Chapter 5

Test Construction

- Determining Function and Form of the Test
- Planning
- Preparing Items
- Reviewing
- Pretesting
- Validation
- Extra Points to Remember

Scanned by CamScanner

Test Construction

To construct a test the following steps should be taken. First, the function and the form of the test should be decided upon. Second, the content of the items should be specified. This step is also called planning. Third, the items should be prepared in accordance with the specified content. Fourth, the items should be reviewed. Fifth, the items should be pretested in order to determine their statistical characteristics. And finally, the test should be validated. Thus the steps in developing a test can be listed as:

1) Determining function and form of the test
2) Planning
3) Preparing items
4) Reviewing items
5) Pretesting items
6) Validation

In this chapter we shall briefly describe each step in the process of test construction as applied to the testing of English as a second language.

1. DETERMINING FUNCTION AND FORM OF THE TEST

In order to determine the function and form of a test, three factors should be taken into account:

- **Characteristics of the examinee:** The nature of the population is important. If the testees are a group of youngsters, for example, test items in pictorial mode would be more appropriate than a test with written modality. If the same mother-tongue is shared by all the testees, the task of sampling is made slightly easier even though they may have attended different schools or followed different course. They will all experience problems of a similar nature as a result of the interference of their first-language habits and thus the test could include items based on the findings of contrastive analysis on the elements of the source and the target language
- **Specific purpose of the test:** According to the test function, the content of the test would differ. For example, the content of a proficiency test differs from that of an achievement test.
- **Scope of the test:** Whether a test is to be used within the scope of a classroom, a school, or a country influences the structure of the test. As the scope of the test widens, the degree of care to be taken along with the amount of time and energy to be spent increases because the decision to be made would influence a larger population.

2. PLANNING[1]

It is important for the tester to decide on the area of knowledge to be measured. The content of the test depends on both function and form. For example, the content of a placement test should logically differ from that of an aptitude test. Moreover, a comprehension test in multiple-choice format allows the examiner to include as many items as he feels necessary, whereas in a production type test, limitations of item and space force the test developer to limit the test to a manageable number of items.

In order to determine the content of a test:

First, instructional objectives should be examined. That is, the course content should be outlined to include a list of major structural points covered during the instruction.
Second, major topics should be divided into their specific components;
Third, *table of specification* should be prepared.

The main purpose of table of specification is to assure the test developer that the test includes a representative sample of the materials covered in a particular course. The following table of specification is prepared for a grammar course:

Instructional objectives/ Content	Number of items
Reported speech	3
Subjunctive	2
Dangling structure	5

As can be seen table of specification specifies what is to be tested (in this case reported speech, subjunctive and dangling structure); the aspect of achievement to be tested; and the number of items (in this case 10 items).

3. PREPARING ITEMS

This is to write items. In the process of item preparation, even the most experienced teacher is apt to make defective items. When constructing MC items[2], test constructor should bear in mind the following guidelines.

a) In constructing multiple-choice items, only one feature/skill at a time should be tested. Such an item is called **pure item** such as listening to a text and then subsequently hearing a series of true/false questions to be marked on the answer sheet. Pure items are contrasted with **Hybrid item** – one in which the examinee tests more than one skill such as dictation which requires testee to *listen* to a text and then *write* it. As an example, the following test item, tests both word order and sequence of tenses:

I never knew where .....
1) had the boys gone
2) have the boys gone
3) the boys have gone
4) the boys had gone

1- Planning or "specifying the content of the test"
2- For guidelines on constracting T/F and matching items refer to FAJAB, chapter 5.

Scanned by CamScanner

b) All MC items should be at a level appropriate to the testee's proficiency level. The content should be at a lower level than the actual problem which the item is testing. For example, a grammar test item should not contain other grammatical features as difficult as the area being tested; or as another example, vocabulary item should not contain more difficult semantic features in the stem than the area being tested.

c) The stem should be quite clear and state the point to be tested unambiguously. The following item is thus deficient:

He is .....
1) one of the students passing the test
2) one of the students failing the test
3) the man who left the testing session early
4) the man who did participate in the test

d) The stem should not contain extraneous information or irrelevant clues, thereby confusing the problem being tested. Unless students understand the problem being tested, there is no way of knowing whether or not they could have handled the problem correctly. For example, in the following example clause after 'because' could be eliminated without making the item impossible to answer.

Children were not ..... to watch the violent part of the movie because in that part two criminals attack a citizen in the street and kill her with a gun.
1) displayed
2) allowed
3) perceived
4) proven

e) The stem should include as much of the item as possible. Any word or phrase that is shared by all alternatives should be placed in the stem. The following item is deficient because *who* should have been included in the stem.

The person ..... is called an author.
1) who writes a book
2) who prints a book
3) who reviews a book
4) who sells a book

f) Negative statements should be avoided because they are likely to be ignored by the examinees.

g) All of the alternatives must be grammatically correct by themselves and consistent with the stem. Choice three in the following item is impossible and thus should be eliminated.

Last year, incoming students ..... on the first day of school.
1) enrolled
2) will enroll
3) will enrolled
4) should enroll

It is worth mentioning that using wrong distractors would expose examinees to wrong forms of the language which might negatively influence the student's language learning process.

h) Every item should have one correct or clearly best answer. In the following item choices 2 and 3 could be taken as the answers.

When he called, I ..... the house.
1) left
2) had left
3) was leaving
4) have left

i) All distractors should be plausible. That is, distractors which do not logically belong to the point being tested will be discarded by the testees. In the following item a verb is to be tested and therefore choice four is unacceptable.

To call on someone means .....
1) to visit
2) to talk
3) to telephone
4) a curious person

j) All distractors should be of similar length and level of difficulty. A relatively long alternative tends to be the correct response.

k) Using 'all of the above' or 'none of the above' as an alternative is not recommended.

l) Correct response should be distributed approximately equally but randomly among the alternatives.

m) The stem should not provide any grammatical clue which might help the examinee find the correct response without understanding the item. In the following example, the article 'an' leads the testee to the correct response

He picked an ..... off the tree and gave it his guest.
1) apple
2) tangerine
3) banana
4) peach

n) The stem should not start with a blank. This recommendation originates from the concept of meaningful learning. According to the cognitive-code learning theory, language processes start with known information and move towards unknown information. Starting a stem with a blank means that the testees should move from unknown to known information. Thus the flow of information is in the direction opposite to the normal flow of information.

4. REVIEWING

It is highly recommended that the produced test be reviewed by an outsider. Through the **reviewing** stage, problems unnoticed by the test developers are likely to be observed by the reviewers. Of course, the reviewer's comments would be subjective and are, thus, not sufficient for the development of a reasonable test. For a test to be scientifically defensible, it should be examined objectively. Such a scrutiny will be possible though the next stage of test construction.

5. PRETESTING

Reviewing is important because it indirectly helps the pretesting stage. That is, improved items would make pretesting more fruitful. **Pretesting** is defined as administering the newly developed test to a group of examinees with characteristics similar to those of the target group. For example, if a test is designed for children, it should be pretested with a group of children. Otherwise, the goal of pretesting will not be achieved. Through pretesting the tester collects the numerical data, based on which the efficiency and/or shortcoming of the item is going to be checked.

The purpose of pretesting is twofold. The first purpose is to determine, objectively, the characteristics of the individual items (i.e. item analysis). The second purpose of pretesting is to

Purpose of pretesting → to determine characteristics of item (Item anal
→ validation)

Scanned by CamScanner

determine the characteristics of the items altogether (i.e. validation). First we will investigate item analysis procedures suitable for NRTs and then CRTs. The effectiveness of NRT items is determined through item facility (IF), item discrimination (ID), and choice distribution (CD); the effectiveness of CRT items is determined through different index (DI) and *B*-index.

NORM-REFERENCED ITEM ANALYSIS

N.1. Item facility (IF)

Item facility refers to the easiness of an item. Technically, it is defined as the proportion of correct responses for every single item. Proportion means that all correct responses should be divided by the total number of responses. This idea can be illustrated by the following formula:

$$IF = \frac{\sum C}{N}$$

IF = item facility
$\sum C$ = sum of the correct responses
N = total number of responses

Note: Item facility equation assumes that items left blank are incorrect answers.

Table below shows the hypothetical results of a grammar test with 10 items. The actual responses are recorded with a 1 for each correct answer and 0 for a wrong answer. As can be seen from the 'Total' column, Haruka and Rie obtained the highest scores i.e. 9 while Kana received the lowest score i.e. 1.

	Items										
Subjects	1	2	3	4	5	6	7	8	9	10	Total
Haruka	1	1	1	1	1	1	1	1	1	0	9
Rie	1	1	1	1	1	1	1	1	1	0	9
Arisa	0	0	1	1	1	1	1	1	1	0	7
Saki	0	1	0	1	1	1	0	1	1	0	6
Natsumi	1	1	1	1	1	1	0	0	0	0	6
Tomomi	0	1	0	0	1	0	0	1	0	0	3
Momo	0	0	0	0	1	0	0	1	1	0	3
Yuuka	0	0	0	0	1	1	0	0	0	0	2
Kana	0	0	0	0	1	0	0	0	0	0	1

Let's calculate IF for items 2, 5, 7 and 10.

In item 2 out of nine students, five students provided the correct answer:

$$IF_2 = \frac{\sum C}{N} = \frac{5}{9} = 0.55$$

In item 5 out of nine students, nine students provided the correct answer:

$$IF_5 = \frac{\sum C}{N} = \frac{9}{9} = 1$$

In item 7 out of nine students, three students provided the correct answer:

$$IF_7 = \frac{\sum C}{N} = \frac{3}{9} = 0.33$$

In item 10 out of nine students, none of the students provided the correct answer:

$$IF_{10} = \frac{\sum C}{N} = \frac{0}{9} = 0$$

It can be understood from the examples above that the maximum item facility, when all examinees get an item correctly, equals 1. By the same token, the most difficult item is the one to which no one gives a correct response, i.e. the item facility is zero. Therefore the range of IF index is:

$$0 \leq IF \leq 1$$

Items with facility indexes beyond 0.63 are too easy and items with facility indexes below 0.37 are too difficult. Therefore, the acceptable range of IF index is:

$$0.37 \leq IF \leq 0.63$$

The ideal index for IF is:

$$IF_{ideal} = 0.5$$

➲ **Example: In a test 20 testees answered an item correctly. If 50 students took the exam, what would be item facility?**

1) 0.3 2) 0.35 3) 0.4 4) 0.45

$$IF = \frac{\sum C}{N} = \frac{20}{50} = 0.4$$

➲ **Example: In a test which consists of 50 items, the first item was answered correctly by 60 students. If 100 students took the exam, item facility would be ---------**

1) 0.35 2) 0.45 3) 0.55 4) 0.60

$$IF = \frac{\sum C}{N} = \frac{60}{100} = 0.6$$

Note: Item facility is also called *facility value* and is shown shortly as FV.

Note: By moving the decimal point two places to the right, teachers can interpret this value as the percentage of correct answers for a given item. For example, a facility value of 0.35 could be interpreted as 35 percent of examinees correctly answered the item.

Note: In calculating item facility, test constructor doesn't need to rank order students based on their scores.

Note: By determining item facility, the test constructor can easily find out item difficulty. Item difficulty can be calculated by using the following formula:

$$\text{Item Difficulty (ID)} = 1 - IF$$

Scanned by CamScanner

➲ **Example: Item ten of a test was answered by 30 students correctly. If there were 80 students in the class, item difficulty is ---------**

1) 0.63 2) 0.57 3) 0.44 4) 0.39

$$IF = \frac{\sum C}{N} = \frac{30}{80} = 0.37$$

$$\text{Item Difficulty} = 1 - IF \rightarrow 1 - 0.37 = 0.63$$

As another solution, to calculated item difficulty, you can put the number of wrong answers (including blank items) in the numerator:

$$\text{Item Difficulty} = \frac{50}{80} = 0.63$$

N.2. Item discrimination (ID)

When a test constructor ranks students according to their scores, we normally expect each item to have been answered correctly more by higher proficiency testees than by lower proficiency testees. In other words, we presuppose those students who performed well on the whole test tend to do well on each item and those who performed badly on the whole test tend to do badly on each item.

	Items										
Subjects	1	2	3	4	5	6	7	8	9	10	Total
Shenan	1	0	1	1	1	1	1	1	1	1	9
Robert	1	0	1	1	1	1	1	0	1	1	8
Millie	1	0	1	1	1	1	1	0	1	0	7
Kimi	1	0	0	1	0	1	0	1	1	1	6
Jeanne	1	0	1	1	0	1	0	0	0	1	5
Corky	0	1	0	0	1	0	0	1	0	1	4
Dean	0	1	0	0	0	0	1	1	1	0	4
Bill	0	1	1	0	1	1	0	0	0	0	4
Randy	0	1	0	0	0	0	0	1	0	0	2
Mitsuko	0	1	0	0	0	0	0	0	0	0	1

Higher Proficiency / Lower Proficiency

Take, for example, item number 2 where item facility is 0.5 which is a perfectly ideal index. However, as can be seen the five top proficiency testees have missed the item and the five low proficiency testees have answered it correctly. This situation renders item number 2 as a low-quality item while by relying solely on IF we should consider it a high-quality item. Therefore it seems that item facility alone wouldn't give us a realistic image of the quality of an item. Besides item facility we need another index called item discrimination.

Item discrimination refers to the extent to which a particular item discriminates more knowledgeable examinees from less knowledgeable ones. The index of discrimination tells us whether those students who performed well in the whole test tended to do well or badly on each item in the test. It is presupposed that the total score on the test is a valid measure of the student's ability (i.e. the good student tends to do well on the test as whole and the poor student badly). On this basis, the score on the whole test is accepted as the criterion measure, and it thus becomes possible to separate the 'good' students from the 'bad' ones in performances on individual items. If 'good' students tend to do well on an item and the 'poor' students badly on the same item, then the item is a good one because it distinguishes the 'good' from the 'bad' in the same way as the total test score. This is the argument underlying the index of discrimination. To compute the item discrimination, the following formula should be used:

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N}$$

C_{high}= the number of correct responses to that particular item given by the examinees in the high group

C_{low} = the number of correct responses to that particular item given by the examinees in the low group

$\frac{1}{2}N$ = the total number of responses divided by 2

Attention: In calculating item discrimination, test constructor, first, needs to rank order students based on their scores.

Let's calculate item discrimination for items 1, 2, 8 and 10.

In item 1, all students in high group gave correct answers while all students in low group gave wrong answers:

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{5 - 0}{\frac{1}{2}(10)} = +1$$

In item 2, all students in high group gave wrong answers while all students in low group gave correct answers:

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{0 - 5}{\frac{1}{2}(10)} = -1$$

In item 8, two students in high group gave correct answers and three students in low group gave correct answers:

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{2 - 3}{\frac{1}{2}(10)} = -0.2$$

In item 10, four students in high group gave correct answers and one student in low group gave correct answer:

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{4 - 1}{\frac{1}{2}(10)} = +0.6$$

Scanned by CamScanner

It can be understood from the examples above that the maximum item discrimination, when all examinees in high group answer an item correctly and all examinees in low group answer wrongly, equals +1. By the same token, the minimum item discrimination, when all examinees in high group answer an item wrongly and all examinees in low group answer it correctly, equals −1. Therefore the range of ID index is:

$$-1 \leq ID \leq +1$$

The acceptable range of item discrimination is:

$$ID \geq +0.4$$

Note: The closer the value of item discrimination to unity, the more ideal the item.

Example: If in a class with 50 students, 20 students in high group and 10 students in the low group answered an item correctly, then ID equals ---------

1) 0.36 2) 0.4 3) 0.47 4) 0.31

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{20 - 10}{\frac{1}{2}(50)} = \frac{10}{25} = +0.4$$

Example: If all the 30 students in the low group answer the first item in a test correctly and half of the students in the high group answer the same item correctly, item discrimination index is ---------

1) − 0.29 2) 0.4 3) − 0.5 4) 0.63

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{15 - 30}{\frac{1}{2}(60)} = -\frac{15}{30} = -0.5$$

The upper and lower groups are sometimes defined as the upper and lower third, or 33 percent. In such type of problems you should divide the total population of testees by three and replace the result in the denominator:

Example: In a group of 120 students, 20 students in the high group and 10 students in the low group answered an item correctly. If each group consists of 33% of testees, then ID is ---------

1) 0.58 2) 0.42 3) 0.33 4) 0.25

$120 \div 3 = 40$

$$ID = \frac{20 - 10}{40} = \frac{10}{40} = +0.25$$

The upper and lower groups are sometimes defined as the upper and lower fourth, or 25 percent. In such type of problems you should divide the total population of testees by four and replace the result in the denominator:

Example: In a group of 100 students, 15 students in the high group and 10 students in the low group answered an item correctly. If each group consists of 25% of testees, ID is-------

1) 0.27 2) 0.3 3) 0.2 4) 0.36

$100 \div 4 = 25$

$$ID = \frac{15 - 10}{25} = \frac{5}{25} = +0.2$$

The upper and lower groups are sometimes defined as the upper and lower fifth, or 20 percent. In such type of problems you should divide the total population of testees by five and replace the result in the denominator:

Example: An item was taken by 50 learners. 7 learners in the high group and 2 learners in the low group produced correct answer. If each group consists of 20% of testees, what is the value of ID index?

1) 0.4 2) 0.2 3) 0.3 4) 0.5

$50 \div 5 = 10$

$$ID = \frac{7 - 2}{10} = \frac{5}{10} = +0.5$$

Note: If all students answer a question correctly (*IF = 1*), it means that the item is not only too easy but also non-discriminating (*ID = 0*). Similarly, if none of the students answer an item (*IF = 0*), it means that the item is not only too difficult but also non-discriminating (*ID = 0*). In summary:

If IF = 1, then ID = 0
If IF = 0, then ID = 0

Thus, there is a relationship between item facility and item discrimination. An item with a too high or low facility index is not likely to have a good discrimination power.

Another formula to calculate ID is:

$$ID = IF_{upper} - IF_{lower}$$

IF_{upper} = item facility for the upper group on the whole test
IF_{lower} = item facility for the lower group on the whole test

Let's calculate item discrimination for items 8 and 9 in the above table using the new formula. In the upper group two students answered correctly and in the lower group three students answered correctly. Notice that since we are calculating *IF* for upper/lower group and each of these groups consist of 5 examinees, *N* equals 5:

$$ID_8 = \frac{\sum C_{upper}}{N} - \frac{\sum C_{lower}}{N} = \frac{2}{5} - \frac{3}{5} = -\frac{1}{5} = -0.2$$

Scanned by CamScanner

In the upper group four students answered correctly and in the lower group one student answered correctly. Notice that since we are calculating *IF* for upper/lower group and each of these groups consist of 5 examinees, *N* equals 5:

$$ID_9 = \frac{\sum C_{upper}}{N} - \frac{\sum C_{lower}}{N} = \frac{4}{5} - \frac{1}{5} = \frac{3}{5} = 0.6$$

N.3. Choice distribution (CD)

Neither item facility nor item discrimination can provide the test constructor with necessary information about the appropriateness of the choices. Therefore, the third characteristic of an item, i.e. choice distribution, should be determined in order to improve the test both qualitatively and quantitatively. **Choice distribution** refers to:

(1) the frequency with which alternatives are assigned as the correct answer by the constructor. For example, if a test consists of 100 four-choice items, options A, B, C and D should each be assigned as correct items 25 times; and

(2) the frequency with which alternatives are selected by the examinees in a multiple-choice item. That is, through choice distribution, the tester examines the efficiency of the distracters. Accordingly there are three types of distractors:

Functioning (or **effective**)**:** If an item contains a distracter which attracts more low-scoring students, who have not mastered the subject, it is a *functioning* choice.

Non-functioning: If an item contains a distracter which attracted no one, not even the poorest examinees, it is a *non-functioning* choice which will increase the chances that some examinees will get the item right by guessing between or among the remaining two or three possibilities. As a rule of thumb, if less than 5% of the total population of examinees selects a distractor, that is a non-functioning distractor.

Mal-functioning: If an item contains a distracter which attracted more high than low scorers, it is a *mal-functioning* choice. Retaining such a distracter will actually harm the test. As a rule of thumb, if the number of examinees who select a distractor equals, or is around the number of examinees who selects the correct answer, the distractor is mal-functioning.

➲ **Example:**

	Choice Distribution			
	A	B	C	D
1	**60 (answer)**	10	30	0
2	37	13	**40 (answer)**	10
3	5	**50 (answer)**	5	40

In spite of the fact all the items have acceptable facility indexes, they have undesirable choice distribution patterns. In item 1, distractor B is selected by 10 and distractor C by 30 people and thus they are functioning. Distractor D, however, is non-functioning since it attracted no examinee.

In item 2, distractors B and D are functioning. On the other hand, distractor A is mal-functioning. It is selected by almost as many examinees as those who selected the correct answer. This implies that choice A is quite deceiving; it is either too close to the correct choice or it involves some unnoticed problem.

In item 3 distractors A and C are non-functioning because few examinees attempted them. Choice D is almost functioning.

➲ **Example**

(Choice C is the answer)

Choice	Highs	Lows	Total
A	3	8	11
B	3	7	10
C	14	5	19
D	0	0	0
	(20)	(20)	(40)

$$IF = \frac{\sum C}{N} = \frac{19}{40} = 0.47$$

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{14 - 5}{\frac{1}{2}(40)} = 0.45$$

As can be seen, the test has acceptable IF and ID indexes. Besides, distracters A and B are functioning since they have attracted a reasonable number of examinees most of whom are from low group. However, choice D is non-functioning because no examinee has selected it.

➲ **Example**

(Choice D is the answer)

Choice	Highs	Lows	Total
A	7	3	10
B	4	3	7
C	1	9	10
D	8	5	13
	(20)	(20)	(40)

$$IF = \frac{\sum C}{N} = \frac{13}{40} = 0.32$$

$$ID = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{8 - 5}{\frac{1}{2}(40)} = 0.15$$

While distracter C appears to be functioning, it is clear that distractors A and B are mal-functioning because they are attracting the wrong candidates (i.e. the better ones). There is no non-functioning distracter in this question. Generally, the item is difficult (IF= 0.32) and doesn't discriminate well (ID = 0.15).

➲ **Example: If choice A is the correct anwer, which item has the most satisfactory choice distribution?**

1) A = 40, B = 2, C = 28, D = 30
2) A = 35, B = 25, C = 20, D = 20
3) A = 60, B = 15, C = 15, D = 10
4) A = 65, B = 15, C = 10, D = 10

Scanned by CamScanner

The item in choice two is difficult (IF=0.35) and the item in choice four is easy (IF=65). In choice one, B=2, i.e. it is non-functioning. However, in choice three item facility is in the acceptable range and the other distractors have attracted almost equl number of testees.

The statistical procedures cited in former sections, i.e. item facility, item discrimination, and choice distribution are commonly used with NRTs. However, recall from chapter 2 that NRTs have characteristics different from CRTs. As a result, CRTs require totally different statistical procedures.

Recall that a central difference between NRTs and CRTs is that NRTs (typically designed to spread students out in percentile terms) are constructed to produce normal distribution, while CRTs (because they are designed to measure what students know or can do in percentage terms) do not necessarily do so.

CRTs may not necessarily produce scores that are normally distributed. In the beginning of the course (pre-test) we expect a positively skewed distribution and in the end of the course (post-test) we expect a negatively skewed distribution. On an ideal CRT designed to test course objectives, all the students would score 0 percent at the beginning of the course (indicating that they need to learn the material) and 100 percent at the end of the course (indicating that they have all learned the material). However, in reality, human beings are never perfectly ignorant at the beginning of a course nor are they perfectly knowledgeable at the end. Such distributions are only ideals.

One consequence of this fundamental difference in score distributions between the NRT and CRT categories of tests is that many of the statistics used for analyzing NRTs, which assume that the test scores are normally distributed, do not word very well for analyzing CRTs.

Item Discrimination in CRTs

If we again consider the ideal situation where students answer all the items wrong in the beginning of the course and answer all the items correct in the end of the course, in both cases item discrimination would be zero for a CRT.

$$ID_{pre-test} = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{\text{none of the testees in high group} - \text{none of the testees in low group}}{\frac{1}{2}N} = 0$$

$$ID_{post-test} = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{\text{all testees in high group} - \text{all testees in low group}}{\frac{1}{2}N} = 0$$

Item Facility in CRTs

Once teachers have selected those items which they judge to have high item quality, the resulting CRTs can be administered and statistical item analysis can proceed. As in CRT item analysis, item facility plays an important role. However, two possible item facilities exist for each item: one for the pre-test and one for the post-test. In CRT development, the goal is to find items that reflect whatever the students have learned, if anything. Hence, an ideal item for CRT purposes is one that has an IF of 0.00 at the beginning of instruction and another IF of 1.00 at the end of

instruction. Such pre-test and post-test IFs indicate that everyone missed the item at the beginning of instruction (that is, they needed to study the content or skill embodied in the term) and everyone answered it correctly at the end of the instruction (that is, they had completely absorbed whatever was taught). Of course, this example is an ideal item, in an ideal world, with ideal students, and an infallible teacher.

CRITERION-REFERENCED ITEM ANALYSIS

Much can be learned about each item in a CRT from comparing the performance on the item of those students who have studied the content (post-test) with those who have not (pre-test). Two different strategies can be used to make such a comparison.

The first approach called an **intervention strategy**, begins by testing students before instruction in a pre-test. At this stage, the students are *uninstructed*. The next step is to intervene with whatever instruction is appropriate and then test the *instructed* students on a post-test.

The second approach is the **differential groups strategy**. This strategy begins by finding two groups of students: one group that has the knowledge or skills that are assessed on the test and another group that lacks them. The test developer can then compare the item facility indexes of the first group, sometimes termed *masters*, with the item facility indexes for the second group, called *non-masters*. In either case, the item statistic that the tester calculates to estimate the degree of contrast between the two administrations of the test is called the difference index.

C.1. Difference index (DI)

The **difference index** indicates the degree to which an item is reflecting gain in knowledge or skill. In contrast to item discrimination, which shows the degree to which an NRT item separates the upper students from the lower students, the difference index indicates the degree to which a CRT item is distinguishing between the students who know the material or have the skill being taught, and those who do not. To calculate the difference index, the IF for the pre-test results (or non-masters) is subtracted from the IF for post-test results (or masters). The following formula is used to calculate difference index:

$$DI = IF_{post-test} - IF_{pre-test}$$

$IF_{post-test}$ = item facility for the post-test
$IF_{pre-test}$ = item facility for the pre-test

Just like item discrimination the range of difference index is:

$$-1 \leq DI \leq +1$$

➲ **Example: In pre-test 3 students answered an item correctly and in the post test 20 students answered the same item correctly. If there were 30 students in the class, different index equals ---------**

1) 0.70 2) 0.48 3) 0.66 4) 0.56

$$DI = IF_{post-test} - IF_{pre-test} \rightarrow \frac{\sum C_{pre-test}}{N} - \frac{\sum C_{post-test}}{N} = \frac{20}{30} - \frac{3}{30} = 0.56$$

Scanned by CamScanner

C.2. *B*-index

One problem that may crop up in using the difference index is that two administrations of the CRT are necessary. To solve this problem we can use *B*-index. The ***B*-index** is an item statistic that compares the *IFs* of those students who passed a test with the *IFs* of those who failed it. In other words, the masters and non-masters on the test are identified by whether or not they passed the test, and then the *B*-index indicates the degree to which the masters (students who passed the test in this case) outperformed the non-masters (students who failed the test) on each item. The following formula is used to calculate *B*-index:

$$B-index = IF_{Pass} - IF_{Fail}$$

IF_{Pass} = item facility for students who passed the test
IF_{Fail} = item facility for students who failed the test

Just like item discrimination the range of *B*-index is:

$$-1 \le B-index \le +1$$

➲ **Example: In a class which comprised of 60 students, 15 failed an exam. Among those who passed the test 40 and among those who failed the test 5 answered item three correctly. What is be *B*-index value of that item?**

1) 0.31 2) 0.55 3) 0.64 4) 0.40

$$B-index = IF_{Pass} - IF_{Fail} \rightarrow \frac{\sum C_{Pass}}{N} - \frac{\sum C_{Fail}}{N} = \frac{40}{45} - \frac{5}{15} = 0.55$$

6. VALIDATION

Through **validation**, which is the last step in test construction process, characteristics of a test as a total unit are determined. These characteristics include reliability, validity, and practicality. These characteristics will be investigated in the subsequent chapter.

Apply Logic

Though indexes of item facility and item discrimination along with information obtained from choice distribution are invaluable pieces of information, test developer should not totally depend on these characteristics. In rare cases, when a reasonable item shows poor item characteristics, by applying logic, test developer may still keep that item for a variety of reasons. Therefore, statistical information should be used as suggestive rather than determining tools in the exclusion or inclusion of a particular item in the test. To make a decision on whether an item should be excluded or included in a test, the theoretical assumptions behind the interpretation of scores must be taken into account. In other words, that a test is CR or NR.

7. EXTRA POINTS TO REMEMBER

7.1. A good test should never be constructed in such a way as to trap the students into giving an incorrect answer. When techniques of error analysis are used, the setting of deliberate traps or pitfalls for unwary students should be avoided. Care should be taken to avoid trapping students by including grammatical and vocabulary items which have never been taught. In the following item by seeing the adverb of time 'since Christmas', students are trapped to choose either choice three or four because adverbials starting with *since* are associated with present perfect or present perfect continuous

When I met Tim yesterday, it was the first time I ..... him since Christmas.
1) saw 2) had seen 3) have seen 4) have been seeing

Similarly, the following item trapped the more proficient students in a group by encouraging them to consider the correct answer, 'safety', as too simple to be right:

The animals tried to find ..... from the fire by running into the lake.
1) sanitation 2) safety 3) saturation 4) salutation

Scanned by CamScanner

State University Questions

1- The guessing-correction formula $R - \frac{W}{2}$ is used when there are -------- options in the item. (State University, 82)

1) two
2) three
3) four
4) five

	Key	Choice Distribution			
		A	B	C	D
1	B	10	60	30	0
2	A	40	50	5	5
3	D	42	10	8	40
4	C	12	15	60	13

2- The following item is poor because --------. (State University, 82)
She needs to get up earlier in order to buy an ..... clock.
a. time *b. alarm* *c. watch* *d. bell*
1) there is more than one correct answer
2) one of the options is a different form of words
3) one of the options is more difficult than the others
4) the correct answer is given away through a grammatical cue

3- A plausible distractor should --------. (State University, 82)
1) be eliminated
2) be incorrect when it stands itself
3) reflect the difference between L1 and L2
4) demand less proficiency than the correct answer languages

4- If 10 out of 50 answers to an item are incorrect, the facility value of the item will be --------. (State University, 82)
1) 10% 2) 20% 3) 40% 4) 80%

5- An item taken by 80 students produced 5 and 15 correct responses in the high and low groups, respectively. Provided that each group consists of 25% of the students, the discrimination index of the item will be --------. (State University, 84)
1) -0.5 2) -0.25 3) +0.25 4) +0.5

6- The item below is poor because there are --------. (State University, 84)
The missing girl was found with a ..... around her neck.
A) fence *B) rope* *C) umbrella* *D) bone*
1) tricky options
2) contextual clues
3) long and short options
4) words of opposite meaning

7- Item No -------- has satisfactory choice distribution. (State University, 84)
1) 1 2) 2 3) 3 4) 4

8- All of the following are among the main purposes of item analysis EXCEPT --------. (State University, 84)
1) locating malfunctioning distractors
2) calculating the discrimination index of test items
3) showing the correlation of an item with the others
4) determining the frequency of correct responses to an item

9- With respect to the following result, choices A, B, C (correct choice), and D are --------, respectively. (State University, 84)
1) malfunctioning, easy, functioning, functioning
2) non-functioning, easy, difficult, malfunctioning
3) non-functioning, malfunctioning, difficult, functioning
4) malfunctioning, difficult, easy, functioning

Choices	Choices distribution
A	0
B	60
C	25
D	15

10- In the following grammar test item, select the one option that is poorly constructed. (State University, 85)
In the morning the streets were flooded. There ..... a cloudburst during the night.
1) has been 2) must be 3) had be 4) must have been

11- In the formula ID = $\frac{H_C - L_C}{1/2N}$ used to determine the discrimination level of test items, N stands for the total number of --------. (State University, 85)
1) the testees in both groups
2) correct answers
3) total number of responses
4) correct answers by both groups

12- You may use contrastive analysis to select test items if --------. (State University, 85)
1) the students have been exposed to the same materials in class
2) language components are the focal point in your test
3) the problems of sampling do not matter to you
4) the students share the same native language

13- Among the following, the most satisfactory reading test item is characterized by --------. (State University, 86)
1) D=0.50, FV=0.75, r_{pb}=0.25
2) D=0.50, FV=0.85, r_{pb}=0.60
3) D=0.65, FV=0.50, r_{pb}=0.50
4) D=0.75, FV=0.45, r_{pb}=0.80

14- If an item in a test consisting of 40 items and taken by 50 students is answered correctly by 6 students in a high group and 2 students in the low group (each group constituting 20% of the total population), the discrimination index would be --------. (State University, 86)
1) 0.1 2) 0.2 3) 0.4 4) 0.8

15- The following item is poor because --------. (State University, 86)
In a multilingual country, people who ..... two languages have a big advantage
(A) speak *(B) talk* *(C) speaks* *(D) talks*
1) the stem is too long
2) it measures the sense of words
3) the correct choice depends on the testee's subjective judgment
4) the stem is too complex for the point to be measured

16- If choice D is the correct one, the item with the best item has the choice distribution --------. (State University, 86)
1) A=10, B=2, C=20, D=68
2) A=22, B=25, C=13, D=40
3) A=40, B=5, C=20, D=35
4) A=40, B=20, C=25, D=15

Scanned by CamScanner

17- **The item below has all of the following shortcomings EXCEPT ---------.** (State University, 86)

It is ..... to pass the exam to get the job.

(A) imperative *(B) unnecessary* *(C) polite* *(D) courteous*

1) there is one pair of opposite words
2) the choices widely vary in terms of the level of difficulty
3) the stem is conductive to more than one correct choice
4) the item measures knowledge of both inflection and derivation

18- **In a test, there were 100 questions, each consisting of three choices. One of the students answered 80 questions out of which as many as 20 were incorrect. After the application of the chance guessing formula, the student's score would be ---------.** (State University, 87)

1) 20 2) 30 3) 50 4) 60

19- **The possible ranges of reliability, item discrimination, and item facility are respectively ------.** (State University, 87)

1) –1 … +1; 0 … +1; –1 … +1
2) –1 … +1; 0 … +1; 0.37 … 0.63
3) 0 … +1; –1 … +1; 0 … +1
4) 0 … +1; –1 … +1; 0.37 … 0.63

20- **The following item is bad due to all of the following EXCEPT ---------.** (State University, 87)

When people ..... pictures of the atrocities on TV, there was a spontaneous reaction against the war, which seemed quite logical

a) saw *b) heard* *c) listened* *d) analyzed*

1) lexical difficulty of the stem
2) options with different lengths
3) options related to different concepts
4) unnecessary phrase/phrases in the stem

21- **To calculate "---------", you need to first rank-order the subjects and divide them into two halves.** (State University, 88)

1) item difficulty
2) item discrimination
3) concurrent validity
4) split-half reliability

22- **The item below is NOT good because ---------.** (State University, 89)

Mary ..... to school in the morning.

(a) goes *(b) went* *(c) had gone* *(d) would go*

1) most of the choices are non-functioning
2) some choices are too short
3) the stem provides insufficient context
4) it assesses the knowledge of both tense and verb

23- **Which of the following gives a proper interpretation for an item with a discrimination Index of 0.75 and a facility value of 0.52?** (State University, 89)

1) The item is at the right level of difficulty and discriminates well
2) The item discriminates well, but it is too difficult
3) The item is too difficult and discriminates in the wrong direction
4) The item discriminates poorly because it is too easy for everyone

24- **Which of the following statements about item facility is TRUE?** (State University, 89)

1) An item facility of +1 indicates that that item is ideal for NRT purposes.
2) Item facility ranges from –1 to +1.
3) An ideal item for CRT purposes is one that has an item facility of 0 at the beginning of instruction and +1 at the end.
4) Item facility is obtained by dividing the number of incorrect responses by the total number of responses.

25- **All of the following can be cited as features of a good multiple-choice test item EXCEPT that ------.** (State University, 90)

1) the stem is long enough to pose the problem
2) the responses are homogeneous, but not mutually exclusive
3) the stem sometimes begins with a blank to evoke specific cognitive processing strategies
4) the distracters are wrong, but plausible and discriminate at the right level

26- **Which of the options in the following multiple-choice grammar item had better be revised?**

No one expected all the trainees to ..... all the words introduced in the previous sessions. (State University, 90)

1) to have learned 2) to have learn 3) to learning 4) to be learned

27- **In case the testees in both low and high groups answer an item correctly in a 20-item test, ------** (State University, 91)

1) variance will be –1
2) mean will be 10
3) item facility will be .5
4) item discrimination will be zero

28- **An average difficult level of --------- will result in maximum differentiation among test-takers.** (State University, 92)

1) 0 2) 0.50 3) 0.75 4) 1

29- **In language testing, the difference index (DI) is used to indicate the degree to which an item------** (State University, 92)

1) discriminates low and high students
2) indicates item distribution
3) reflects gain in knowledge
4) indicates item facility

30- **If all the high-ability students in a class answer a test item correctly, and all the low-ability students fail to answer it correctly, the item discrimination index of this item would be --------.** (State University, 93)

1) –1 2) 0 3) 0.5 4) +1

31- **Why is the following test item poorly constructed?** (State University, 93)

Everyone was ..... with John's performance on the test.

a. amazed *b. satisfied* *c. discouraged* *d. ashamed*

1) There is a grammatical clue which gives away the correct answer.
2) The options are not at the same level of difficulty.
3) There is more than one correct answer.
4) The distractors are very tricky.

Scanned by CamScanner

State University Answers

1. Choice 2
☞Based on the guessing correction formula $Score = Right - \frac{Wrong}{n-1}$, when the denominator is 2, the number of options is three: $n - 1 = 2 \rightarrow n = 3$

2. Choice 4
☞Determiner 'an' is a contextual clue by the help of which testee selects *alarm*.

3. Choice 4
☞Distractors should not be too difficult nor demand higher proficiency in the language than the correct option. If they are too difficult, they will succeed only in distracting the good student, who will be led into considering the correct option too easy.

4. Choice 4

$$IF = \frac{\sum C}{N} = \frac{40}{50} = 0.8 \times 100 = 80\%$$

5. Choice 1
☞Since each group consists of 25% of the population, so: $80 \div 4 = 20$

$$ID = \frac{5 - 15}{20} = -0.5$$

6. Choice 2
☞Determiner 'an' is a contextual clue by the help of which testee eliminateds *umbrella*.

7. Choice 4
☞In item 1, choice D is nonfunctioning. In item 2, choice B is malfunctioning. In item 3, choice A is malfunctioning. In choice four most testees have selected the correct option and the distracters have almost equally attracted the same number of examinees.

8. Choice 3
☞Choice one and three refer to choice distribution. Choice two refers to item discrimination.

9. Choice 3
☞Choice A is nonfunctioning because it hasn't attracted anyone; Choice B is malfunctioning because it has attracted more students than the correct answer; choice C is difficult because IF = 0.25.

10. Choice 3
☞In an item of grammar, all distractors should stand grammatically correct, therefore choice three is poorly constructed.

11. Choice 3
☞Refer to section N.2.

12. Choice 4
☞Refer to section 1.

13. Choice 4
☞In choice one, item is easy and point biserial correlation (rpb) is low. In choice two, item is easy. Though the indexes in choices three and four are standard, the indexes in choice four is more desirable.

14. Choice 3
☞Each group consists of 20% of the total population → $50 \div 5 = 10$

$$ID = \frac{6 - 2}{10} = \frac{4}{10} = 0.4$$

15. Choice 4
☞The test constructor who is measuring the use of third person singular *-s* has chosen a difficult stem for this purpose.

16. Choice 2
☞In choice 1 *B* is nonfunctioning; In choice 3 *A* is malfunctioning; In choice 4, *A* is malfunc-tioning.

17. Choice 4

18. Choice 3
☞The student provided 60 correct answers:

$$score = R - \frac{W}{n - 1} = 60 - \frac{20}{3 - 1} = 50$$

19. Choice 3
☞Refer to section 5.

20. Choice 2
☞Some test writers argue that the options should be related to the same general topic or area, e.g.
Apparition means ..... a) skeleton b) ghostc) corpse d) nightmare

21. Choice 2
☞Refer to section N.2.

22. Choice 3
☞Because the stem is not long enough, there is more than one correct response to the item.

23. Choice 1
☞Refer to section 5.

24. Choice 3
☞Refer to section Item Facility in CRTs.

25. Choice 3
☞Refer to section 3.

26. Choice 2
☞One of the guidelines for MC construction is "All distracters should be plausible". In the given item 'to have learn' is grammatically poorly cpnstructed.

27. Choice 4
☞If there were, for example, 20 students in the class then,

$$\text{Item discrimination} = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{10 - 10}{\frac{1}{2}20} = 0$$

28. Choice 2
☞If a test is too easy or a test item m that is too difficult can tell us nothing about the differences in ability within the test population. A difficulty level falling the in the middle of the range guarantees some variance in scores among the examinees.

29. Choice 3
☞Refer to section C.1.

30. Choice 4
☞When all the examinees in the high group and none of the examinees in the low group answer an item correctly, ID = +1.
When all the examinees in the low group and none of the examinees in the high group answer an item correctly, ID = –1.

31. Choice 1

Scanned by CamScanner

Azad University Questions

1- If item facility equals 1, item discrimination will be --------. (Azad University, 83)
1) 0 2) 1 3) 0.33 4) 0.67

2- Some of the shortcomings of the items in a test can be accounted for through item --------. (Azad University, 83)
1) facility 2) discrimination 3) difficulty 4) analysis

3- If 6 testees from among a total number of 24 answer a question correctly, what is the item facility index of it? (Azad University, 83)
1) 0.18 2) 0.25 3) 0.75 4) 0.50

4- If the number of the correct responses in the lower group exceeds that of the higher one, the test is regarded to be --------. (Azad University, 83)
1) very easy 2) very difficult 3) inappropriate 4) valid and reliable

5- 15 testees in the lower group and 45 in the higher group answered a question correctly. If all of the subjects in the upper group have answered correctly, what is the item discrimination of this given item? (Azad University, 83)
1) 0.75 2) 0.90 3) 0.86 4) 0.67

6- A distractor is mal-functioning if it attracts --------. (Azad University, 84)
1) high and low students equally 2) more low than high students
3) more high than low students 4) some of the students

7- Determining the characteristics of an item altogether is called --------. (Azad University, 84)
1) item facility 2) choice distribution 3) item discrimination 4) validation

8- If the item discrimination of an item equals 1, what would its item facility be? (Azad University, 84)
1) 1 2) -1 3) 0.0 4) 0.5

9- The most important task in pretesting is --------. (Azad University, 84)
1) analyzing the result 2) reviewing the items
3) preparing directions 4) planning the test

10- To decide on the form of a test, certain elements such as the following are involved EXCEPT the --------. (Azad University, 85)
1) nature of the attribute 2) function of the test
3) authenticity of the attribute 4) physical appearance

11- The purpose of pretesting is to determining the characteristics of the individual items including the following EXCEPT --------. (Azad University, 85)
1) IF 2) ID 3) CD 4) SD

12- In constructing a test, which of the following points must be determined in advance? (Azad University, 85)
1) the form of the test 2) the function of the test
3) the content of the test 4) the time needed

13- Which one refers to the power of distinguishing among the testees on the basis of their language proficiency? (Azad University, 86)
1) item facility 2) item analysis 3) item difficulty 4) item discrimination

14- A number of scholars believe that an impure MC item is one which --------. (Azad University, 87)
1) involves non-English options 2) tests more than one thing at a time
3) presents a single problem 4) has lengthy distractors

15- In the process of -------- a test, it undergoes some changes to fit the needs and purpose of the particular language program. (Azad University, 89)
1) scoring 2) developing 3) administering 4) adapting

16- In the multidimensional process of test construction, which of the following steps should be taken first? (Azad University, 89)
1) Determining the content of the test 2) Determining the form and function of the test
3) Preparing the test items 4) Examining the instructional objectives

17- If the IF value of a test is 0.1, we can conclude --------. (Azad University, 91)
1) it is a very easy item for all tests takers
2) it provides the maximum differentiation among tests takers
3) it is a very difficult item for all except the most qualified applicants
4) it is an acceptable value to differentiate qualified from unqualified applicants

Azad University Answers

1. Choice 1
☞Refer to section N.2.

2. Choice 4
☞Refer to section 5.

3. Choice 2

$$IF = \frac{\sum C}{N} = \frac{6}{24} = 0.25$$

4. Choice 3
☞In calculating item discrimination, when we divide students into two halves, it is supposed that more students in the high group would answer an item than the low group. If the case is reversed, there is some problem with the item.

5. Choice 4
☞Since all 45 students in the high group answered the item correctly, we infer that in the low group there are 45 students, i.e. the total number of testees is 90:

$$\text{Item discrimination} = \frac{C_{high} - C_{low}}{\frac{1}{2}N} = \frac{45 - 15}{\frac{1}{2}90} = 0.66$$

6. Choice 3
☞Refer to section N.3.

7. Choice 4
☞Refer to section 6.

8. Choice 4
☞When item discrimination of an item is 1, it means all the students in high group answered it correctly and all the students in low group answered incorrectly, that is, *half the learners answered correctly*. Therefore item facility equals 0.5.

Scanned by CamScanner

9. Choice 1

☞Refer to section 5.

10. Choice 3

☞Refer to section 1.

11. Choice 4

☞Refer to section 5.

12. Choice ?

☞According to FJB, p. 83 choices 1, 2, and 3 should be determined in advance.

13. Choice 4

☞Refer to section N.2.

14. Choice 2

☞Refer to section 7.

15. Choice 4

☞A newly developed test may work fairly well in a program, but perhaps not as well as was originally hoped. Such a situation would call for further adapting of the test so that it better fits the needs and purposes of the particular language program. The process of adapting a test to a specific situation will involve some variant of the following steps:

1) Administer the test in the particular program, using the appropriate teachers and students;
2) Select those test questions that word well at spreading out the students (for NRTs), or are efficient at measuring the learning of the objectives (for CRTs) in this particular program;
3) Develop a shorter, more efficient revision of the test – one that fits the program's purposes and works well with its students; and
4) Evaluate the quality of the newly revised test.

16. Choice 2

☞Refer to section 1.

17. Choice 3

☞The closer the index of IF to zero, the harder the item.

Scanned by CamScanner