

Items and Item Scores

Measurement Methods in Psychological
Research

Lecture 3 – 02/02/2006



Today's Lecture

- An account of nine main types of items that compose objective tests.
 - How each type of item is scored.
 - How scoring each item has implications for analysis.

Definitions

- ❑ **Objective Tests** are tests composed of objectively scorable items.
- ❑ An **Item** includes item stem (stimulus material) and all response options.
- ❑ **Item Scoring** is the method of assigning numbers to the item responses.
- ❑ An **Item Score** is a number assigned to an item response.
- ❑ A **Test Score** is a combination of item scores.

Item Type #1 – Completion or Constituted-Response Item

- A completion or constituted response item is an item where the respondent constructs answer.
- Example:
 - $2 + 2 = \underline{\quad\quad} ?$

Item Type #1 – Completion or Constituted-Response Item

- Scoring method:
 - 1 for correct answer.
 - 0 for incorrect answer.
- Benefits:
 - Virtually eliminates guessing by examinees on cognitive tests.
- Problems:
 - Can be difficult to define all possible “correct” answers.
 - No credit for partial knowledge

Item Type #2 – Multiple Choice Item

- A multiple-choice item consists of a keyed (correct) option and a set of distractors (or foils).
- Example:
 - If items on an achievement test are scored 0 (incorrect) and 1 (correct), then the mean score on each item is:
 - A. Equal to the item difficulty.
 - B. Equal to (1-item difficulty).
 - C. Equal to item difficulty divided by sample size.
 - D. None of the above.

Item Type #2 – Multiple Choice Item

- Scoring method:
 - 1 for keyed (correct) answer.
 - 0 for non-keyed (incorrect) answer.
- Benefits:
 - Understandable to examinees.
- Problems:
 - Examinees can guess correctly if they do not know the answer.
 - Difficult to imagine non-cognitive items developed this way.

Item Type #3 – Two-choice Dichotomous Item

- ❑ A two-choice or dichotomous item has just two response options.
- ❑ Examples:
 - The mode is a measure of central tendency TRUE/FALSE.
 - Compared to water, the density of ice is GREATER/LESS.
 - I often suffer from sleeplessness TRUE/FALSE.
 - I am generally HAPPY/UNHAPPY.
 - Assault weapons do not belong in private hands AGREE/DISAGREE.
 - I feel anxious right now TRUE/FALSE.

Item Type #3 – Two-choice Dichotomous Item

- Scoring method:
 - 1 for keyed (correct) answer.
 - 0 for non-keyed (incorrect) answer.
- Benefits:
 - Understandable to examinees.
- Problems:
 - Examinees can guess correctly if they do not know the answer (this is not a problem in non-cognitive tests).
 - Difficult to imagine non-cognitive items developed this way.

Item Type #4 – Checklist

- A checklist item has three or more options, each having the possibility of being keyed (correct) or not (incorrect).
- Examinees must select *all* correct options.
- Example:
 - Which of the following is/are measures of central tendency?
 - A. Mean.
 - B. Median.
 - C. 50th centile.
 - D. Student's *t*.
 - E. Standard Deviation.
 - F. All of the above.
 - G. None of the above.

Item Type #4 – Checklist

☐ Scoring:

- Each option can be score individually.
 - ☐ 1 for keyed (correct) answer.
 - ☐ 0 for non-keyed (incorrect) answer.
- The item can be scored by the sum of the correct answers.

☐ Problems:

- Each option allows a guessing strategy.
- Examinees commonly find this type of item “unfair.”

Item Type #4 – Checklist

- In the case of non-cognitive items, checklists can be more appropriate.
- Example:
 - In which of the following activities have you participated?
 - A. Signed a petition.
 - B. Written my Congressperson.
 - C. Written the President.
 - D. Participated in a legal demonstration.
 - E. Participated in an illegal demonstration.
 - F. Been arrested for demonstrating.
- A simple count could suffice for a score.
 - Although we may want a differential weighting of the items, it is unlikely we may do better than the simple count.

Item Type #5 – Matching

- ❑ In a matching item, stimuli in one list are matched to stimuli in another list.
- ❑ Example:
 - For each person named in the left column, record the letter from the right column that identifies the field with which the name is associated.

Person	Field
____ 1. Marie Curie	A. Linguistics
____ 2. Margaret Mead	B. Psychiatry
____ 3. Sigmund Freud	C. Anthropology
____ 4. Noam Chomsky	D. Paleontology
	E. Physics

Item Type #5 – Matching

□ Scoring:

- Item score is number of correct matches.
- Can also subtract number of incorrect matches.

□ Problems:

- Not certain how such an item could be applied to non-cognitive items.

Item Type #6 – Ordered-Category

- An ordered-category (or graded-response) item is an item where there are at least three response options (categories) that have a natural order.
- Example:
 - Assault weapons do not belong in private hands.
 - Strongly Agree.
 - Agree.
 - Neither agree nor disagree.
 - Disagree.
 - Strongly Disagree.

Item Type #6 – Ordered-Category

- Scoring:
 - Integer assigned to category (most frequently used).
 - Other variations of options include a numeric or graphic rating.
- Note: such items are frequently referred to as “Likert” items (pronounced Lick-ert).
- A true Likert scale is not an integer-valued set of responses (Likert, 1932).
 - We have come to call these Likert items because in his 1932 paper, Likert showed that a very sophisticated scaling system was not a clear improvement over integer-valued ratings.

Item Type #7 – Forced-Choice

- A forced-choice item forces the respondent to prefer one of two options.
- Example:
 - I prefer:
 - Strawberry ice cream.
 - Vanilla ice cream.

Item Type #7 – Forced-Choice

□ Scoring:

- A very complex problem (particularly when intransitivities are present).
- Not a focus of this course (requires scaling theory – briefly touched on in Chapter 18).

□ Benefits:

- May avoid halo effects.
 - Halo effect is a term used to describe a tendency for judgments of one rated characteristic to influence judgments of other characteristics in a positive or negative direction.

Item Type #8 - Rankings

- In a rankings item, respondents order a set of stimuli on an attribute.
- Example:
 - Rank the following ice creams according to your order of preference, using 1 for your “most preferred” and 4 for your “least preferred”
 - Vanilla.
 - Strawberry.
 - Butter Pecan.
 - Chocolate.



Item Type #8 - Rankings

- Scoring:
 - Very difficult.
 - Again, requires scaling theory.

Item Type #9 – Unordered Categorical

- In an unordered categorical item, respondents make a single choice from a set of unordered options.
- Example:
 - Which of the following colors do you like the best?
 - A. Red.
 - B. Yellow.
 - C. Blue.
 - D. Green.



Item Type #9 – Unordered Categorical

□ Scoring:

- No clear way to do so.
- Sometimes can be accomplished via optimal scaling techniques.



Guidelines for Item Writing

- Item stems should :
 - Be short.
 - Be unambiguous.
 - Be definite (not general).
 - Be simple (one proposition).
 - Avoid negatives.
 - Use simple vocabulary.
 - Use conventional vocabulary.

Item Scores

- ❑ A **binary item score** is an item score taking two scoring values only (usually zero and unity).
- ❑ A **quantitative item score** is an item score that takes more than two values.
 - Usually never have enough values to be treated as a continuous variable.
 - This is usually approximated by a normal distribution.

Concepts

- Two or more items taken together form a **test**.
- We expect to combine the item scores in some reasonable way to form a **test score**.
- The test score is generally intended to measure a **psychological attribute**.
 - Such as an ability, a personality trait, an emotional state, or an attitude.

More Concepts

- The sets of items forming a test can sometimes comprise **subsets**.
 - Measuring distinct attributes.
- Each subset forms a **subtest**.
 - The subtest item scores can be combined to provide a subtest score.
- One form of a subtest is a **item bundle** or **testlet**.
 - A set of questions about a common object, such as a paragraph requiring interpretation.

More Concepts

- ❑ A number of tests, called a **test battery**, may be administered to an examinee.
- ❑ A **profile** of test scores is a list of test scores from each test.
- ❑ Generally, the item is the ultimate atomic subtest (atomic is Greek for uncuttable).

Example Test

1. I am satisfied with my life.

- ☐ Strongly agree.
- ☐ Agree.
- ☐ Slightly agree.
- ☐ Neither agree nor disagree.
- ☐ Slightly disagree.
- ☐ Disagree.
- ☐ Strongly disagree.

Example Test

2. The conditions of my life are excellent.

- ☐ Strongly agree.
- ☐ Agree.
- ☐ Slightly agree.
- ☐ Neither agree nor disagree.
- ☐ Slightly disagree.
- ☐ Disagree.
- ☐ Strongly disagree.

Example Test

3. In most ways, my life is close to the ideal.

- ☐ Strongly agree.
- ☐ Agree.
- ☐ Slightly agree.
- ☐ Neither agree nor disagree.
- ☐ Slightly disagree.
- ☐ Disagree.
- ☐ Strongly disagree.

Example Test

4. So far I have gotten the important things I want from life.

- ☐ Strongly agree.
- ☐ Agree.
- ☐ Slightly agree.
- ☐ Neither agree nor disagree.
- ☐ Slightly disagree.
- ☐ Disagree.
- ☐ Strongly disagree.

Example Test

5. If I could live my life over, I would change almost nothing.

- ☐ Strongly agree.
- ☐ Agree.
- ☐ Slightly agree.
- ☐ Neither agree nor disagree.
- ☐ Slightly disagree.
- ☐ Disagree.
- ☐ Strongly disagree.



Example Test Distinctions

- By looking at the items, we can surmise that items 1, 2, and 3 represent a subtest measuring satisfaction with the present.
- Items 4 and 5 represent a subtest measuring satisfaction with the past.
- Without any empirical evidence we would conclude these two subtests are measuring different things.

Test Scores

- ❑ Assume test homogeneity – all items measure the same thing.
- ❑ Are used as a measure of an attribute that possesses some generality beyond that of a single item.
- ❑ From our example, it appears that the sum of all the items would not be feasible because the items are not homogeneous.

Wrapping Up...

- Today we discuss nine different types of commonly used items.
 - Some were easily scorable.
 - Some were not easily scorable.
- The definitions used today will follow us throughout the course as we evaluate items and tests.



Next Time

- Item statistics.
- Test statistics.