Week 2 Lecture Outline
Test Development Process

Context

- The challenge here is that we are looking at an overview of a process that involves activities that we have not yet looked at in detail.
- Think of this chapter as scaffolding for future chapters.

Steps

Here is the overview from the chapter:

- 1. State the purpose of the scale
- 2. Define the domain of the construct to be measured
- 3. Determine whether a measure already exists
- 4. Determine the item format
- 5. Develop a test blueprint or test objectives
- 6. Create the initial item pool
- 7. Conduct initial item review (and revisions)
- 8. Conduct large-scale field test of items
- 9. Analyze items
- 10. Revise items
- 11. Calculate reliability
- 12. Conduct second field test of items
- 13. Repeat steps 8 – 11 as necessary
- 14. Conduct validation studies
- 15. Prepare guidelines for administration

Knocking: A toy example

1. State the purpose of the scale	To determine whether someone is inside his or her office.
2. Define the domain of the construct to be measured	Construct: physical presence within one’s office. Domain: Successful response to audible door knock when present (no response when absent). E.g., responding verbally or opening the door.
3. Determine whether a measure already exists	Yes, it exists but this is just an example.
4. Determine the item format	Knock on central region of outer surface of door with sufficient force to be audible inside.
5. Develop a test blueprint or test objectives	Knock at horizontal center of outer surface of door between 3 and 6 feet from the ground. Loudness between 60 and 80 decibels. (Could sample from a 4-by-3 table of height by decibels.) Knock on unobstructed portion of door.

6. Create the initial item pool	E.g., <3ft, 60db>, <4ft, 80db>, <5ft, 70db>, <6ft, 90db>
7. Conduct initial item review (and revisions)	Oops, 90db is too loud.
8. Conduct large-scale field test of items	Knock on 1000 office doors. (One item per office.) Record responses.
9. Analyze items	Frequency distribution of responses (Yes/No).
10. Revise items	Any items with very high or very low response frequencies?
11. Calculate reliability	This example may require a separate study for test-retest reliability.
12. Conduct second field test of items 13. Repeat steps 8 – 11 as necessary	

14. Conduct validation studies	Crosstabulate responses with independent measures of presence in office (e.g., shout through door, peek under door). Conduct cognitive interviews about further constraints on appropriate use of knocking.
15. Prepare guidelines for administration	Test manual for door knocks. How to knock properly. How to interpret the results. Normative data. Summary of validity and reliability evidence.

Transitions between steps

- We will look at most steps in more detail later in the term.
- So, for the moment, let's focus on the connections between the steps and the logic of their order.
- To facilitate this, let's focus on the transitions from one step to the next.

1. State the purpose of the scale
2. Define the domain of the construct to be measured

- Even seasoned test developers report difficulty defining domains and constructs.
- Stating the purpose first helps focus the problem and make it more concrete.
- A domain and construct are not necessarily the same thing.
  - Nor are they exclusive of one another.
  - You can think of the construct as determining a corresponding domain.
- The purpose can help determine which you want to emphasize.
  - E.g., if the purpose is to assess competency for some scope of material, then emphasizing the domain may fit well.
  - If the purpose is to assess a dimension without a well-defined domain, then emphasizing the construct may work better.

2. Define the domain of the construct to be measured 3. Determine whether a measure already exists

- You want to be thinking about existing tests from the start.
- However, it is hard to identify and existing test until you pin down the purpose and domain/construct.
  - It is not uncommon for two tests to share a construct but differ in construct labels.
  - *Jingle Fallacy*: Assuming that the same name ensures same referent.
  - Likewise, tests can share the same construct label but differ in constructs.
  - *Jangle Fallacy*: Assuming that referents must differ because names do.
- Even if there is an existing alternative, you may still decided that you want to develop something that would work better.
- Return to Step 3 throughout the development process whenever new developments lead to revisions in earlier steps.

3. Determine whether a measure already exists
4. Determine the item format

- If form follows function, the same can be said of format.
- The purpose and domain/construct should influence the choice of format.
- You may also learn from competing tests (both from their successes and their mistakes).
- E.g, the purpose includes the intended test taker population.
  - Different formats may work better for different populations (e.g., literacy, cultural expectations, familiarity)

4. Determine the item format
5. Develop a test blueprint or test objectives

- This transition is particularly prone to cycling back and forth.
- Working out the blueprint may stimulate reconsideration of formats.
- E.g., The LSAT exam reflects a small number of item formats out of a much larger pool of formats developed and evaluated for the test.
- Use the format choices to guide the initial blueprint....
- ...but also use the blueprint to clarify and evaluate the choices of format.

5. Develop a test blueprint or test objectives
6. Create the initial item pool

- The blueprint provides the specifications for the item pool.
- If you encounter problems in the item writing process, consider revising or refining your blueprint (especially item specifications).
- It can be tempting to skip the item pool step and just draft the test directly.
  - Resist that temptation.
  - It can be very hard to predict which items will work best.
  - If you only draft what you need, you will never know how your items compare.
- For domain based tests, creating an item pool can test your understanding of the domain, possibly inviting revisions.
- So, even unused items contribute to test development.
  - They do not represent wasted effort.

6. Create the initial item pool
7. Conduct initial item review (and revisions)

- If you skimp on the item pool, you run the risk of Step 7 forcing you back to Step 6.
- If you have a large pool, it has a better chance of getting you through Step 7.
- However, there is nothing wrong with writing more items after Step 7 if that seems valuable.
- Reviewing the items from the pool will also help develop the understanding of the specifications, ensuring common understanding.
- As such, it can be useful to have the item writers involved in item review as well.

7. Conduct initial item review (and revisions)
8. Conduct large-scale field test of items

- Field tests are expensive.
- Careful review at Step 7 avoids wasted resources in Step 8.
- Step 7 can also include some small scale item testing (preliminary item tryouts).
  - These might also occur in earlier stages, such as experimental item formats.
- The goal is to invest time and effort before the first large-scale field test in order to maximize its value and minimize the need for more field tests.
- Also, use the earlier steps including item review to formulate research questions for the field test.

8. Conduct large-scale field test of items 9. Analyze items

- This pair represents the standard relationship between research design and data analysis.
- The design must collect data to support the analyses.
- The analyses must answer the questions that guided the research design.
- Different field tests may focus on different questions and different analyses, even for the same test.
- Earlier field test may focus more on item analysis whereas later field tests focus more holistically on the internal structure of the test.

9. Analyze items
10. Revise items
11. Calculate reliability

- Including the whole pool rather than just a draft test lets you evaluate items in the context of the other items.
  - E.g., what is the range of item difficulty?
  - what s the range of item intercorrelations?
- It also allows you to choose the best items from the pool based on the field test data.
- The analyses need to provide adequate information to guide item revision.
- The more you learn from the field test, the more you are able to strengthen the pool.
- Systematic item revisions may require revisiting the domain/construct, blueprint, or item specifications.
- Reliability estimation is really a parallel process that also follows Step 9.
  - Step 10 focuses on the item level.
  - Step 11 focuses on the test level.

10. Revise items
11. Calculate reliability
12. Conduct second field test of items
13. Repeat steps 8 – 11 as necessary

- Item revisions are designed to fix problems.
- Subsequent field test are needed to confirm that the problems have been satisfactorily fixed.
- This often involves successive approximations.
- It is also possible that revisions aimed at fixing one thing can create a problem someplace else.
  - E.g., revisions to adjust item difficulty might inadvertently reduce reliability.
- The blueprint includes specifications for acceptable item and test functioning.
- The process stops when the test meets those specifications.

13. Repeat steps 8 – 11 as necessary 14. Conduct validation studies

- If you have not thought about validity until this point, you are in trouble.
  - Defining the construct, developing the blueprint, and choices about item design and revision all involve validity.
  - Documentation from the beginning of the test development process provides important validity evidence.
- The test should be in nearly final form before conducting validation research.
  - Should study test scores from the proposed test, not just the item pool.
- Nonetheless, validation evidence can lead to a return to earlier steps in order to revise the test to improve validity.

14. Conduct validation studies
15. Prepare guidelines for administration

- The test manual will normally contain a summary of reliability and validity evidence.
- In order to produce the kinds of scores that you have validated, test users must use the test the same way.
- Validation can include research on the interaction between the test and test users (e.g., varying in experience or training).
- Validation can also include investigation of systematic variation in test administration.
  - E.g., guidelines for test accommodations.