Item-Level Development

PSY 395 – Oswald

Outline

- Constructing questionnaires (review)
- Item Content
- Item Formation
- Item Sequencing
- Item Statistics
- Uses for Item Statistics

Constructing Questionnaires

Given a research question and related theory…
- Determine content of the questionnaire
- Determine relative importance of the content areas
  - Include numbers of items in proportion to the relative importance.
- Determine whether measures exist for the content or need to be constructed
Constructing Questionnaires

- Write items
  - Open or closed items?
- Sequencing of items on the questionnaire
- Expert review (for knowledge, for bias)
- Pre-testing - pilot testing
  - Representative sample from population
  - Item analysis (discuss later)
  - Timing
  - Item scoring issues

Constructing Questionnaires

- Revise questionnaire based on pretest
- Develop guidelines for administration and interpretation of results
  - Train individuals who will administer the questionnaire
- Administer questionnaire
  - As standardized as possible
  - Participants should know what you are looking for
- Analyze results and construct a report on the results

Constructing Questionnaires

Example: College Board

- Develop Situational Judgment Test
- Based on dimensions of college performance
  - Knowledge
  - Leadership
  - Interpersonal Skills
  - Ethics
  - Career Orientation
  - Psychological and Physical Health
Constructing Questionnaires
Example: College Board

Sample SJT Item for Leadership
You are assigned to a group to work on a particular project. When you sit down together as a group, no one says anything.

a) Look at them until someone eventually says something
b) Start the conversation yourself by introducing yourself
c) Get to know everyone first and see what they are thinking about the project to make sure the project’s goals are clear to everyone
d) Try to start working on the project by asking everyone’s opinion about the nature of the project
e) You would take the leadership role by assigning people to do things or asking questions to get things rolling

What would you be most likely to do?
What would you be least likely to do?

Constructing Questionnaires
Example: College Board

- Based on dimensions, have one group of students generate items
- See whether items can be sorted back into the dimensions they are supposed to come from
- Have another group of students generate responses to the items that you keep
- Pilot test the items (evaluate content, reliability)
- Edit and revise the items

Item Content

- Facts (e.g., knowledge/ability tests)
  - Get right answers from experts or sources on the topic
  - Be careful…facts change!
  - Recognize that individuals may not be able to provide objective/correct statements about their own or others’ behaviors
Item Content

- Attitudes or beliefs
  - No right answer—can’t score as correct/incorrect
  - Assess the extent or magnitude of the attitude
    - Strongly agree to strongly disagree
    - Is there a neutral point?
  - Should try to separate two aspects of attitudes
    - Strength of attitude ("How strongly do you feel about...")
    - Direction of attitude

- Attitudes or beliefs (cont')
  - Wording of attitude items is absolutely critical
    - Subtle differences can result in very different responses (see book example on abortion)
    - Example “When is the death sentence justified?”
      - Answer yes/no
        - Version 1
          1. “If they commit murder.”
          2. “If they commit a treasonous act that threatens US security.”
        - Version 2
          1. “If they commit a treasonous act that threatens US security.”
          2. “If they commit murder.”
  - Can rarely be assessed with a single item
    - Multi-faceted, complex (e.g., abortion attitudes)

- Behavior
  - Good if asking close in time to the behavior
    (1 month?)
  - Memory problems limit the usefulness of this for distant memories
  - Frequency estimation of behavior is very difficult
  - In the past week have you… (Been to the museum? Chewed gum? Exercised? Gotten angry at someone?)
Item Formation

- Easy to read
  - Concise
  - No unnecessarily difficult words
  - Detailed enough to get across the purpose of the question
- One clear theme…
  - no double barrel questions
- Important information first
- Don’t use language or forms that suggest a right answer (e.g., inflammatory words, big business)

Item Formation

- The response options are just as important as the question
  - Don’t overly restrict responses
  - Don’t use inflammatory words or phrases in the response.
  - Easy to read
  - Allow a no opinion response if possible.

Item Sequencing

- The order of items matter!
  - Demographics last generally works best
  - Easy to answer at the beginning…builds efficacy
  - Keep similar items together – nothing tricky
    - Funnel principle
  - Tell them what you are trying to get from them…what’s the theme
  - Make it clear when you are moving to a different set of items.
  - Include random or careless response detection items
Item Statistics

- **Item Difficulty**
  - Proportion of people getting an item correct (the average for the item if it’s scored 1=correct, 0=incorrect)
  - (1) “The capital of France is (Paris).” $\Rightarrow p = .88$ (not difficult)
  - (2) “Paris is on the (Seine) River.” $\Rightarrow p = .23$ (difficult)
  - Therefore higher difficulty values mean easier items.

- **In Lab: Scoring Office Difficulty**
  - $(1 - p) \times 100$
  - $(1 - p)$ is what % got the item incorrect, x 100
  - Higher Scoring Office Difficulty means more difficult items.
  - “The capital of France is (Paris).” $\Rightarrow$ SOD = 12 ($p = .88$)

- **Item Discrimination**
  - Does the item response tell you about the person’s ability?
  - If everyone gets an item correct, that is a ceiling effect, and you can’t tell the difference in people’s ability levels.
  - If no one gets an item correct, that is a floor effect, and you also can’t tell the difference in people’s ability levels.
Item Statistics

- Item Discrimination
  - Item discrimination tends to be maximized when $p = .50$, in other words when half the people get an item correct and half get it wrong. This is when you have the best chance of knowing who is of higher ability and who is of lower ability.
  - Why is this so? Consider the extremes ($p$ near .00 or $p$ near 1.00)

Item Statistics

- Item Discrimination (cont’)
  - Say you had a customer-service test given to employees at Wal-Bart.
  - Take the top 1/3 and bottom 1/3 of scorers on the test.
  - Now look at the test items.
    - Sample item: “Wal-Bart offers ______ prices every day.”
      - (a) low, (b) decent, (c) exclusive, (d) outrageous
  - How might the top 1/3 do on that item compared with the bottom 1/3?

Item Statistics

- Item Discrimination (cont’)
  - Sample item:
    “Wal-Bart offers ______ prices every day.”
    - (a) low, (b) decent, (c) exclusive, (d) outrageous
  - Top 1/3 $\rightarrow$ $p = .98$, Bottom 1/3 $\rightarrow$ $p = .40$
  - The greater the differences between the high-scorers and low-scorers on an item, the higher the item discrimination.
The previous example split the sample into high and low scoring groups.

What if you didn’t split the sample into groups but instead correlated the response to that item (1=correct, 0=incorrect) with their total score on customer service (high scores = high knowledge, low scores = low knowledge)

…What would you expect?

The correlation between a dichotomous variable (the item scored 1/0) and a continuous variable (the scale score goes from 0-50 and is fairly continuous) as we have here is called a point-biserial correlation or an item-total correlation.

This correlation is one measure of item discrimination (what would the correlation be if everyone got an item correct, or if everyone got an item incorrect)?

This correlation ranges from 0 to 1 and has the potential to be highest when item difficulty is \( p = .50 \).
Uses for Item Statistics

- **Item Refinement (edit items, write new ones)**
  - Are items too difficult/too easy?
  - Do items predict the total score? (Does the item contribute to alpha?)

- **Theory Refinement**
  - Are the criterion-related validities relatively strong where you expect them to be (convergent validity)?
  - Are they weak/zero where you expect them to be (discriminant validity)?