

Lecture: 23

Creating Fixed-Choice Test Items: Matching Exercises

Matching test items are also selection items specially used to measure a student's ability to identify the relationship between a set of similar items, each of which has two components.

Advantages Matching Exercises

- Ease of construction
- Measure a large amount of related factual material in a relatively short time

Limitations of Matching Exercises

- Restricted to the measurement of factual information
- Highly susceptible to the presence of irrelevant clues
- Difficulty of findings homogenous material

Lecture: 24

Suggestions for Constructing Matching Exercises

- Use only homogeneous material in the single matching exercise
- Include an unequal number of responses and premises
- Arrange the list of responses in logical order
- Indicate the directions the basis for matching the responses

Lecture: 25

Forms of Supply Type

- Short answers
- Completion items

They only differ in form of presentation. Short answers are presented in the form of direct questions while completion items are incomplete statements.

Both of them can be answered with a word, phrase, number, or symbol.

Lecture: 26

Advantages of Short Answer/Completion Questions

- Teacher wants to control the possibility of guessing
- Teacher wants students to complete a large number of items in a fairly short time
- The possibility of guessing the correct answers to these questions is greatly reduced when compared with true-false questions

Disadvantages

- Difficult to frame questions for one specific answer
- Not usable to measure complex learning outcomes

Lecture: 27

Suggestions for Constructing Short Answer/Completion Questions

- If the answer is to be expressed in numerical units, indicate the type of answer
- When completion items are used, do not include too many blanks or start with a blank space

Lecture: 28

Constructed Response Test Items

Subjective test items (essay questions) are constructed response type questions that can be the best way to measure the students' higher order thinking skills. Such as applying, organizing, synthesizing, integrating, evaluating, or projecting.

Essay type questions are divided into two types

1. Restricted Response Items
2. Extended Response Items

Examples:

- Analysis of relationship.
- Compare and contrast positions.
- Explain cause-effect relationship

Guidelines for Constructing Restricted Response Essay Type Items

- Prepare a tentative scoring key in advance
- Do not employ optional questions
- Verify a question's quality by writing a trial response to the question
- Evaluate essay responses anonymously
- Specify the value and an approximate time limit for each question

Lecture: 29

Advantages of essay type items (Extended response questions)

- Effective for assessing higher order abilities
- Comparatively less time consuming
- Emphasizes essential communications skills

Scoring Rubrics provide two benefits

- They support the examination to the extent to which the specified criteria has been reached.
- They provide feedback to students concerning how to improve their performance

Types of Scoring Rubrics

- The analytic (scoring key, point or rating) method
 - The analytic scoring requires developing a list of major elements that students are expected to include in an ideal answer of a question
 - Decide the number of points awarded to students when they include each element
 - When assigning number to each element, be sure the total points should match with the essay's total value
- The holistic (also called global, sorting or rating) method
 - The holistic scoring rubric requires making a judgment about the overall quality of each student's response to an item.
 - No need to mark each specific content element that a student included in the answer.

Lecture: 30

Item Analysis

It reveals if an item is:

- Too easy
- Too difficult
- Failing to show a difference between skilled and unskilled examinees

The two most common statistics reported in Item Analysis

1. **Item difficulty** (which is a measure of the proportion of examinees who responded to an item correctly)
2. **Item discrimination** (which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area and those who are not).

Distractor analysis - provides a measure of how well each of the incorrect options contributes to the quality of a multiple choice item.

Appropriate Time for Item Analysis

1. After administration, but before scoring – it helps the test developers identify errors in the scoring key or defects in the items.
2. After scores have been reported – it helps test developers select items for reuse in future forms of the tests.

Lecture: 31

Theories of Test Development/Item Analysis

1. Classical test theory (CTT)
2. Item response theory (IRT)

Classical test theory is more focused on finding the true score of an individual on a test and called as true score theory
 $\text{Observed Score} = \text{True score} + \text{Error}$

Item response theory is known as latent trait theory as its focus is to find the item characteristics as a function of ability

Item Discrimination in CTT

The ability of an item to discriminate between higher ability examinees and lower ability examinees is known as item discrimination.

It is the relationship between how well students did on the item and their total exam score. The range is from –1.00 to 1.00.

The higher the value, the more discriminating the item. A highly discriminating item shows that the students who had high exams scores got the item correct students who had low exam scores got the item incorrect.

Lecture: 32

Item Characteristic Curve (ICC) in Item Response Theory

The item characteristic curve is considered as the basic building block of item response theory.

Methodological properties of an ICC

- Item difficulty - The application of Item difficulty in IRT is defined as the ability at which the probability of success on the item is .5 on a logit scale.

- Discrimination power - The items on a test might also differ in terms of the degree to which they can differentiate individuals who have high trait levels from individuals who have low trait levels.
- Probability of answering correctly by guessing - Guessing means giving an answer or making a judgment about something without being sure of all the facts

Types of Guessing

1. **Blind guessing:** Where an examinee chooses an answer at random from among the alternatives offered.
2. **Informed guessing:** Where the examinee draws upon all his knowledge and abilities to choose the answer most likely to be correct.

Lecture: 33

Issues in test administration

- Cheating
- Test anxiety
- Poor testing conditions
- Errors in test scoring procedure

Packing of test involves:

- Grouping together items of similar format
- Arranging test items from easy to hard
- Properly spacing items
- Keeping items and options on the same page
- Position illustration near description (if diagram given)
- Randomness of answer key
- Determine how students record answers (separate sheet or same sheet)

Lecture: 34

Reproducing the Test

Reproduction of test involves:

- 1) **Knowing the photocopying machine**
 - Expert operator
 - Toner quality
 - Atomization facility in machine
- 2) **Specifying copying instructions**
 - Randomly checking of every 19th copy
 - Paper size
 - Margins
 - Ordering of pages
- 3) **Filing original test**
 - Master copy for reference and reuse
 - Use as reference in random checking

Lecture: 35

Things to remember when administering test

- Maintain a positive attitude
- Maximize achievement motivation

- Monitor students
- Equalize advantages
- Avoid surprises
- Rotate distribution
- Remind students to check their copies
- Give time warning
- Collect test uniformly

Lecture: 36

Scoring Criteria - Planning how responses will be scored leads to rethinking and clarifying the questions so that students have a clearer idea of what is expected.

Scoring Rubric - A rubric is an explicit set of criteria used for assessing a particular type of work or performance and provides more details than a single grade or mark.

Steps to design Rubric

- Identify a learning goal.
- Choose outcomes that may be measured.
- Develop or adapt existing rubric.
- Share it with students

Elements of Rubric

- Score
- Criteria
- Level of performance
- Descriptors

Holistic Scoring Rubric - These rubrics are good for evaluating overall performance on a task. All criteria are assessed as a single score.

When to use:

- There is no correct answer/response to ask a task
- Focus is on overall quality, proficiency
- The assessment is summative

Lecture: 37

Approaches to Apply Holistic Scoring Rubric

1. Implement holistic rubric is to decide beforehand on the number of quality categories into which you will sort the student's answers.
2. Craft a holistic rubric, which defines the qualities of paper that belong in each category
3. Select specimen papers, which are good examples of each scoring category
4. Read the answer completely and one with another to decide which are the best

Levels of Rubric

1. Ineffective
2. Adequate
3. Effective

4. Outstanding

Advantages

- Quick scoring
- Provides overview of student achievement

Disadvantages

- Does not provide detailed information.
- May be difficult to provide one overall score.

Analytic Scoring Rubric - Each criterion is assessed separately by using different descriptive ratings. Each criterion is given a separate score. Final score is made up of adding each component parts

When to use:

- Several faculties are collectively assessing student work.
- Outside audience will be examining rubric scores.
- Profiles of specific strength/weakness are desired

Advantages

- More detailed feedback.
- Scoring more consistent across students and grades.

Disadvantages

- Time consuming to score.

Suggestions for scoring constructed response questions

- Prepare an outline of the expected answer in advance.
- Use the scoring rubric that is most appropriate.
- Decide how to handle factors that are irrelevant to the learning outcomes being measured.
- Evaluate all responses to one question before going on to the next one.
- When possible, evaluate the answers without looking at the student's name.

Lecture: 38

Standardized achievement test - a fixed set of items to measure defined achievement domain, specific direction for administering and scoring test. These tests are often customized to include characteristics of both norm and criterion referenced test.

Characteristics of Standardized Achievement Test

- The test items are of highly technical quality
- A test manual is used as guides for administering the test
- Directions for administering and scoring are so precisely stated
- Equivalent and comparable forms of the test are usually provided
- Norms based on national samples of students in the grade where the test is intended for use are provided

Standardized Test vs. Informal Classroom Test

- Quality of test items
- Reliability of the tests
- Procedure for administering and scoring

- Interpretation of scores
- The nature of the learning outcomes and the content measured

Lecture: 39

Standardized test batteries

Standardized achievement tests are frequently used in the form of survey test batteries. A **battery** consists of a series of individual tests all standardized on same national sample of students. The selection of battery should be based on its relevance to the schools objectives.

Reading test

Such tests commonly measure

- 1) Vocabulary
- 2) Reading comprehension
- 3) Rate of reading

Reading survey test measures only some of the outcomes of reading instruction, the mechanics of reading is measured by diagnostic reading tests.

Concept of Interpreting Test Scores

Test scores can be interpreted in the terms of:

1. Types of tasks that can be performed (Criterion referenced or standard based)
2. Relative position held in some reference group (norm reference)

Lecture: 40

Raw scores - It is simply the number of points received on a test. It is a numerical summary of a student's test performance.

Derived Scores - It is a numerical report of test performance on a score scale that has well-defined characteristics and yields normative meaning.

Criteria Most Desired in Norms

Test norms should be

1. Normal
2. Representative
3. Up to date
4. Comparable
5. Adequately described

Cautions in Interpreting Test Scores

A test score should be interpreted

- In terms of the specific test form which it was derived.
- In light of all of the student's relevant characteristics.
- According to the type of decision to be made.

Lecture: 41

High Stake Testing - The use of test and assessment to make decisions that are of prominent educational, financial, or social impact.

Decision based on high stake testing

- Promotion to next grade
- Awarding of diploma or degree
- Evaluation of school performance
- Incentives and accountability of school staff

High Stake Testing in Pakistan

In school education in Pakistan, high stake testing is conducted at each level from primary to higher secondary.

- 1) Grade 5 and 8 examination by Punjab Examination Commission.
- 2) Grade 10 and 12 examination by BISEs

Recommendations for effective High Stake Testing

- Adequate resources
- Validation for each intended separate use
- Full disclosure of likely consequences
- Alignment between test and curriculum

Preparation for effective High Stake Testing

- 1) Focus on task
- 2) Inform presents
- 3) Teach test taking skills
- 4) As the test day approaches respond to students questions openly and directly
- 5) Take advantage of whatever preparation material is available

Lecture: 42

Institutions involve in assessment

1. Examination commission
2. Board of intermediate and secondary education (BISEs)
3. Board of Technical Education
4. National Education Assessment System (NEAS)
5. Provincial Education Assessment Centre (PEAS) in KPK and Sindh

National Education Assessment system (NEAS) works for promoting quality learning among children of Pakistan by carrying out fair and valid national assessment.

Objectives of NEAS

- Informing policy
- Monitoring standards
- Identifying correlation of achievement
- Directing teachers' efforts

