ACTIVITY 4

Item Analysis- Discrimination and difficulty index

A. Give the term described/explained.

____Item analysis__ 1. Refers to a statistical technique that helps instructors identify the effectiveness of their test items.

___Item Difficulty__ 2. Refers to the proportion of students who got the test item correctly.

Discrimination Index 3. Which is the difference between the proportion of the top scorers who got an item correct and the proportion of the bottom scorers who got the item right.

___Item difficulty ___4. Which one is concerned with how easy or difficult a test item is?

____Item Analysis__5. Which adjective describes an effective distracter?

B. **Write TRUE if the statement is correct and FALSE if it is wrong**.

_TRUE__ 1. Difficulty index indicates the proportion of students who got the item right.
FALSE 2. Difficulty index indicates the proportion of students who got the item wrong.
_TRUE__3. A high percentage indicates an easy item/question and a low percentage indicates a difficult item.
_TRUE__4. Authors agree, in general, that items should have values of difficulty no less than 20% correct and no greater than 80%.
_TRUE_5. Very difficult or very easy item contribute greatly to the discriminating power of a test.
_FALSE_6. The discrimination index range is between -1 to +2.
_FALSE_7. The farther the index is to +1, the more effectively the item distinguishes between the two groups of students.
TRUE 8. When the items discriminates negatively, such item should be revised and eliminated from se scoring.
_FALSE__ 9. A positive discrimination index indicates that the lower performing students actually. Selected the key or correct response more frequently than the top performer.
_TRUE__10. If no one else selects a distracter it is important to revise the option and attempt to make the distracter more plausible choice.

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/

C. Solving problem

1. Solve for the difficulty index of each test item.

ITEM NO	1	2	3	4	5
No, of correct responses	2	10	20	30	15
No, of students	50	30	30	30	40
Difficulty index	**0.04**	**0.33**	**0.67**	**1**	**0.38**

1. **What is most difficult? Most easy?**
   = The most difficult is **Item # 1** and the most easy is the **item # 4**
2. Which need revision? Which should be discarded?
   = The item # 1 need revision and the item # 4 need discarded

D. Solve for the discrimination indexes of the following test item?

Item no	1		2		3		4		5	
	UG	**LG**	**UG**	**LG**	**UG**	**LG**	**UG**	**LG**	**UG**	**LG**
No, of Correct Responses	12	20	10	20	20	10	10	24	20	5
No, of students	25	25	25	25	25	25	23	25	25	25
Discrimination index	**- 0.32**		**- 0.4**		**0.4**		**-0.56**		**0.6**	

D1 = 12-20 / 25 = **- 0.32** **D**2 = 10-20 / 25 = **-0.4** **D**3 = 20-10 / 25 = **0.4** **D**4= 10-24 / 25 = **-0.56** **D**5 = 20-5 / 25 = **0.6**

Based on the computed discrimination index, which are good test items? Not good item?

= The good item is the D3 = 0.4 because good test should be > 0 and the not good test are D1, D2, and D4 because its < 0 and they are all negative.

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/

E. A multiple choice type of test has 5 option. The table below indicates the number of examinees out of 50 who choose each option.

OPTION				
A	**B**	**C**	**D**	**E**
0	20	15*	5	10

***- Correct answer**

*** Which is option are plausible? Which one is implausible?**

= The most plausible is letter **B** because it has most option to take the exam and it is also most answerable exam compared to other, and the implausible **A** because it is the most difficult to answer.

F. Study the following data. Compute for the difficulty index and the discrimination index of each set of scores

A. N = 80, number of wrong answers: upper 25%=2 lower 25%=9
B. N = 30, number of wrong answers: upper 25%=1 lower 25%=6
C. N = 50, number of wrong answers: upper 25%=3 lower 25%=8
D. N = 70, number of wrong answers: upper 25%=4 lower 25%=10

	1		2		3		4	
Item No 1.	UG	LG	UG	LG	UG	LG	UG	LG
No. of Correct Answer	38	31	14	9	22	27	31	25
Total Number of Students	40	40	15	15	25	25	35	35
Difficulty Index	0.86		0.777		0.78		0.80	
Discrimination Index	0.18		0.33		0.2		0.17	

G. Write test items to check if the following learning outcomes were realized.

H. Enumerates the three types of validity evidence. Which of these types of validity is the most difficult to measure?

- Content- related evidence of validity
- Criterion-related evidence of validity
- Construct-related evidence of validity

= The most difficult to measure is the Construct-related evidence of validity because several measures are usually required to demonstrate it, including pilot studies and clinical traits.

I. What is the relationship between validity and reliability? Can a test be reliable and yet not valid? Illustrate

= Validity refers to whether you are measuring what you are supposed to be measuring, while reliability refers to whether your results are consistent**.** If an instrument or experiment is

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/

valid, it is usually also reliable if it is built properly to control all variables save the one being researched.

J. Discuss the different measures of reliability. Justify the use of each measure in the context of measuring reliability. Can a text be valid and not reliable?

Type of reliability	Measures the consistency of	What is my methodology?
Test-retest	The same test over time	Measuring a property that you expect to stay the same over time
Interrater	The same test conducted by different people	Multiple researchers making observations or ratings about the same topic
Parallel forms	Different versions of a test which are designed to be equivalent	Using two different tests to measure the same thing
Internal consistency	The individual items of a test.	Using a multi-item test where all the items are intended to measure the same variable

1. **Test-retest reliability**- test-retest reliability measures the consistency of results when you repeat the same test on the same sample at a different point in time. You use it when you are measuring something that you expect to stay constant in your sample. A test of colour blindness for trainee pilot applicants should have high test-retest reliability, because colour blindness is a trait that does not change over time.

   Why it's important

   Many factors can influence your results at different points in time: for example, respondents might experience different moods, or external conditions might affect their ability to respond accurately. Test-retest reliability can be used to assess how well a method resists these factors over time. The smaller the difference between the two sets of results, the higher the test-retest reliability.

   How to measure it

   To measure test-retest reliability, you conduct the same test on the same group of people at two different points in time. Then you calculate the correlation between the two sets of results.

   Test-retest reliability example

   You devise a questionnaire to measure the IQ of a group of participants (a property that is unlikely to change significantly over time).You administer the test two

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/

months apart to the same group of people, but the results are significantly different, so the test-retest reliability of the IQ questionnaire is low.

Improving test-retest reliability

When designing tests or questionnaires, try to formulate questions, statements and tasks in a way that won't be influenced by the mood or concentration of participants.

When planning your methods of data collection, try to minimize the influence of external factors, and make sure all samples are tested under the same conditions.

Remember that changes can be expected to occur in the participants over time, and take these into account.

2. **Interrater reliability**- Interrater reliability (also called interobserver reliability) measures the degree of agreement between different people observing or assessing the same thing. You use it when data is collected by researchers assigning ratings, scores or categories to one or more variables.

   In an observational study where a team of researchers collect data on classroom behavior, interrater reliability is important: all the researchers should agree on how to categorize or rate different types of behavior.

   Why it's important

   People are subjective, so different observers' perceptions of situations and phenomena naturally differ. Reliable research aims to minimize subjectivity as much as possible so that a different researcher could replicate the same results.

   When designing the scale and criteria for data collection, it's important to make sure that different people will rate the same variable consistently with minimal bias. This is especially important when there are multiple researchers involved in data collection or analysis.

   How to measure it

   To measure interrater reliability, different researchers conduct the same measurement or observation on the same sample. Then you calculate the correlation between their different sets of results. If all the researchers give similar ratings, the test has high interrater reliability.

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/

Interrater reliability example

A team of researchers observe the progress of wound healing in patients. To record the stages of healing, rating scales are used, with a set of criteria to assess various aspects of wounds. The results of different researchers assessing the same set of patients are compared, and there is a strong correlation between all sets of results, so the test has high interrater reliability.

Improving interrater reliability

- Clearly define your variables and the methods that will be used to measure them.
- Develop detailed, objective criteria for how the variables will be rated, counted or categorized.
- If multiple researchers are involved, ensure that they all have exactly the same information and training.

3. **Parallel forms reliability**- Parallel forms reliability measures the correlation between two equivalent versions of a test. You use it when you have two different assessment tools or sets of questions designed to measure the same thing.

Why it's important

If you want to use multiple different versions of a test (for example, to avoid respondents repeating the same answers from memory), you first need to make sure that all the sets of questions or measurements give reliable results.

In educational assessment, it is often necessary to create different versions of tests to ensure that students don't have access to the questions in advance. Parallel forms reliability means that, if the same students take two different versions of a reading comprehension test, they should get similar results in both tests.

How to measure it

The most common way to measure parallel forms reliability is to produce a large set of questions to evaluate the same thing, then divide these randomly into two question sets.

The same group of respondents answers both sets, and you calculate the correlation between the results. High correlation between the two indicates high parallel forms reliability.

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/

Parallel forms reliability example

A set of questions is formulated to measure financial risk aversion in a group of respondents. The questions are randomly divided into two sets, and the respondents are randomly divided into two groups. Both groups take both tests: group A takes test A first, and group B takes test B first. The results of the two tests are compared, and the results are almost identical, indicating high parallel forms reliability.

Improving parallel forms reliability

- Ensure that all questions or test items are based on the same theory and formulated to measure the same thing.

4. **Internal consistency**- Internal consistency assesses the correlation between multiple items in a test that are intended to measure the same construct. You can calculate internal consistency without repeating the test or involving other researchers, so it's a good way of assessing reliability when you only have one data set.

Why it's important

When you devise a set of questions or ratings that will be combined into an overall score, you have to make sure that all of the items really do reflect the same thing. If responses to different items contradict one another, the test might be unreliable. To measure customer satisfaction with an online store, you could create a questionnaire with a set of statements that respondents must agree or disagree with. Internal consistency tells you whether the statements are all reliable indicators of customer satisfaction.

How to measure it: Two common methods are used to measure internal consistency.

- **Average inter-item correlation**: For a set of measures designed to assess the same construct, you calculate the correlation between the results of all possible pairs of items and then calculate the average.
  - **Split-half reliability**: You randomly split a set of measures into two sets. After testing the entire set on the respondents, you calculate the correlation between the two sets of responses.

Internal consistency example

A group of respondents are presented with a set of statements designed to measure optimistic and pessimistic mindsets. They must rate their agreement with each statement on a scale from 1 to 5. If the test is internally consistent, an optimistic respondent should generally give high ratings to optimism indicators and low ratings to pessimism indicators. The correlation is calculated between all the responses to the "optimistic"

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/

statements, but the correlation is very weak. This suggests that the test has low internal consistency.

Improving internal consistency

- Take care when devising questions or measures: those intended to reflect the same concept should be based on the same theory and carefully formulated

This study source was downloaded by 100000893346523 from CourseHero.com on 11-04-2024 09:15:26 GMT -06:00

https://www.coursehero.com/file/113675708/MIDTERM-ACTIVITY-PAGE-103-105docx/