Measurement
Types of Instruments

Y520
Strategies for Educational Inquiry

Robert S Michael

First Questions

■ What is the stated purpose of the instrument? Is it to:
  • Collect information?
  • Predict a specified performance (e.g., The purpose of the SAT is to predict first college semester GPA).
  • Measure achievement?
  • Measure personality?
  • Measure attitude toward [an attitudinal object]?
■ The adequacy of a test can be judged only if you know its stated purpose.
Background Terms & Ideas

- **Instruments**: Collections of items or questions intended to measure, collectively (total score) levels of theoretical variables (constructs) not easily observable directly. Examples:
  - intelligence
  - optimism
  - quality of life
  - job satisfaction
  - depression
  - anxiety
  - political orientation

Instruments vs Indices

- **Instrument** (aka, test, scale): Type of response to items on an instrument are presumed to be caused by the underlying construct you are trying to measure.
- **Index**: In contrast, an index consists of items that *define* the construct.

Instrument example: Given a scale that measures optimism, we believe that more positive responses (i.e., higher total scores) are caused by someone who has higher levels of optimism.
Instruments vs Indices

- Index example: A socioeconomic status (SES) index might include a measurement of education level. But having more education is not caused by higher SES; instead having more education influences SES.

- An SES index measures SES, by definition. No attempt is made to measure a latent construct.

Three types of Instruments

- **Questionnaires.** Goal: Collection of information. No attempt to measure underlying latent construct.
  - Teacher Turnover Survey
  - Early Literacy Intervention Grant Prog. Survey

- **Cognitive domain** instruments. Goal: Measure achievement or aptitude.

- **Affective domain** instruments. Goal: Measure personality constructs, attitudes, beliefs, interests, etc.
Questionnaire Development Guidelines

- Clear, concise, unambiguous statements.
- Be aware of respondent reading levels.
- Pilot test, pilot test, pilot test.
- Frary, Robert (nd). A brief guide to questionnaire development. [Retrieve from http://ericae.net/ft/tamu/vpiques3.htm]

Example of Wording

- When presented with difficult professional situations where a superior censures you for an act for which you are not responsible, how frequently do you respond in an assertive way?
  - All of the time
  - Some of the time
  - None of the time

- Notice the vocabulary level and sentence complexity.
Alternate Wording

- When your boss blames you for something you did not do, how often do you stick up for yourself?
  - All of the time
  - Some of the time
  - None of the time

Which wording is preferable? Why?

Example 2: Alternate Wording

- During the past 4 weeks, how often have you felt a sense of loneliness?
  - All of the time
  - Sometimes
  - From time to time
  - Never
Example 2: Alternate Wording

- How often in the past month have you felt alone in the world?
  - Every day
  - Some days
  - Occasionally
  - Never

- Is this an improvement? Might some problems still exist?

Steps in Constructing an Instrument

- Determine clearly what it is you want to measure.
- Generate a pool of possible items (about 60 for a 20 item instrument).
- Decide on question format (yes/no, strongly agree, agree, etc)
- Decide how to score the instrument (e.g. on an optimism scale, all responses that indicate optimism should get higher scores, so that the more optimistic you are, the higher your total score).
- Administer the instrument to a pilot sample of people: 100 - 300.
- Scale purification: Evaluate each item using the pilot data:
  - The goal is to select the smallest possible subset of items that is internally consistent. This reduces respondent burden (few questions) and high reliability (internal consistency). Use Reliability procedure in Spss.
  - Calculate Cronbach’s alpha for various subsets of items and calculate the “item-remainder” coefficients (the correlation of an item with the sum of the remaining items).
  - Drop items that lower Cronbach’s alpha
  - If alpha remains low, use Spearman Brown Prophecy formula to tell you how many more items you need.
Constructing Instruments

- Preceding steps work well for questionnaires, attitude measures, personality scales.
- Some what different procedures are needed for cognitive domain instruments.

Cognitive Domain Instruments

- Achievement Instruments
  - Performance limited to specific domain (e.g., music)
  - Classroom Tests (teacher developed)
  - “High stakes” tests
    - Iowa Test of Basic Skills
    - ISTEP

- Aptitude Instruments
  - Performance “generalized”
Achievement Tests

- Classroom tests: measure student achievement in spelling, math, English, history, science, etc.
- The latent construct is “achievement”
- The operational definition is “number correct on a particular test.”
- Items on a test are a sample from all possible questions that could be asked from a particular academic domain.
- For test results to be interpretable, test (or subscale) should include items from only one domain.

Achievement Tests (continued)

- Classroom tests (continued): If a test contained both spelling and math items and number correct were summed, what would the score mean?
- Instead, sum the correct spelling items separately from the correct math items.
- Achievement tests can be either norm-based or criterion-based.
- Can be objectively scored items (true/false, multiple choice, matching, ordering) or performance assessments (aka, essay), or authentic assessments
Achievement Test Items

- Highly sophisticated guidelines for item writing exist. See [enter reference here]
- Items differ in the level of cognitive functioning required to answer them:
  - Knowledge
  - Comprehension
  - Application
  - Analysis
  - Synthesis/Evaluation

Bloom’s Taxonomy: Test Items

- Bloom’s Taxonomy describes different levels of cognitive functioning. Items can be classified depending on the level of cognitive functioning required to answer.
- Objectively scored items (e.g., multiple choice, true/false, matching, short answer) can be written to require either knowledge, comprehension, application, or analysis.
- Carefully constructed essay questions are used to require synthesis/evaluation answers.
Bloom’s Taxonomy: Levels of Knowledge

- **Knowledge**: Recognize or recall key facts, definitions, concepts, rules, principles. ("lowest level")
  - Key words: Define, identify, label, list, name
- **Comprehension**: Understanding of abstract terms. Recognize or explain similarities and differences.
  - Key words: Translate, Compare, contrast, relate, differentiate, distinguish.
- **Application**: Apply a concept to a concrete example.
  - Key words: Solve, apply,

---

Bloom’s Taxonomy: Levels of Knowledge

- **Analysis**: Divide a whole into its components.
  - Key words: Categorize, sort, subdivide, separate
- **Synthesis/Inference**: Recognize or explain the evidence relating to a generalization.
  - Key words: Infer, generalize, deduce, predict, speculate, apply, conclude
- **Evaluation**: Judge quality, credibility, worth or practicality, generally use established criteria and explain how those criteria are or are not met.
  - Key words: Evaluate, judge
Constructing a Classroom Test

- Table of Specifications
  - Crosses content area (rows) with level of cognitive functioning (columns).
  - Intersecting cells show either number or percent of items to be written on a particular subject and level of cognitive functioning.
  - Insures that test coverage reflects emphasis of topics in class and avoids including too many items that function at lower cognitive levels.

Table of Specifications: Example

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Knowledge</th>
<th>Comprehension</th>
<th>Application</th>
<th>Analysis</th>
<th>Synthesis</th>
<th>Evaluation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative Research</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research Traditions</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ethnography</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative Data Collection</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Qualitative Data Analysis</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
<td>12</td>
<td>0</td>
<td>16</td>
<td>5</td>
<td>2</td>
<td>50</td>
</tr>
</tbody>
</table>
Achievement Test Validity

- Validity: Does the instrument measure what we intend to measure (and what it was designed to measure)?
- “I am going to measure knowledge of research methods by placing each student’s term paper on a postage scale and recording the number of ounces.”
- The scale reading is reliable (able to be repeated) but ounces is not related to knowledge of research methods.
- However, if an instrument is valid, it must also be reliable.

Achievement Test Validity

- **Content Validity** is our concern for achievement tests.
- Validity is not “all or nothing.” Instead we judge based on the accumulation of pieces of evidence.
- Pieces of validity evidence:
  - Validity is “built in”
  - Table of specifications serves as “audit trail”
  - Items judged by experts as representing the domain
  - Performance of experts vs novices
  - Factor analysis for dimensionality
Achievement Test Reliability

- **Reliability:**
  - means consistency, stability, repeatability.
  - is established before validity.
  - sets an upper limit on validity.
  - always a correlation coefficient.
  - a test can be highly reliable but lack validity.
  - a valid test is also reliable.
  - sacrifice some reliability for validity.

Types of Achievement Test Reliability

- **Stability:** Test produces same results with repeated testing.
  - Test - retest
  - Parallel forms
  - Alternate (equivalent) forms

- **Internal consistency** or homogeneity: Degree to which all items in the test measure the same construct (achievement).
  - Item - total correlation
  - Split - half reliability
  - Cronbach’s Alpha
Types of Achievement Test Reliability

- **Equivalence**: Test produces the same result as does an equivalent instrument.
  - Parallel items on alternate forms

Indicators of Item Performance

- **Item Difficulty**: Percent of test takers who answered item correctly.
  - Too easy and too difficult items provide no information. “Ideal” difficulty depends on item type.

- **Item Discrimination**:
  - Identify 27% of test takers with the lowest total test score; likewise, top 27%.
  - For each item (proportion of top who got item right) – (proportion bottom who got item right).
  - Good items: few in the bottom answer correctly; many in the top group do.