
complexity of language testing, grappling with 626–628
componential approach to reading assessment 255
Comprehension, Attitude, Strategies, Interest (CASI) 190–191
computer adaptive tests (CATs) 471–472, 531
computer-assisted language testing (CALT)
computer-based testing (CBT): accommodations and 315–316; administering tests 425–426, 530, 611–612; collaboration of language

assessment and 540; construct representation and 539; critical issues 534–535; development of 531–534; digital assessment and 614; dynamic assessment and 534; efficiency, improving (phase 1) 531–532; further reading 540–541; future directions 539–540; historical perspectives 530–534; innovative task design and new delivery modes, introducing (phase 2) 532–533; instruction, learning, and assessment, integrating (phase 3) 533–534; integration with learning 626; learning and 532–533, 626; linking assessment with instruction and 539–540; multimedia's role in language testing and 537; multiple-choice format and 612; overview 12, 530; paper-based tests versus 535–536; practice recommendations 539; research 535–539; second language acquisition research and 540; speaking assessment and 537–539; task types and 611–612; technology and 530; virtual reality in language testing and 537; vocabulary tests and, piloting 442
computerized dynamic assessment (CDA <or C-DA>) 5, 178–180, *179*, **179**, 534
computer testing *see* computer-based testing (CBT)
concept-based instruction (CBI) 180–181
conceptual assessment framework (CAF) **291**, 294–295, 298
conceptual scaffolding 196
concurrent validity 456
confirmatory factor analysis approach 310, 536
Congo State English Exam 379, 381–382
Coniam, D. 547
Conrad, S. 296, 551
consequential validity 565
Constitution of United States 100
constructed response (CR) 407
constructed-response items 481–485
construction-integration model 255
construct-irrelevant factors 568
construct-irrelevant variance 41, 262, 496
construct representation 261–262
constructs: alternative tasks measuring defined 408–411; assessment 239; corpus linguistic 553; defining 24, 66, 296, 363, 568; enhancements to existing, evaluating 408–409; language 24, 26; multilingual 161; prototyping new item types and 401–408; social 65–66, 76; speaking assessment and measured 211–214; test 273–275; underrepresentation 570; validity 456; viability of measuring new and expanded 401–408; writing assessment 239, 241; writing integrated tasks 362–363
contamination 9
Content Language Integrated Learning (CLIL) 161, 358
content-responsible tasks 363–364, 407
content validity 456
context alternative task 409
context in evidence-centered design 297
contextualization cues 69
contributions, current *see* research
conventional test items, piloting 431–433
conversation analysis (CA) 215–216
Coombe, C. 425
cooperative writing 246–247
core processes 605–606, *606*
Corpus of Contemporary American English 89
corpus linguistics: analyses 550–552, 555; annotation of 547–549; automated scoring and 554; caveats 555–556; collocation and 551; constructs 553; critical issues 547–553; defining 545; development of 547–549; feedback tools and 554; further reading 556–557; future directions 556; historical perspectives 545–547; keyness in 550; language at different levels of proficiency and 554; language tests/testing and 556; longitudinal learner corpora and 556; n-grams and 551; overview 12, 546; practice recommendations 554–556; research 545, 553–554; technology and 556; test development/validation and 552–553; tools 555–556; types of 546; validation of test and 554
Corpus Resource Database 555
Corpus of Supreme Court Opinions 555
Corrigan, M. 420
Cortes, V. 551
Council of Europe 585
Courtney, L. 164
covariance 481
COVID-19 epidemic 4, 6, 351, 533, 536, 623, 625
Crawford, J. 98
Cremer, M. 442
criterion measure 456
criterion outcomes 49–50
criterion-referenced measurement (CRM) 271, 282–283
criterion-referenced testing (CRT) 125, 128, 448–449
criterion-related validity 310–311, 456
critical approach 345–346
critical issues: automated scoring of performance tests, validity of 515–516; classical test theory 449–450; classroom-based assessment 123–125; computer-based testing 534–535; corpus linguistics 547–553; dynamic assessment 172–174; evidence-centered design 296–297; IRT 463–471; item writers/writing 346–349; language for specific purposes testing 82–84; listening assessment 224–229; performance test scoring 499–503; pre-operational testing 417–420; quality management in test production and administration 602–607; raters/rating 324–327; reading assessment 257–258;

speaking assessment 211–215; test-taking strategies 378; validity 22–26; writing assessment 237–244; writing integrated tasks 363–367; young learners' assessment 155–159
critical language testing 345, 568
Crocker, L. 148
Cronbach, L. 138
Cronbach, L. J. 37, 44
Cronbach's alpha 453–454, 480–482
Crosthwaite, P. R. 552
Crump, P. 420
"cult of ethnicity" 97
cultural bias 25
Cumming, A. 360–361
curriculum-based standards 586–587
curriculum integration 181
Cushing, S. T. 12, 623
customized dictionary 313–314
CVRi 264

Dagneaux, E. 549
DALF French language proficiency test 364–365
Damböck, B. 27, 272, 279–280
data truncation problem 49–50
Davidson, F. 4, 7, 76, 195, 271–273, 275, 283–285, 290, 344–346, 372, 417, 565–566, 599, 605, 610, 623, 629
Davies, A. 19, 281, 430, 566–567, 574–575
Davies, J. A. 51
Davies, M. 555
Davin, K. J. 180
Davis, L. 8, 324–325, 329, 421, 627
Deane, P. 242
De Ayala, R. J. 465, 473
Declaration of Independence 100
deep learning (DL) 525
deep-thinking processes 200
de Jong, N. H. 217–218
Deming, W. E. 600
dependability: concept of 486–487; further reading 491–492; future directions 490–491; overview 11, 477; *see also* reliability
depth of vocabulary (DVK) test 436
descriptional inadequacy 579
descriptive statistics 260, 457
design of language tests for specific social uses *see* language for specific purposes (LSP) testing
design of tasks *see* task design
De Valle, S. 101
developmental portfolio 248
Developmental Reading Assessment (DRA) 190–191
development stage of validation 34–35
Deville, C. 531
Deygers, B. 3, 13, 628
diagnostic assessment: advancements in 187, 197–198; assessment tasks and 192–193; authenticity in 197; characteristics of 188–193; classroom-based assessment and 125–126; cognitive 193–194; distractors and 192; dynamic assessment and 172, 178; feedback, use of 195–196; further reading 200–201; future directions 198–200; historical perspectives 187; inferences and 188–190; mediation and 196–197; misconceptions about 188; overview 6; performance assessment and 187; purposes of 188; quality of 199–200; research 199; scaffolding and 198; specificity level and 192; standardized tests 190–192; teachers' assessment competence and 199; technology in 197–199; TOEFL and 190; vocabulary tests and, piloting 442; washback effects of 194–195
diagnostic competence 162
diagnostic judgments 189
diagnostic language assessment (DLA) 278
DIALANG system 187, 190, 532
dialogic mediation 196
dictionaries as accommodations 312–313
Diederich, P. B. 323
differential groups study 456
differential item functioning (DIF) studies/analysis 260, 416, 472, 569
digital assessment 611–615
Diplôme approfondi de langue française (DALF) 364–365
Direct Method approach 210
direct range restrictions 486
discourse analysis 523, 551
discourse synthesis 362
discrete assessment 431
discrete multiple-choice format 489
discrete-point testing 223–224
discriminant analysis 52–53
discursive practice, language testing as 66–71
distractor efficiency analysis 452
distractors 192, 452, 552
diverse needs of children 159–162
diversification of employee teams and quality management 604–605, 612
diversity, language tests in suppressing 73–76
Document 4444 (IACO) 84
Document 9835 (ICAO) 82–83
domain analysis 291–293, **291**, 296
domain modeling **291**, 292–293
Dörnyei, Z. 422
Doughty, C. J. 54
Douglas, D. 225, 329, 530, 535, 539–540
Downing, S. M. 599
drag-and drop responses 409–410
Dred Scott v. Sandford 100
Duijm, K. 329
Dunlea, J. 504
Dunning-Kruger effect 195–196

Duolingo English Test 623
Dutch learners and word associates 438–439, *438*
dynamic assessment (DA): approaches to 172–173; assistance and 158; assumptions underlying 172; at-risk learners and 174–175; computer-based testing and 534; computerized 5, 178–180, *179*, **179**, 534; critical issues 172–174; defining 171–172; diagnostic assessment and 172, 178; further reading 182; future directions 181–182; goal of 172; group 177–178; historical perspectives 171–172; interactionist 173, 180–181; interventionist approaches to 173, 180–181; learner corpora and 127; Lidz's definition of 172; overview 5; peer assessment and 177–178; practice recommendations 181; program assessment and 180–181; prompts for classroom 177, **177**; research 174–182; role in assessment 173–174; second language and 174–177; similar approaches to 174; teacher education/development and 180–181; test-taking strategies and 392; theoretical background 171–172
dyslexic students 8, 309

Earl, L. 121
Eberharter, K. 10, 430
Eckes, T. 463
Edelenbos, P. 162
Edgeworth, F. Y. 498
educational measurement 271
educational performance indicators 13, 580–581, **580**
Educational Testing Service (ETS) 297, 330, 374, 379–380, 399, 547, 587–588, **589**
education and classical test theory 448
effect-driven language testing 282–284, *283*, 629–630
Elder, C. 3, 19, 84–85, 87, 217
electro-encephalography (EEG) technology 506
Elementary and Secondary Education Act (ESEA) 75
elicited imitation (EI) tasks 211–212
embedding assessment 625
Embretson, S. E. 290–291
empirical approaches to inferences about test validity 48–49
empirically derived, binary-choice, boundary definition scales (EBBs) 497, 500–502
empirical studies 139, 242, 348, 359
employee skills and quality management 604–605
engineering assessment 290–291
English for the Children (Proposition 203) 99
English dictionary accommodation 312–313
English as a Foreign Language (EFL) 386–388, 409, 484–485
English glossary accommodation 313
English Language Development Classroom Assessment 27
English Language Institute (ELI) 434
English Language Learners (ELLs): accommodations and 306–307, 310–317; evidence-centered design tests and 297; increase in 316; in Libya 144–145; NCLB and 75
English as a Lingua Franca (ELF) 626, 628
English proficiency tests 159; *see also specific name*
English as a Second Language (ESL) 486, 567; *see also* second language (L2)
English for specific purposes (ESP) context 537
Ennis, R. H. 489
Enright, M. 403, 418
ePortfolios 126, 128–129
equity *see* fairness
Erben, T. 180
Erickson, F. 64
error analysis (EA) 549
errors in measurement 447–448, 478
error variance 447
estimation procedures 469–470
Estival, D. 85
Estonia 71–72, 76
Estonian Law on Citizenship 71
ethics/ethical codes: concept of 567; defining *ethics* and 563; digital assessment and 615; fairness in language testing and 567–570; further reading 575; future directions 572–573; ILTA 566; in language testing 345, 563–567; objective approaches to 564–565; overview 12–13; practice recommendations 573–575, *574*; research 570–572; subjective approaches to 564
Etzioni, A. 112
European Association for Language Testing and Assessment (EALTA) 416, 619
European Language Portfolio for Young Learners 158
European Survey on Language Competences 500
evaluation component of evidence-centered design 295
evaluation inference 418
evaluation strategies 388
Evans, B. A. 75
Every Student Succeeds Act (ESSA) 187
evidence: English language learners and 297; evidence-centered design and 298; integration of 26; model 295; for validity 18–20, **20**; writing assessment 238–241, **238**
evidence-centered design (ECD): assessment as argument and 289–290; assessment delivery layer and **291**, 296; assessment engineering and 290–291; assessment implementation layer and **291**, 295–296; assessment use arguments

and 300; claims and 289, 298; conceptual assessment framework and **291**, 294–295, 298; constructs and, defining 296; context and, role of 297; critical issues 296–297; defining 289, 291–296; design patterns 293, **293**, **294**; domain analysis and 291–293, **291**, 296; domain modeling **291**, 292–293; evaluation component and 295; evidence and 298; expansions to 297–298; further reading 302–303; future directions 301–302; historical perspectives 289–291; language tests developed using 297; measurement model and 295; methodology of research 298–300; modularity and 299–300; overview 7–8; practice recommendations 300–301; research 297–298; reusability and 299–300; systematic approach of 344; test specifications and 271, 290
Ewers, N. 309
examiners: community of 323–324; input language of 216; latitude of, broad 323; naturalization 102–104, 107, 110–111; standardising rating judgments of 210, 323; training/certification of 331, 609; video-conferencing situations and 538; *see also* raters/rating
expeditious reading 382
explanation inferences 418
extended time accommodation 312
external testing 123–124
extrapolation inference 40, 418
eye-tracking technology 198, 231, 261, 332, 506, 623
Eynon, B. 126

FACETS software 463
factor analysis study 457
Fairburn, J. 504
FAIR Guiding Principles 615
fairness: accessibility and 570; accommodations and 8; argument 571; bias and, absence of 105, 107; concept of 567; construct-irrelevant factors 568; defining 567–570; folk 567; in language testing 345, 563–567; micro-analytic approach and 573, 575; overview 12–13; procedural 570; procedural due process and 570; reliability and 487–490; research 572–573; social consequences and 573; substantive 570; substantive due process and 570; test 567; as test characteristic 569; Test Context Framework and 570–573; in test use 27; USNT and principles of 106–109; validity and 487–490, 569–571; variability and 8; washback and 148; *see also* ethics/ethical codes
Fan, J. 11, 625
Farrington v. Tokushige (1927) 99
Feldt, L. S. 479, 483
Feng, M. 298
Ferne, T. 472
Feuerstein, R. R. 172–173
Field, J. 227
field testing 9–10, 273, 400, 415, 417, 419, 422, 424
Filipović, L. 549
first language (L1) 153, 256
fit-for-purpose scoring instruments 505
Fitzgerald, K. R. 514
folk fairness 567
footing 70
Ford, C. E. 68
foreign languages
Foreign Service Institute (FSI) 210, 323, 498–499, 581–582
formal standards 579
formative assessment 120, 122, 157, 188
form-meaning-use mapping 297
Foucault, M. 70, 72, 74
four-skills approach to language testing 225
14th Amendment 100
Fox, J. 4, 121, 125, 140
Francis, D. J. 255, 264
Freedle, R. 230, 258
French learners in the Netherlands 437–438
Friginal, E. 546, 553
Froehlich, V. 140, 277
Froetscher, D. 139–140
Fulcher, G. 72, 75–76, 195, 209–210, 271–275, 281–284, 344–345, 401, 417, 430, 496–497, 499, 503, 531, 539, 565–566, 584, 599, 605, 610, 623, 629
full-test scoring 497
further reading: accommodations 317–318; argument-based validity 45–46; automated scoring of performance tests, validity of 526–527; classical test theory 459–460; classroom-based assessment 130–131; computer-based testing 540–541; corpus linguistics 556–557; dependability 491–492; diagnostic assessment 200–201; dynamic assessment 182; ethics/ethical codes 575; evidence-centered design 302–303; immigration and citizenship 113; inference 57–58; IRT 475; item writers/writing 352–353; language proficiency measurement standards 591; listening assessment 231–232; LSP testing 93–94; performance test scoring 506; prediction 57–58; pre-operational testing 427; prototyping new item types 413; quality management in test production and administration 618; raters/rating 333; reading assessment 264–265; reliability 491–492; social dimensions of language testing 77–78; speaking assessment 218; test specifications 285–286; test-taking strategies 393; validity 28; vocabulary tests, piloting 442–443;

washback 148; writing assessment 248–249; writing integrated tasks 367–368; young learners' assessment 164–165
future directions: accommodations 316–317; automated scoring of performance tests, validity of 522–526; classical test theory 458–459; classroom-based assessment 128–130; computer-based testing 539–540; corpus linguistics 556; dependability 490–491; diagnostic assessment 198–200; dynamic assessment 181–182; ethics/ethical codes 572–573; evidence-centered design 301–302; IRT 474–475; item writers/writing 351–352; listening assessment 231; LSP testing 92–93; performance test scoring 505–506; pre-operational testing 426–427; prototyping new item types 412–413; quality management in test production and administration 617; raters/rating 332–333; reading assessment 263–264; reliability 490–491; social dimensions of language testing 76–77; speaking assessment 217–218; test specifications 284–285; test-taking strategies 392; validity 27; vocabulary tests, piloting 442; washback 147–148; writing assessment 244–248; writing integrated tasks 367; young learners' assessment 163–164

Gablasova, D. 547
Galaczi, E. D. 11, 215, 217, 331, 500, 503
Gambino, C. 99
Gambino, L. 126
Ganschow, L. 174
gap-fill tasks/tests 257, 386
Garcia, J. 110
Garcia, P. N. 5, 180–181
Gass, S. 422
Gebril, A. 361–362
General Data Protection Regulation (GDPR) 615
General English Proficiency Test-Intermediate (GEPTI-S) 387
generalizability 67, 418, 506, 535
generalizability theory (G-theory) 329, 448–449
generalization inference 418
Georgetown University Round Table (2016) 420
GEPT 143
Ginther, A. 49
globalization 154
Goegan, L. D. 307
Goffman, E. 70
Goodwin, S. 547
Google Docs 7, 246
Grabe, W. 240, 263
Graduate Record Examination (GRE) 513
Graham, S. 164
Great Society 75
Green, A. 137, 142, 280, 346–347
Greene, K. 175
Greidanus, T. 437, 439–440
Gries, S. T. 550
Grigorenko, E. L. 172
group-based analysis 52
group-differentiated rights 97
group dynamic assessment 177–178
Guidelines for Good Practice (EALTA) 416
Guidelines of the International Language Testing Association (ILTA) 416
Gu, L. 156
Guttman, L. 448
Gu, Z. 192
Gyllstad, H. 442

Haladyna, T. M. 290, 599
Halleck, G. B. 4, 85, 87, 545, 547, 550, 623, 626, 628
Halliday, M. A. K. 67
Hall, K. 70
Hamp-Lyons, L. 237, 327, 345, 500, 563–566
Harding, L. 194, 625
Hargreaves, P. 605
Harrington, M. 442
Harsch, C. 503
Hartwick, P. 127
Hatcher, W. S. 564–566
Hatto, P. 579
Hawkey, R. 346–347
Hawkins, J. A. 549
He, A. W. 70
Heitink, M. C. 120
Henning, G. 463
Henson, R. 194
Heritage, J. 216
Heyworth, F. 602
high-language-proficiency (HLP) 385–386, 388
high-stakes tests 272–-277, 400, 413n1
Hill, K. 122, 236
Hirai, A. 497, 503
historical perspectives: accommodations 306–307; automated scoring of performance tests, validity of 515; classical test theory 448–449; classroom-based assessment 121–123; computer-based testing 530–534; corpus linguistics 545–547; diagnostic assessment 187; dynamic assessment 171–172; evidence-centered design 289–291; IRT 462–463; item writers/writing 342–346; listening assessment 223–224; performance test scoring 497–499; pre-operational testing 416–417; quality management in test production and administration 598–601; raters/rating 323–324; reading assessment 254–257; speaking assessment 209–211; test-taking strategies 376–377; validity 18–22; vocabulary tests, piloting 431; washback 137–138; writing assessment 236–237; writing integrated tasks 358–359; young learners' assessment 153–155

Hively, W. 290
holistic scoring 496
Holland, P. W. 459
Holton, D. 196
Holzknecht, F. 10, 430
homogeneity 97, 99, 481, 502
Honer, S. M. 563–564, 566
Hong Kong Certificate of Education Examination (HKCEE) 137
Hoque, M. E. 141
Horák, T. 145
Hornberger, N. H. 75
Hoskens, M. 459
Hsieh, C. N. 156, 329
Huang, H. T. 360
Hudson, T. 486
Hughes, A. 137, 431, 450
Huhta, A. 92
humanist-scepticist descriptive approach 488
Hung, S. T. 360
Hunter, D. M. 214
Husserl, E. 67
Hyland, K. 239–240
Hymes, D. H. 343

ideational metafunction 67
IELTS Indicator 533
ILTA Code 566
Immigration Act (1924) 101
immigration and citizenship: African Americans' great migration from South 107; aliens and, illegal 101; asylum seekers and 74, 98; "cult of ethnicity" and 97; educational requirements and 102; further reading 113; Immigration and Nationality Act and 101–102, 111; involuntary immigrants and 96–97; language assessment and 101–105; language policy and 97–98; monolingual language policy in US and 97–99; naturalizations from six leading countries 109, **109**; overview 4; permitting 99–101; restricting 99–101; transnational immigrants and 96; US policies and 99–102, 113n1; voluntary immigrants and 96–97; xenophobia and 99, 113n1; *see also* US Naturalization Test (USNT)
Immigration and Nationality Act (INA) (1952) 101–102, 111
Immigration Reform and Control Act (1986)102, 114n5
Immigration Restriction League 103
impartiality 418, 535
Inbar-Lourie, O. 564–565
independent tasks 212, 359–360
Indian Citizenship Act (1924) 101
inductive inferences 37
Infante, P. 176
inferences: approaches to 48–49; criterion outcomes 49–50; diagnostic 188–190; evaluation 418; explanation 418; extrapolation 40, 418; further reading 57–58; generalization 418; inductive 37; interpretation/use argument and 33; interpretive 50–51; overview 3; placement decisions 43–44; placement testing 41–42; from test score to end-of-course grade 43; from test scores to level of achievement 33, 42–43; test validity 48–49, 57; validity 610–611
inflection points 467
information integration 325
information transfer 257
In'nami, Y. 258
Inoue, C. 3, 6
inputs 600
institutionalized language tests 13
instruction, composing 365–366
integrated reading testing 263–264, 379–380
integrated skills 357–367, 406–408
integrated tasks 9, 212–213, 263–264, 360, 367, 379–380
integrated writing tasks *see* writing integrated tasks
integration: of ACTFL into US educational system 584; of computer-based testing with learning 626; of curriculum 181; of evidence 26; of information 325; of language skills 357–367; of models 50; social 97, 572; of teaching, learning, and assessment 178
integrative testing 224
interactional resources 68
interactionist approaches 173, 180–181
interactive approach to language testing 225
interactive compensatory model 255
interactive rubrics 128
interactive speaking tasks 213
interactive task types 213–214
Interagency Language Roundtable (ILR) 52, 210, 274, 298, 582–583
interconnectedness of teaching, learning, and assessment 157–159
interlocutors 212–213, 322–323, 326, 331, 351, 409, 538, 627; *see also* raters/rating
internal consistency 479–481
International Association for the Evaluation of Educational Achievement (IEA) 579
International Bureau of Weights and Measurements 591
International Civil Aviation Organization (ICAO) 82–86
International Corpus for Learner English (ICLE) 546
International English Language Testing System (IELTS): certification perspective 587–590; computer-based, piloting 538; language function analysis and 216; listening subset from 388–389; misuse/misperceptions about 142; preparation for 142, 388–389; purposes of 273; scoring 587–590; speaking strategies

in completing 383–384; washback and 138; writing assessment and 241
International Language Testing Association (ILTA) 154, 619
international performance assessment 579
International Standardization Organization (ISO) 578, 601
Internet-based College English Test (IB-CET) 535
Internet-based testing (iBT)
interpersonal metafunction 67
interpretation: of reliability estimates 454–455; of rubrics 8; of test scores 19, **20**, 502–503; writing assessment 243
interpretation/use argument (IUA) 2–3, 32–33, 35–36, 44
interpretive approaches to inferences about test validity 49
interpretive argument 22, 517–518
interpretive inferences 50–51
interventionist approaches 173, 180–181, 196–197
intervention study 456
introspective questionnaires 261
involuntary immigrants 96
Iranian National University Entrance Exam for the MA in TESL (INUEMA) 383
Irvine, S. H. 290–291
Isaacs, T. 218, 329
Isbell, D. 624
ISO 9000 601
ISO 9001 607
Israel, Avi 379
item characteristic curves (ICCs) 464, 466–467, *466*
item discrimination (ID) 451
item facility (IF) 451–452
item response theory (IRT): advantages of 472–473; assumptions underlying 471–472; category response curves and 470–471, *471*; computer adaptive tests and 471–472, 531; conceptualization of 463–465; critical issues 463–471; defining 462; differential item functioning studies and 472; disadvantages of 473; estimation procedures and 469–470; further reading 475; future directions 474–475; historical perspectives 462–463; item characteristic curves and 464, 466–467, *466*; language testing and 448; listening assessment and 230; MFRM and 463, 470–474; multiple-choice format and 462–464; one-parameter Rasch model and 465–467, *466*; overview 10–11; practice recommendations 472–473; probability theory and 423; research 471–472; sample size and 472; statistical performance and 423; two and three parameter Rasch models and 467–469, *469*; use of 462
items: analyses of 451–452; audio options for 402; authenticity of 92, 351; basic understanding 402–403; bias in 260, 472, 484–485, 569; characteristics of 258; constructed-response 481–485; contamination and 9; conventional, piloting 431–433; discrete 223–224; distractors and 192, 452; interaction between text and 258; selected-response 479–481, 484–485; strictly parallel 449; variance 449; *see also* prototyping new item types
item trialling 400
item tryouts 400
item variance 449
item writers/writing: anti-psychometric approaches 343; authentic approach 343; authenticity of items and 343; automated item generation and 347–348, 352; bias and 348; cloze format 343; COVID-19 epidemic and 351; critical approach 345–346; critical issues 346–349; further reading 352–353; future directions 351–352; historical perspectives 342–346; identifying writers 346–347; job description of writers 347–348, 350; multiple-choice format 342–343, 345; multiple-choice format and 341–343, 345, 347–348, 352; non-native speakers and 350; overview 8–9, 341–342; plagiarizing and 349; process 341, 347; psychometric approach 341–343; for reading test, stages of 347; research 349–351; as social practice 348–349; socio-political thinking and 345–346, 351; systematic approach 344–345; technology and 347–348, 351; test specifications and 348; textmapping and 263; training 349–351; whole picture of test development and, seeing 344–345

Jamieson, J. 531–532
Jang, E. 567
Jang, E. E. 6, 158, 190–192, 194–195, 431, 623, 626
Japanese Language Testing Association (JLTA) 619
Jeon, E. H. 258
Jewish immigrants 101
Jiang, X. 263
Jin, Y. 7, 535–536
Johnson, Lyndon 75
Johnson, R. C. 27
justice 570
justice principles 110–111

Kane, M. 499, 515, 518, 534–535, 569–570, 572–573, 575, 610
Kane, M. T. 2, 21–22, 32, 48, 50–51, 138, 144, 590
Kaplan, R. B. 240
Kaulfers, W. V. 581–582
Kelly, P. A. 517
Kempf, A. 124
Kennedy, G. 551
Kenyon, D. M. 10, 415, 417–419, 426, 430
keyboarding responses 409–410

keyness in corpus linguistics 550
Khabbazbashi, N. 3, 6, 215, 500, 503
Khalifa, H. 256–257, 264
Kim, A. Y. 193
Kim, E. Y. J. 142–143
Kim, H. 85, 87
Kim, H. J. 329
Kim, J. 284, 344, 346, 348
Kim, N. 350
King, J. H. 615
Klemke, E. D. 564
Knoch, U. 7, 92, 239–240, 247, 325–326, 329, 357, 421, 625
knowledge base 256
Koh, K. 147
Koizumi, H. 497, 503
Koizumi, R. 258
Korean EFL students 386, 409
Korean TOEFL 142–143
Košak-Babuder, M. 309
Kostin, I. 230, 258
Koyama, D. 472
Kramsch, C. 67
Kremmel, B. 10, 258, 430, 553, 624
Kubanek-German, A. 162
Kuder, G. F. 448
Kuder-Richardson (K-R) formula 453, 480–481
Kunnan, A. J. 4, 27, 104, 567, 569–570, 572–573, 575, 628
Kyllonen, P. C. 290–291
Kymlicka, W. 97–98

Lado, R. 6, 18, 209, 223
LaFlair, G. T. 554
Lane, S. 599
language: appropriateness of 92; aptitude 53–54, *54*; constructs 24, 26; defining 358; development 25, 126–127; early, in China 64; national 72–73; native 99; policies 97–99, 279; real-world spoken 227–228; social life of 64–65; vocabulary and 430–431
language assessment 101–105; *see also* language tests/testing
language assessment literacy (LAL) 124, 162
Language Assessment in Practice 22
Language Assessment Quarterly 154
language function analysis 216
language learner strategies 373–374
language proficiency: English 99; frameworks 581–587; high 385–386, 388; levels 585, **585**; low 385–386; target domains and 40–41; test-taking strategies and 383–387; universe of generalization for 40–41; *see also specific test*
language proficiency measurement standards: ACTFL and 583–584; Canadian benchmarks and 586, **587**; can-do perspective 581–587; CEFR and 584–586, **585**; certification perspective 587–590; Chinese standards of English and 586–587, **588**; criteria of standards and 590–591; curriculum-based 586–587, **588**; definition of *standard* and 578–579; descriptional inadequacy and 579; educational performance indicators and 580–581, **580**; in foundational years (1950s-1970s) 581–582; frameworks 581–587; further reading 591; impact of 578; large-scale language tests 587–590, **589**; overview 12–13; policy perspective 580–581; Standardization Agreement 6001 and 582–583, 591n1; task-based 586, **587**; types of 579–580, 590; use of 578
language skills 6–7, 357–367; *see also* language tests/testing
language for specific purposes (LSP) testing: aviation context questions 91–92; aviation English 82, 84–92; communicative functions and 86–87; critical issues 82–84; Document 4444 and 84; Document 9835 and 82–83; further reading 93–94; future directions 92–93; goal of 81; ICAO proficiency and 83, 85–86, 90–91; negotiation of meaning 88; overview 4, 81–82; phraseology 84, 92; practice recommendations 84–92; radiotelephony question 91; VAET 89, 92; work-related communication 81, 87–91
Language Teaching for Young Learners 154
Language Testing 154, 463, 518
Language Testing and Assessment 22
Language Testing Research Colloquium (LTRC) 530
language tests/testing: classical test theory in 448–449; collaboration of 540; complexity and, grappling with 626–628; constructs 273–275; corpus linguistics and 556; design features 275–277, **276**, 279–280; as discursive practice 66–71; diversity and, suppressing 73–76; effect-driven 282–284, *283*, 629–630; ethics in 345, 563–567; evidence-centered design and 297; fairness in 345, 567–570; field 273; four-skills approach to 225; interactive approach to 225–226; IRT and 448; learning and, connecting 625–626; linguistically modified 314–315; longitudinal research and, lack of 622; multimedia's role in 537; national language and, preserving prestige of 72–73; overview 3–4; political context of 74–75; prediction and 51–53; purposes 272–273, 277–278; quality management in test production and administration and 598–600; social consequences of 71–76; social constructs in 65–66, 76; socio-political nature of, theorising 628–629; specifications for high-stakes 272–277; standards and, maintaining 73; technology and, increasing role of 623–625;

test specifications and 284–285; themes 13, 622–630; usefulness of 20–21; users of 603; uses, overview of 3–4; validity and 18, 25–26, 48; values of 25–26; virtual reality in 537; washback and 137–138; *see also* language for specific purposes (LSP) testing; social dimensions of language testing; *specific name*
LanguEdge Courseware 380–381, 408
Lantolf, J. P. 65, 174, 178, 190, 534
large-scale assessment 123, 587–590, **589**, 597–598
latent variables 462
Lau v. Nichols (1974) 99
Lazaraton, A. 215
Lazarus, Emma 99
LCS patterns 297
learner corpora 127, 546
learner engagement 279
learning: assessment as 121; assessment and, connecting 625–626; assessment for 120, 122, 147, 157–158, 279; assessment of 120; computer-based testing and 532–533, 626; deep 525; integration of computer-based testing with 626; integration of teaching and assessment and 178; interconnectedness with teaching and assessment 157; motivation 52; outcomes 125; washback and 147–148
learning driven teaching 125
learning-oriented assessment (LOA) 124–125, 129, 198, 279, 540, 625
learning propensity assessment device (LPAD) 173
Lee, H. 2, 361
Lee-Williams, B. 101
Lee, Y. W. 278
Leighton, J. P. 192–193
Levinson, S. C. 69
Lewkowicz, J. A. 360
lexical density/coherence of texts 365
LexTutor website 551, *551*
Libya 142, 144–145
Lidz, C. 172
Li, J. 361
Likert-type rating rubric 315
Lim, G. S. 11, 329, 490
Lim, H. 490
limited-English-proficient (LEP) 384
Lin, C. 420
Ling, G. 536
Lingua Franca 82
linguistically modified test 314–315
linguistic interdependent hypothesis 255
linguistics 64
linguistic threshold hypothesis 255
Lipton, P. 51
listening assessment: accents and 223, 228–229; accent varieties in spoken texts and 228; audio-visual texts and 226–227; critical issues 224–229; eye-tracking technology and 231; further reading 231–232; future directions 231; historical perspectives 223–224; interactive speaking and 225–226; IRT and 230; methodology of research 229–231; MFRM and 230; multiple-choice format 463–464, 464, **464**; overview 6, 223; real-world spoken language and 227–228; second language 223–231; technology and 231; verbal reports and 230–231; washback and 229
listening proficiency scale 581–582, **582**
listening/speaking task 364–365
listening test strategies 384–385, 387, 391
literacy assessment/measurement 39, 102–103, *103*, 114n8, 147
Literacy Test Bill 103
Litman, D. 214, 217, 538
Llosa, L. 27, 300
local document register 607
Locke, J. 72
logistic modeling, binary/multinomial 52
logits 464, 466–467
Lomax, R. 359
longitudinal approach to prediction 55–57
longitudinal learner corpora 556
Lord, F. M. 448, 463
Löwenadler, J. 258
Lowenberg, P. H. 73
Lower Certificate in English exam 498
low-language-proficiency (LLP) 385–386
Lüdeling, A. 549
Luecht, R. M. 291, 298
Lumley, T. 504
Lund, A. 177
Lu, X. 179, 552
Lynch, B. K. 271, 283, 290, 344–345, 346, 565

MacGregor, D. 10, 415, 417–420, 426, 430
Mackey, A. 422
Macqueen, S. 239–240
macro analysis approach 573
Ma, J. 142
management science 600
Manalo, J. R. 536
Manual for Language Test Development and Examining, The 416
many-facet Rasch measurement (MFRM): bias analyses 472; IRT and 463, 470–474; listening assessment and 230; overview 10–11; pre-operational testing and 423; as quantitative technique 329; raters graphical feedback based on 328, 331; raters severity and 327; use of 10–11, 470
Mariel Boatlift 101
Mari, M. A. 141
marking models 214–215
marking schemes 495–497
Martinez, J. 110

matching format 257
Mathew, R. 141
May, S. 97–98
McCarthy, M. 227
MccKay, P. 153
McCray, G. 198, 257–258, 261
McElwee, S. 13, 623
McEnery, T. 547
McNamara, T. F. 26, 70–71, 75, 84–85, 92, 122, 236, 240, 343, 345, 351, 463, 490, 499
meaningfulness argument 418, 535
Meara, P. 433
measurement: criterion-referenced 271, 282–283; deviations and, accidental 447; educational 271; errors 447–448, 478; of literacy 39; model 295; norm-referenced 271; overview 10–11; physiological 478; pragmatic understanding 401–403; psychological 478; reliability and 477–478; standard error of 455–456, 481; standards 17; validity 17; *see also* language proficiency measurement standards
mediated development (MD) 176
mediated diagnostic assessment interventions 196–197
mediated language experiences (MLEs) 173, 175–177
mediation, dialogic 196
Meehl, P.E. 37, 44
Mehan, H. 67
Mell, J. 86
Memon, N. 142
Messick, S. A. 18–21, 26, 66, 138–139, 225, 284, 290, 300, 457, 517, 565–566, 573, 575, 610
metacognitive activity 256
methodological pluralism 129
methodology of research: accommodations 307–308; classical test theory 451–457; evidence-centered design 298–300; listening assessment 229–231; pre-operational testing 422–423; raters/rating 329–330; reading assessment 259–261; speaking assessment 215–217; survey 230; test-taking strategies 389–390; washback 145–147
Meurers, D. 625
Meyer v. Nebraska (1923) 99
Michigan English Language Assessment Battery (MELAB) 554
micro-analytic approach 573, 575
micro-linguistic analysis 216–217
minimum standards 616
Min, K. 110
misalignment 139–141
Mislevy, R. J. 7, 24, 50, 192, 489
mixed-methods scale development 503
Miyazaki, I. 72
models 50, 462; *see also specific name/type*
Moder, C. L. 4, 85, 87, 545, 547, 550, 623, 626, 628
modern test theory *see* item response theory (IRT)
modularity 299–300
Molesworth, B. 85
monitoring strategies 388
monolingual language policies 97–99
Morton, J. 55
Moss, P. 489
Moss, P. A. 129–130, 326, 525–526
motivation for learning 52
multi-facet Rasch measurement *see* many-facet Rasch measurement (MFRM)
multilingual constructs 161
multilingual corpora 546
multilingualism/multilingual constructs 97, 161
multimedia's role in language testing 537
multiple-choice (M-C) format: computer-based testing and 612; discrete 489; distractor efficiency analysis 452; distractors for 552; ESL placement test in UK 531; guessing in 468–469; IRT and 462–464; item writers/writing and 341–343, 345, 347–348, 352; listening assessment 463–464, **464**; pedagogical knowledge mastery and 295; prototyping 400, 402, 405; rationalist-empiricist approach and 488; reading assessment 462, 478; in second language reading assessments 257–258, 262; self-report on 389; test-deviousness strategies for 375; test-management strategies for 374; TOEFL 436; TOEFL iBT 532; US switch to 236–237; vocabulary 484–485, *484*; vocabulary tests in 431–432
multiple distractors 192
multiple referencing 285
multiple text content-responsible task 364
multiple-trait scales 497
multi-traits multimethod study 456–457
Muravyova, Olga 63, 65–66, 70–72, 76
Murray, N. L. 210
Myers, R. 329

Nakatsuhara, F. 3, 6, 538
National Council on Measurement in Education (NCME) 17
National Examinations and Qualifications Centre 71
national language, preserving prestige of 72–73
National Standards for Foreign Language Learning 584
Nation, I. S. P. 431
native languages 99
native language testing accommodation 314
native speakers (NSs) 401, 415
Naturalization Act (1790) 100
naturalizations from six leading countries 109, **109**
naturalization test *see* US Naturalization Test (USNT)

needs analysis 239–240, 363
Nepal 140, 142
Netherlands, the, French learners in 437–438
Newfields, T. 143
Newton, P. E. 18
n-grams 551
Nichols, P. D. 188
Nienhuis, L. 437
Nikolov, M. 389–390
Nissan, S. 9–10, 623
No Child Left Behind (NCLB) Act (2001) 74–75
non-native speakers (NNSs) 85, 350, 401
nonverbal resources 68
normal distribution in classical test theory 457
norm-referenced measurement (NRM) 271
Norris, J. 22
Norris, J. M. 66
North American studies of word associates 436; second language reading tests and 257
Novick, M. 448, 463
Nowlan, N. S. 127

objective approaches 48, 564–565
O'Brien O'Keeffe, K. 64
observable attributes 37–40
observed performances 40
observed scores 447
Occupational English Test (OET) 240, 273, 499–500
Ockey, G. J. 10–11, 214, 226–227, 231, 423, 472, 537–538
O'Hagan, S. R. 323
Oklahoma State University Aviation Corpus 89, 92
Oller, J. W. 18, 223–224, 273, 358
one-parameter Rasch model 465–466, *466*
online learning spaces 127
Ontario Secondary School Literacy Test (OSSLT) 26, 140
Oral Method approach 210
Oral Proficiency Interview (OPI) 210, 331, 532, 537
Organisation for Economic Co-operation and Developments (OECD) 123, 579, 590
organisation innovation 616
organisation theory 616
O'Sullivan, B. 216, 372–373
other-initiated self-repair 69
outcomes-based approach 125
outcomes, learning 125
outline completion task 404–405
outputs 600, 606
overrepresentation 161

PADI tools 298
Page, E. B. 513, 515
Pakistan 141–142
Pallotti, G. 217
Palmer, A. S. 20–22, 32, 273–274, 276, 282, 290, 344, 417, 450, 534–535, 537, 553, 586, 590, 598–599, 603
Pāṇini 64
Pan, Y. C. 143
Papageorgiou, S. 160
paper-based test (PBT) and scores 402, 405, 531, 534–536, 612
Paquot, M. 553
parallel-form reliability 479–480
parallel growth model 56–57, *56*
parallel tests 449–450
parameter 462
Park, E. 9–10, 623
Park, K. 547, 556
Park, M. 537
participant framework 70
parts of speech (POS) tagging 548–549, *548*, *549*
path analysis models 52
Paulus, D. H. 513
Pavlenko, A. 98
Pearson, K. 448
Pearson publishing company study 588, **589**, 590
Pearson Test of English (PTE) Academic 27, 210, 513, 554
peer assessment 177–178
Pekrun, R. 198
Pellegrino, J. W. 123, 130, 148
Peña, E. 175
pen-and-paper tests, administering 425
Perfetti, C. A. 255
performance assessment: argument-based validity and 39; aural 582–583, **583**; claims, evidence and inference in; diagnostic assessment and 187; international 579; *see also* automated scoring of performance tests, validity of; performance test scoring
performance data 230, 401
performance decision trees (PDTs) 497, 500–502
performance difference 310
performance-level descriptors (PLDs) 189
performance test scoring: approaches to, advantages/limitations of 500; assessment criteria and, developing 501; considerations, key 495–496; criteria 495–497; critical issues 499–503; design stage 505; development stage 505; further reading 506; future directions 505–506; generalizability dilemma in 506; historical perspectives 497–499; instruments for, choosing contextually appropriate 500; interpretation of score 502–503; mixed-methods scale development and 503; operational stage 505; overview 11; practice recommendations 504–505; rater effects and 496, 502; raters' cognition and decision-making processes 503–504, 506; rating scales and,

limitations of 501–502; reliability and 505; research 501, 503–504; "strong" performance tests and 499–500, 506; "weak" performance tests and 499–500, 506; *see also* automated scoring of performance tests; automated scoring of performance tests, validity of
Perry, G. 246
Phakiti, A. 261
physiological measurement 478
Pickus, N. 103, 107
piloting/pilot testing 400, 415, 417, 422, 424, 430, 433; *see also* vocabulary tests, piloting
placement decisions 43–44
placement testing 41–42
plagiarizing 349
Plakans, L. 9, 360–363
Plan, Do, Check, Act model 607
planning strategies 388
Plough, I. 539
pluralism, methodological 129
Poehner, M. 173–178, 190
Poehner, M. E. 534
Pollitt, A. 210
pop-up glossary on computer 315–316
portfolio assessment 129
portfolio prisons 129
positive impact by design 615
power 70–71, 76
power relations in language testing 568
Powers, D. E. 515, 533
Prabhu, R. N. S. 122
practice recommendations: automated scoring of performance tests, validity of 518–522; classical test theory 457–458; classroom-based assessment 128; computer-based testing 539; corpus linguistics 554–556; dynamic assessment 181; ethics/ethical codes 573–575, *574*; evidence-centered design 300–301; IRT 472–473; LSP testing 84–92; performance test scoring 504–505; pre-operational testing 423–426; quality management in test production and administration 615–617; raters/rating 330–332; reading assessment 261–263; test-taking strategies 390–391; writing integrated tasks 363–367; young learners' assessment 162–163
pragmatic realism 274, 281, 282–284
pragmatic strategies 376
pragmatic understanding (PU), measuring 401–403
prediction: aptitude and 53–55; further reading 57–58; language testing and 51–53; longitudinal approach to 55–57; overview 3; of second language acquisition 55
predictive validity 51–53, 57, 456
pre-operational testing: administration of 424–425; assessment use argument and 418; critical issues 417–420; data retrieval, analysis, and storage 425–426; defining 415; differential item functioning analysis and 416; digitalization of language assessment and 426–427; documentation/dissemination and 426; forms of 416; Fulcher's tripartite model and 417; further reading 427; future directions 426–427; generalizability and 418; historical perspectives 416–417; impartiality and 418; meaningfulness argument and 418; methodology of research 422–423; MFRM and 423; overview 10; planning 424; practice recommendations 423–426; prototyping and 417; relevance and 418; research 420–422, 426; Standards for Educational and Psychological Measurement and 416; sufficiency and 418; terminology 415–416; TOEFL and 418
pre-testing 400, 430
primary test users 603
primary-trait scales 497
private standards 579–580
probability theory 423
procedural due process 570
procedural fairness 570
processes, defining 606–607
process interactions map 606
process owners 607
proficiency: cognitive academic language 161; English 99, 187; FSI test 498–499; ICAO 83, 85–86, 90–91; parallel growth model for 56–57, *56*; *see also* language proficiency; *specific test*
proficiency frameworks 13
program assessment 180–181
Programme for International Student Assessment (PISA) (OECD) 123, 193, 580–581, **580**, 590
Progress of International Reading Literacy Study (PIRLS) 580–581, 590
Project Essay Grade 513–514
prototyping: beta 415; defining 415; multiple-choice format 400, 402, 405; overview 9–10; pre-operational testing and 417; reading-to-learn 406; sampling for 400–401, 424; second phase 418; task types, new 399, 403–407, *405*, *406*, 409, 412–413
prototyping new item types: alternative tasks measuring a defined construct and, evaluating 408–411; constructs and, viability of measuring new and expanded 401–408; decisions about scope and type of 399–400; further reading 413; future directions 412–413; high-level reading abilities 403–406, *404*; for high-stakes tests 400, 413n1; integrated skills 406–408; native speakers and 401; overview 9–10, 399; pragmatic understanding 401–403; sampling for 400–401, 424; scaffolding and 410–411; scoring rubrics and 411–412; terminology 400

Pryor, J. 153–154
psychological measurements and 478
psychology and classical test theory 448
psychometrics/psychometric approach 341–343, 477
PTE Academic 259
Purpura, J. E. 50, 124, 277–279

Qian, D. D. 436–437, 440
qualitative analysis 260
qualitative techniques 329
quality assurance (QA) 598, 603–604
quality control (QC) 598, 603–604
quality control checks 612–613
quality management (QM) 600–607
quality management systems 618–619
quality management in test production and administration: administration of test procedures 599, 609–610; assessment cycle and 599, 605–607, *608*, *613*; background to quality management systems and 618–619; change processes and 616; classroom-based assessment and 597–598; codes of practice and 619; critical issues 602–607; digital assessment and 611–615; diversification of employee teams and 612; employee skills and 604–605; enhanced reporting and, managing 614; further reading 618; future directions 617; historical perspectives 598–601; language tests/testing 598–600; large-scale assessment and 597–598; overview 13, 597, 617–618; practical considerations and 598; practice recommendations 615–617; process and, defining 606–607, *606*; quality assurance and 598, 603–604; quality control and 598, 603–604; quality control checks and 612–613; quality management and 600–604; quality triangle and 599–600; team development and 604–605, 612; terminology 598; test material procedures, assembling 608–609; validity and 597, 610–611
quality standards 600
quality triangle 599–600

Ramaswamy, R. 599–600
Raquel, M. 552
Rasch, G. 463, 482
Rasch partial credit analysis 434
raters/rating: accents and 328, 332, 568; approaches 214–215; bias and 8, 11, 242, 327–328, 331, 342, 471, 474, 502, 568, 571; cognition/decision-making processes 332, 503–504, 506; critical issues 324–327; effects 496, 502; expertise 324–325, 327–328; further reading 333; future directions 332–333; historical perspectives 323–324; hyper-homogeneity and 502; impact of 322; linguistic sophistication of 328–329; methodology of research 329–330; MFRM and severity of 327; overview 8, 322–323; practice recommendations 330–332; professional background 329; reliability 479, 490–491; research 327–329; rubrics 295, 315, 490; scales 496, 501–502; scoring experience 329; severity of 327–328; training 242, 322, 325–328; validity 325–326; variability 8, 326–327; *see also* examiners
rating scales 496, 501–502
rationalist-empiricist approach 488
Razı, S. 329
read-aloud accommodations 309
reader-purpose perspective 255
Rea-Dickins, P. 130
reading 254–258, 263–264
reading abilities, measuring higher-level 403–406
reading assessment: authenticity and 259; central processing core and 256; challenges 263; cloze format 257; cognitive processing model and 256; componential approach to 255; computerized 259, 263; construct-irrelevant variance and 262; construct representation and 261–262; critical issues 257–258; CVRi and 264; descriptive statistics and 260; differential item functioning studies and 260; eye-tracking methodology 261; first language 256; further reading 264–265; future directions 263–264; gap-fill tasks 257; historical perspectives 254–257; information in the text to target and 262; integrated 263–264; introspective questionnaires 261; item characteristics and 258; knowledge base and 256; matching format 257; metacognitive activity and 256; methodology of research 259–261; multiple-choice format 257, 462; overview 7, 254; practice recommendations 261–263; process 257; qualitative analysis and 260; research 258–259; second language 254–256, 257, 264; sentence-completion tasks 257; short-answer questions 257; stimulated recalls and 261; technology and 259, 263; test-taking strategies 385–386, 388, 391; text characteristics and 258; TOEFL 2000 Reading Framework and 255; true-false-not given format 257; validity 380–382; verbal protocols and 260–261; written-text sourcing and 262
reading-listening-writing (RLW) task 408
reading test-taking strategies 385–386, 388, 391
reading-to-learn prototype study 406
reading-to-write task 357–358
Read, J. 10, 92, 434, 436, 438, 551
real-world spoken language 227–228
rebuttal data 290
reciprocal scaffolding 196–197
Reed, D. J. 430

reference corpora 546
relevance 418
reliability: classroom-based assessment and 491; conceptualizations 479–483; for constructed-response items 481–483, 485; defining 479; error and 477–478, 483, *483*; estimates 452–455; factors influencing test 483–486, *483*; fairness and 487–490; further reading 491–492; future directions 490–491; internal consistency 479–481; measurement and 477–478; measurement error and 478; operationalizations of 479–483; overview 11, 477; parallel-form 479–480; performance test scoring 505; physiological measurements and 478; in psychometrics 477; raters/rating 479, 490–491; score variance and 479; scoring 242; for selected-response items 479–481, 484–485; split-half 453, 479, 481; standard error of measurement and 481; test-retest 479–480; types of 479; validity and 477, 483, 487–490; *see also* dependability
repair in trouble talk 69
replicability of reading passages 406
research: accommodations 311–316; automated scoring of performance tests, validity of 513, 516–518; bias in 457; classical test theory 450–451; classroom-based assessment 125–127; computer-based testing 535–539; confirmatory 3, 282, 415; contextualized 161; corpus linguistics 545, 553–554; diagnostic assessment 199; dynamic assessment 174–182; empirical 139, 242, 348, 359; ethics/ethical codes 570–572; evidence-centered design 297–298; fairness 572–573; IRT 471–472; item writers/writing 349–351; longitudinal, lack of 622; meta-analysis of 258; performance test scoring 503–504; pre-operational testing 420–422, 426; raters/rating 327–329; reading assessment 258–259; second language acquisition 25, 53, 210, 247, 540; test validity 49; transdisciplinary 129–130; validity 26–27; washback 139–145; writing assessment 237–244; writing integrated tasks 359–363; *see also* methodology of research
re-sequencing words tasks 410
reusability 299–300
Revesz, A. 230
Ricento, T. 98
Richards, N. M. 615
Richardson, M. W. 448
Riconscente, M. M. 298
Rind, I. A. 141
RITCME (recruitment, induction, training, co-ordination, monitoring, and evaluation) framework 350
Roach, J. O. 498–499
Roever, C. 71, 74–75, 92, 343
Römer, U. 553
Roosevelt, Franklin Delano 98, 101
Rosenbusch, M. H. 75
Ross, S. J. 3
Roy, E. L. 514
rubrics: assessment and 302; criterion-references 125, 128, 448–449; developing new 411–412; interactive 128; interpretation of 8; rating 295, 315, 490; SAT essay 274–275, **275**, **276**; scoring 33, 39, 238, 241–242, 245, 322, 325, 330, 341, 366, 400–401, 409, 411–412, 495–497, 552, 555–556
Ruch, G. M. 271
Rupp, A. 472
Rupp, A. A. 261
Ryan, E. 348
Ryan, K. 490

Sacks, H. 68
Şahan, Ö. 329
Saito, K. 329
Salisbury, K. 347
sample size 473
sampling 40–41, 400–401, 424
sandwich format 172–173
Sarig, G. 262–263
SAT Essay Test 274–275, **275**, **276**
Sato, T. 143
Saudi Arabian EFL students 388
Saussure, F. de 65
Saville, N. 13, 605, 623
Sawaki, Y. 12, 362, 626
scaffolding: assessment for learning and 158; diagnostic assessment and 198; in mediated diagnostic assessment interventions 196–197; prior knowledge 367; prototyping new item types and 410–411; second language and 164; writing assessment and 244, **245**; written corrective feedback and 244; young language learners and 534
scales, rating 496, 501–502
Scarcella, R. 161
scatterplots 53, *54*, *55*, 485, *485*
Schedl, M. 403, 437
Schegloff, E. A. 68
schematic table completion 403, *404*
Schleppegrell, M. J. 68, 161
Schlesinger, A. 97
Schmidgall, J. E. 533
Schmitt, N. 439–440
Schneider, E. 174
School Leaving Certificate (SCL) examination 140
Schoonen, R. 438, 440, 442
Schroeders, U. 536
scientific data management/stewardship 615
score tests 25
score variance 479, 489

scoring: analytic 496–497; approaches, advantages/limitations of 500; automated 242; automated essay 242; criteria 329–330, 495–497, 505; full-test 497; holistic 496; IELTS 587–590; instruments, choosing contextually appropriate 500; interpretation of 502–503; model 505; observed scores 447; paper-based test 402, 405; reliability 242; rubrics 33, 39, 238, 241–242, 245, 322, 325, 330, 341, 366, 400–401, 409, 411–412, 495–497, 552, 555–556; second language acquisition and 502; TOEFL 587–590; true scores 448; true score variance 448–449; validity 499; writing assessment 241–242; writing integrated tasks 366–367; *see also* performance test scoring; performance test validity and automated scoring
Scriven, M. 119
Secondary Education Certificate Examination (SECE) 144
secondary test users 603
second language (L2): accents and 228, 484; computer-based testing and research on 540; dynamic assessment and 174–177; listening assessment 223–231; mediated experiences 175–177; reading assessment 254–258, 264; scaffolding and 164; writing assessment and 236, 240, 246; *see also* English as a Second Language (ESL)
second language acquisition (SLA): individual differences in 53; long-term versus short-term interventions and 56; prediction of 55; prototyping of new items and 409; research 25, 53, 210, 247, 540; scoring and 502; theory 25, 121, 583
selected response (SR) 407
selected-response items 479–481, 484–485
selected response task types 192
self-observation 389
self-regulated learning (SRL) 121
self-report 389, 506
self-revelation 389
self-scaffolding 197
sentence-completion tasks 257
sentence/paragraph re-ordering 259
serial piloting 433
Seyferth, S. 503
Shanahan, T. 359
Shaw, P. 346
Shaw, S. 328
Shaw, S. D. 18
Shewhart, W. A. 600
Shih, C. 143
Shin, D. 8–9, 350, 627, 629
Shindoll, R. R. 290
Shohamy, E. 72–73, 76, 153, 345, 349, 567–568, 572–573
short-answer question format 257, 262
short-essay question format 295
showcase portfolios 129
Sinclair, J. 6, 623, 626
Sireci, S. 17, 417
Sitajalabhorn, W. 357
skill area, test-taking strategies by 387–388
Slomp, D. H. 247
social consequences of language tests 71–76
social constructs 65–66, 76
social dimensions of language testing: discursive practice and, language testing as 66–71; further reading 77–78; future directions 76–77; overview 3; social consequences of language testing and 71–77; social constructs in language testing and 65–66, 76; social life of language and 64–65; vignette 63; *see also* design of language tests for specific social uses
social integration 97, 572
social and interactional trait 210
socio-cognitive framework 282
socio-political nature of language testing 628–629
socio-political thinking 345–346, 351
Solano-Flores, G. 314
Sollenberger, H. E. 581
Spaan, M. 273, 567–568
speaking assessment: advancements in 209; computer-based testing and 537–539; conversation analysis and 215–216; critical issues 211–215; elicited imitation tasks and 211–212; further reading 218; future directions 217–218; historical perspectives 209–211; independent tasks and 212; integrated tasks and 212–213; interactive tasks and 213; language function analysis and 216; marking models and 214–215; measured constructs and 211–214; methodology of research 215–217; micro-linguistic analysis and 216–217; overview 6, 217; questions to research 217–218; rating approaches and 214–215; test delivery modes and 213–214; test-taking strategies 386–387; in United Kingdom 210–211; in United States 210; validity 382–383; video-conferencing situations for 538
speaking descriptors 582, **583**
speaking task types 211
speaking test-taking strategies 386–387
Spearman-Brown prophecy formula 448, 453
Spearman, C. 447–448
spec-driven test development
specialized corpora 546
specification-driven testing 283–284, *283*
specifications *see* test specifications
specificity level, determining 192
speech-to-text technology 351
split-half reliability 453, 479, 481
spoken dialogue systems (SDS) 214, 538

spoken language corpora 548
Spolsky, B. 187, 194–195, 198, 272, 448, 488
Standard Corpus and Present-Day Edited American English 546
standard error of measurement (SEM) 455–456, 481
Standardization Agreement 6001 (STANAG 6001) 582–583, 591n1
standardized assessments 39–40
standardized diagnostic tests 190–192
standard operating procedure (SOP) 607
standards: assessment 39–40; for automated scoring of performance tests 522; curriculum-based 586–587; defining 578–579; diverse needs of children and, reconciling 159–162; English, in China 586–587, **588**; formal 579; ICAO proficiency 83, 85–86, 90; ISO 9000 601; language testing, maintaining 73; measurement 17; minimum 616; movements 122–123; private 579–580; quality 600; task-based 586; types of 13; validity 19; *see also* language proficiency measurement standards
Standards for Educational and Psychological Measurement 416
Standards for Educational and Psychological Testing 569
Stanford parser 549, *550*
Staples, S. 554
statistical determinism 372
statistical performance 423
statistics, descriptive 260
Stemler, S. E. 482
stem-option matching format 375
Stimpson, G. 74
stimulated recalls 261
student model 294–295
subjective approaches 48, 564
substantive due process 570
substantive fairness 570
sufficiency 418
Sultana, N. 3, 5, 124, 140–141, 142, 144–145, 147–148, 628
summary task 403–405, *405*
summative assessment 119, 120
Sun, Y. 26, 142–143, 145
support processes 605
Supreme Court 99, 101
survey techniques 230, 329
sustainability of reading passages 406
Suvorov, R. 231
Swain, M. 360
symptomatic assessment 171–172
synchronous computer-mediated communication (SCMC) 246
Syntactic Analyzer 552
systematic approach 344–345
Tacitus 71
Taguchi, T. 422
Taiwan 143
Taiwanese EFL students 387, 409
"take-at-home" tests 623
TALE project 277, 280
Talk2Me Jr. 197
Tan, M. 140
Tannen, D. 227
target domains 37–38, 40–41
target-language (TL) 255, 373
target language use (TLU) domains 213, 225, 239, 276, 292, 344, 535, 537, 548
task-based language teaching (TBLT) 122, 125, 358
task-based standards 586
task bias 471
task design: assumptions about 372–373; innovative 532–533; relationship between language proficiency and test-taking strategy use 383–387; skill area test-taking strategies 379–383; test preparation and 388–389; test-taking strategies and 378; validity of measures 379–383; washback and 387–388
task model 295
task types: alternative, measuring defined construct 408–411; appropriateness of 92; cognitive processing/strategy across different, varying 198; computer-based testing and 611–612; considering 364–365; elicited imitation 211–212; GEPTI-S 387; independent 212, 360; innovative 237, 367, 408, 532–533; integrated 9, 212–213, 263–264, 360, 367, 379–380; interactive 213–214; new reading 256; prototyping new 399, 403–407, *405*, *406*, 409, 412–413; range of 6; reading-to-learn 406; re-sequencing words 410; restricting range of 489; score variances with 489; scoring rubrics 412; selected response 192; sentence/paragraph re-ordering 259; speaking 211–214; strategy use and performance and 383; summary 403–405, *405*; tension/incompatibility between 519, 532; TOEFL iBT 532; TOEFL Junior 412; *see also specific type*
Taylor, F. W. 600
Taylor, L. 331, 625
teachers: development 181; education and dynamic assessment 180–181; training 141, 162, 180–182, 346, 450; understanding of curricula 279; young learners' assessment and 164
Teachers of English to Speakers of Other Languages (TESOL) 153, 488
team development and quality management 604–605, 612
Teasdale, A. 87
technical approach to validation 282

technical-formalist approach 572–573
Technical Recommendations and Standards 17
technical test development strategies 127
technologically mediated classroom-based assessment 126–127
technology: artificial intelligence assistants 351; automated writing evaluation 244, 524; challenges in testing field 624–625; classroom-based assessment and 128–129; computer-based testing and 530; corpus linguistics and 556; in diagnostic assessment 197–199; electro-encephalography 506; eye-tracking 198, 231, 261, 332, 506, 623; item writers/writing and 347–348, 351; language development and 126–127; language tests/testing and, increasing role in 623–625; listening assessment and 231; overview 11–12; rater cognition and 506; reading assessment and 259, 263; speech-to-text 351; young learners' assessment and 154–155, 163–164
telemetrics 127
Templin, J. 194
Teo, A. 534
test bias 416, 569–570, 572
test constructs 273–275
Test Context Framework 570–573
test design/administration 7–8; *see also* quality management in test production and administration
test-deviousness strategies 9, 374–375, 379
test-driven teaching 125
Test of English as a Foreign Language (TOEFL): automated scoring 554; certification perspective and 587–590; computer familiarity and 535; diagnostic assessment and 190; interpretive argument and 22; Korean 142–143; pre-operational testing and 418; scoring 587–590; speaking tasks 360, 382; test-deviousness strategies and 379; vocabulary assessment 431, 436; washback and 138; writing assessment and 241
Test of English for International Communication (TOEIC) 73, 298, **299**, 409
test fairness 567
test information function (TIF) plots 467
testing agency 599
testing populations, new 410–411
test-management strategies 374, 381, 391
test material procedures, assembling 608–609
test preparation 388–389
test-retest reliability 453, 479–480
tests *see* language tests/testing; *specific name*
test scores 19, **20**; *see also* scoring
test sensitivity review 568–569
test specifications: for classroom-based language assessment 277–280; educational measurement and 271; effect-driven testing and 282–284, *283*; evidence-centred design and 271, 290; further reading 285–286; future directions 284–285; for high-stakes language testing 272–277; internet and 285; item writers/writing and 348; language tests/testing and 284–285; literature on 271–272; multiple referencing and 285; origin of *specification* and 271; overview 7; pragmatic realism and 274, 281, 282–284; specification-driven tests and 283–284, *283*; test syllabus versus 281; for validity/validation of tests 281–284
test syllabus 281
test-taking strategies: automated scoring of performance tests and, adapting 523; caveat 375–376; in cloze format 376–377, 388, 390; clustering of 389–390; communication strategies and 383; critical issues 378; dynamic assessment and 392; further reading 393; future directions 392; for gap-filling test 386; for high-stakes test 383–384; historical perspectives 376–377; instruction 391; investigating 389–390; language learner strategies and 373–374; language proficiency and 383–387; listening 384–385, 387, 391; main idea and, understanding 377; methodology of research 389–390; in multiple-choice format 375–376, 388, 389–391; overview 9, 372; practice recommendations 390–391; pragmatic strategies and 376; reading 385–386, 388, 391; by skill area 387–389; speaking 386–387; statistical determinism and 372; stem-option matching format 375; task design and 378; test-deviousness strategies and 9, 374–375, 379; test-management strategies and 374, 381; test preparation and 388–389; validity of measures and 379–383; verbal reports and 380, 383, 389; washback and 388; writing 384–386
test usefulness 20–21, 276, 599
test validity 49; *see also* validity
test-wiseness 9, 374–375, 379
text characteristics 258
textmapping 263
text-responsible tasks 363–364
text selection for writing integrated tasks 365
textual authenticity 347, 350, 626
textual malfunction 67–68
themes in language testing 13, 622–630
theory-driven approach 156
Thewissen, J. 549
13th Amendment 100
Thompson, F. 103
Thompson, S. A. 68
Thomson, R. I. 329
Thorndike, E. L. 498
Thorne, S. 174
3D virtual learning environments (3DVLEs) 127

Thurston, L. L. 463
time, extended 312
"tiny tricks" 385–386
TOEFL 2000 Reading Framework 255, 401, 403, 408
TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) 547
TOEFL ITP4 12, 413n4
TOEFL Junior speaking test 412, 413n5
TOEFL paper-based test (PBT) 402, 413n2
TOEIC Bridge writing test 409–410
TOEIC speaking test 410, 413n3
TOEL iBT 26, 49, 211, 224, 297–298, **299**, 352, 380–382, 401, 408–409, 412, 532–533
top-down approach 140, 156
top-down processing 255
Toprak, T. E. 193
Torrance, H. 153–154
Total Quality Management (TQM) 600
Toulmin, S. E. 3, 50, 289, 292, 417
traits 3, 38, 210, 456–457
transdisciplinary research 129–130
transmission model 121
transnational immigrants 96
Trevarthen, C. 67
trialling stage 417, 430
Tribble, C. 555
Trim, J. 585
Trinity's Integrated Skills in English (ISE) examinations 211
trouble talk 69
true-false-not given format 257, 262, 345
true scores 448
true score variance 448–449
Truman, Harry S. 101
Tsagari, D. 138, 143, 145
t-statistic 551
Tunks, J. 346
Tunmer, W. E. 255
turn-constructional unit (TCU) 68–69
Turner, C. 124
Turner, C. E. 4, 127–128, 140, 277–279, 497
two- and three-parameter Rasch models 467–469, *469*

UCREL Semantic Analysis System 548
Umashankar, S. 139, 143, 145
underrepresentation 161, 570
unitary competence hypothesis (UCH) 273
unitary concept of validity 610
United Kingdom: British Academic Spoken English Corpus and 546; British Standards Institute in 600–601; speaking assessment in 210–211; University of Cambridge Local Examinations Syndicate in 210, 449, 498
United Kingdom ESL placement test in 531
United States: Bureau of Immigration and Naturalization 102; Census Bureau 99; Chinese Exclusion Act and 100–101; Civil Service Commission 581; Constitution 100; Cuban and Central American immigrants and 101; Declaration of Independence 100; ESSA and 187; Foreign Service Institute 210, 323, 498–499, 581–582; Immigration Act and 101; immigration applications, naturalizations, and denials 109, **109**; immigration policies 99–102, 113n1; Immigration Reform and Control Act and 102; INA and 101–102, 111; Indian Citizenship Act and 101; Japanese Americans' incarceration and 101; Jewish immigrants and 101; language policy 98–99; Literacy Test Bill and 103; Mariel boatlift and 101; monolingual language policy in 98–99; multiple-choice format in, switch to 236–237; Naturalization Act of 1790 and 100; NCLB and 74–75; speaking assessment in 210; Supreme Court 99, 101; 13th Amendment and 100; writing assessment in 236–237
universally designed accommodations 8, 307, 312
universe of generalization for language proficiency 39–41
universe score 39
University of Cambridge Local Examinations Syndicate (UCLES) 210, 498, 588
University Word List 434
Upshur, J. 497
usability testing 400
USAS tagger 548, *549*
US Citizenship and Immigration Services (USCIS) 106
usefulness framework 20–21, 599
user-acceptance testing (UAT) 613
US Naturalization Test (USNT): accents and 104; access and 107–108; accommodations and 107–108; administration of 105; arguments against 112–113; benefits of, questioning 105; bias and 107; consequences 108–109, **109**; content 104–105; development of 104; fairness principles and 105–109; format 104–105; ideological motives and 628; justice principles and 105, 110–111; justification 110; learning opportunities and 106; literacy test 102–103, *103*, 114n8; meaningfulness and 106–107; reform suggestions 111–112
Utiyama, M. 551

Valette, R. A. 18
validation: appraisal stage of 35; development stage of 34–35; priorities, determining 521–522; technical approach to 282; of test and corpus linguistics 554; tests specification for 281–284

validity: of accommodations 309–311; analyses 456–457; argument 32–36, 571; of automated scoring, frameworks for using 525–526; coefficients 49–50; co-existing perspectives on 23–26; cognitive 372–373; comparable 570; concurrent 456; consequential 565; consequential bases for 573; construct 456; construct-related evidence for 573; content 456; criterion-related 310–311, 456; critical issues 22–26; defining 17–20; evidence for 18–20, **20**; fairness and 487–490, 569–571; further reading 28; future directions 27; historical perspectives 18–22; inferences 610–611; of integrated reading/writing tasks 379–380; Kane's conception of 21–22; language constructs 24; language development and 25; language testing 18, 25–26, 48; measurements 17; of measures 379–383; Messick's conception of 18–19, **19**; objective approach to testing 48; overview 2–3; predictive 51–53, 57, 456; quality management in test production and administration and 597, 610–611; questions, asking relevant 518–521; raters/ rating 325–326; reading assessment 380–382; reliability and 477, 483, 487–490; research 26–27; scoring 499; social consequences of 573; speaking assessment 382–383; standards 19; studies 49; subjective approach to testing 48; technical approach to, test specifications for 282; test specifications for 281–284; test usefulness and 20–21; theory 44–45; types of 18, 20; unitary concept of 610; value implications of 573; washback and 138–139, 284; *see also* argument-based validity; automated scoring of performance tests, validity of
values of language testing 25–26
van Compernolle, R. A. 178
Van Moere, A. 87, 89
variability of rater 8, 326–327
variables, latent 462
variance, item 449
Veda 64
verbal protocols 192–193, 260–261, 329
verbal reports 230–231, 379–380, 383, 389
verbal resources 67
Verhallen, M. 438, 440, 442
Versant Aviation English Test (VAET) 89, 92, 547
vertical discrepancy assessment 189
video-conferencing situations for speaking assessment 538
virtual reality in language testing 537
vocabulary assessment 431, 436; *see also* vocabulary tests, piloting
Vocabulary Knowledge Scale 436
vocabulary and language 430–431
Vocabulary Size Test 431
vocabulary tests, piloting: collocations and 442; computer-based testing and 442; conventional test items 431–433; diagnostic language assessment and 442; discrete assessments and 431; embedded assessments and 431; further reading 442–443; future directions 442; historical perspectives 431; overview 10, 430; pilot testing and 430; scope of 430–431; *see also* word associates
Vogt, K. 147, 277
voluntary immigrants 96–97
Voss, E. 19, 23, 536, 552
Vygotsky, L. S. 158, 171–172, 392
Vyn, R. 143

Wagner, E. 6, 226–227, 229, 231, 537, 626
Walker, M. 417, 419
Wall, D. 136–137, 138, 143, 145–147, 229
Walters, F. S. 3, 12
Wang, H. 27
Wang, L. 420
Wang, Y. 156
warrant 289
washback: alignment and 139–141; complexities 138–139; contextual factors 141–143; critical issues 138–139; defining 136; design of research 145–147; diagnostic assessment and 194–195; evolution of 139; fairness and 148; frameworks, emerging conceptual/theoretical 143–145; further reading 148; future directions 147–148; historical perspectives 137–138; IELTS and 138; language testing and 137–138; learning and 147–148; listening assessment and 229; literacy assessment and 147; methodology of research 145–147; misalignment and 139–141; negative 5; overview 5, 136–137; positive 5; research 139–145; task design and 387–388; test-taking strategies and 388; test validation and 284; TOEFL and 137, 138; validity and 138–139, 284
Washington, George 99–100
Watanabe, T. 362
Webb, N. L. 144
Weigle, S. 553
Weigle, S. C. 518, 547
Weir, C. J. 144, 256–257, 264, 282, 372–373, 499, 534
Wendler, C. 417, 419
Wesolowski, B. C. 504
White, E. M. 324, 332, 500
wider processes 605
Wiggins, G. P. 280
Wigglesworth, G. 323, 328
Wild, C. L. 599–600
Wilhelm, O. 536
Wiliam, D. 122

Willingham, W. W. 570–571
Willner, L. S. 307
Wilson, A. 547
Wilson, Woodrow 98, 103
Winke, P. 160, 490
Wolfe, E. W. 325, 328–329, 333, 536
Wolf, M. K. 156
word associates, piloting: design features, further research on 439–440; Dutch learners, adapting format for 438–439, *438*; French learners in the Netherlands 437–438; initial studies, devising with 433–436; North American studies 436; review of, for summarizing piloting of vocabulary tests 440–442
work instructions and quality management 607
World-Class Instructional Design and Assessment (WIDA) 187
Wright, B. 463
Wright, W. 98
writing assessment: authenticity and 359; automated 244; CEFR and 243; in China's Chou period 236; collaborative writing and 246–247; considerations 237, **238**; constructs 239, 241; critical issues 237–244; development of writing 247–248, *247*; evidence 238–241, **238**; further reading 248–249; future directions 244–248; historical perspectives 236–237; IELTS and 241; interpretation 243; over time 247–248; overview 7, 236, 248; process 237, *237*; research 237–244; scaffolding and 244, **245**; scoring 241–242; second language and 236, 240, 246; test-taking strategies 384–386; TOEFL and 241; in United States 236–237; use of 243–244; written corrective feedback and 241, 244–246, **245**, 625
writing integrated tasks: challenges 357; characteristics of tasks 360–361; constructs 362–363; content-responsible tasks 363–364; critical issues 363–367; defining *construct* and 363; describing integrated task and 357; examples 357–358, 364–365; further reading 367–368; future directions 367; historical perspectives 358–359; independent skills tasks versus 359–360; instructional composition and 365–366; needs analysis and 363; performance features 361–362; practice recommendations 363–367; research 359–363; scoring rubrics for 366–367; task type and, considering 363–365; text-responsible tasks 363–364; text selections and 365; TOEFL 360
writing items/tasks *see* item writers/writing
writing test-taking strategies 384–386
written corrective feedback (WCF) 241, 244–246, **245**, 625
written texts corpora 548
written-text sourcing 262

xenophobia 99, 113n1
Xi, X. 11, 517, 555–556, 569–575, 623

Yamashita, J. 258
Yanagawa, K. 227
Yancey, K. B. 237
Yang, H. C. 362–363
Yang, Y. 515–517
Yan, M. 535–536
Yan, X. 11, 49, 211–212, 625
Yin, C. 7
young language learners (YLLs) 153, 158–164, 534
young learners' assessment: age-appropriate, development and practice 155–157; critical issues 155–159; defining 153; diverse needs of children, reconciling with standards 159–162; further reading 164–165; future directions 163–164; globalization and 154; historical perspectives 153–155; interconnectedness of teaching, learning, and assessment 157–159; *Language Assessment Quarterly* and 154; *Language Teaching for Young Learners* and 154; *Language Testing* and 154; overview 5; practice recommendations 162–163; teachers and 164; technology and 154–155, 163–164
Young, R. F. 3, 67, 70
Youn, S. J. 376
Yu, G. 346, 536

Zenisky, A. L. 417
Zhang, D. 440
Zhang, J. 179, 325, 504
zone of proximal development (ZPD) 158, 171–172, 176