Module 3: Validity and Reliability
Ensuring the Rigor of Your Assessment Study
January, 2014

Jeremy Penn, Ph.D.
Director

DIVISION OF STUDENT AFFAIRS

Follow-through
1. Asking a Question (Purpose)
2. Designing an Assessment Study
3. Assessment Instruments (validity and reliability)
4. Analyzing Results
5. Using and Communicating Results
6. Celebrating results (follow-through and reflection)

Plan

Implement

Resources
• Schuh, chapter 5: Instrumentation
What will we do?

- Introduction to issues of reliability and validity
  - Choose follow-up session on instrument(s) of choice
- In module 3, each team will need to address:
  - What specific assessment instruments will be used
    - Will instrument be adapted in some way? Created new? Purchased?
  - A final copy of the instrument should be included
  - How reliability and validity (quant study) or credibility and trustworthiness (qual study) will be supported
    - Inferences the team hopes to make from the data and how those inferences will be supported
    - Will anyone believe your inferences?

The Importance of Rigor

Father Sarducci’s “Five Minute University”
http://www.youtube.com/watch?v=kO8x8eoU3L4
“strategic student affairs professionals need to evaluate – and just as important, need to effectively articulate – the contributions of the profession and those in it. However, they need to do so not simply through their own eyes, but through objective measures and through the eyes of their colleagues within the colleges and universities in which they work and the eyes of the external constituencies they ultimately serve” (Bresciani, 3)

Student Centered Division

• Important to draw accurate and meaningful conclusions about our programs and services
• Good stewards of the funds we are provided
• Perception of the public and others outside the division regarding the value and worth of what we do – credibility!
Establishing Rigor

- Although a bit different in qualitative and quantitative studies, concept is the same
- No checklist or one-size-fits-all approach
- The more severe the consequences for the planned use, the more evidence needed for rigor

Professional Judgment is Critical

- Work with your colleagues (or with me) on these issues
- Consulting additional resources helpful
  - *Standards for Educational and Psychological Testing*
  - Creswell’s books on qualitative research methods
    • Can borrow these materials from me

Quantitative Studies

- Closed-ended studies focusing on cause and effect, use of theory, use numbers (most assessment studies are QUAN)
- Reliability and validity
  - Reliability necessary, but not sufficient for establishing validity
Reliability

“Consistency”
Could be consistently good…

...or consistently bad

Reliability

“Reliability refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups.” (Standards, 1999)

• Multiple items
• Multiple raters
• Multiple forms
• Multiple administrations
Reliability

- Careful consideration should be given to the training of reviewers or scorers
- Inter-rater reliability should be examined (this is another presentation) and quality control processes implemented as part of the scoring process
  - (E.g., periodically checking that scoring is consistent between raters)

Reliability

Efforts taken to control error in instrument design
- Items not prone to multiple interpretations
  - Football vs. "Football"
- Carefully proofread
- Similar instructions given to all students
- Equal difficulty of multiple forms of the same instrument

Reliability

Careful thought given to the timing of the administration of an instrument
- Would students respond in a similar manner if the instrument was administered on a different date?
- Are students stressed out due to finals week? Or relaxed because it is spring break?
Issues in Reliability

FAIL: Staff team evaluates students using the same rubric.

FIX: Staff are trained on the rubric and practice so they will score students consistently.

Validity

“the degree to which evidence and theory support the interpretations of test scores entailed by proposed use of tests” (Standards, 1999, p. 9)

“It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself” (Standards, 1999, p. 9)

Example

• Use ACT score to predict which students are likely to be successful in college
  – Substantial evidence exists from research to support the use of scores in this manner
• Use ACT score to identify which students should become members of the marching band
  – No evidence this is a reasonable use of ACT scores
Supporting your Interpretations and Inferences

• Start with instrument itself
  – Does the content match the domain / construct of interest?
    • Depth (not just a shallow connection) and breadth (all important aspects of the domain / construct are included)
  – Does the response process match the desired process
    • Working “backward” on a math problem is not a match

• Consider the technical quality of the instrument
  – If it is a commercially published instrument, look for reviews (such as on www.buros.org) and read the technical manual
  – Look for information regarding recommended uses and population of individuals with whom the instrument was tested

• Consider a pilot study for your instrument
  • Show your instrument to a representative from the target group and have the person respond to the items out loud (what does this question mean to you? How would you respond? What are you considering when you answer this question?)
  • Have experts review your instrument
Supporting your Interpretations and Inferences

• Use multiple measures or sources of information
  – Instruments that claim to measure the same thing should be highly correlated
  • Perceptions of learning should correlate with learning, although not as strongly and perceptions are often not as accurate
  – Don’t want your instruments correlated with constructs of non-interest (such as leadership ability correlated with knowledge test on the history of Major League Baseball)

Supporting your Interpretations and Inferences

• Be thoughtful regarding the consequences of the planned use of your results (intended and unintended)
  – The higher the consequences, the more evidence is needed for your validity argument
  – Assessment can have bigger consequences than you think:
    • Adjusting a program before it is offered again: fairly benign
    • Ending a program, evaluating a staff member, substantially changing how resources are allocated: more severe consequences (if you are wrong!)

Validity - summary

• When selecting an instrument, must consider the possible uses of scores and how scores will be interpreted
• When sharing results, consider how different audiences may misinterpret findings
• Should gather evidence from other sources (multiple measures) to support interpretations
Validity - summary

- Clearly identify the construct (or constructs) the instrument is intended to measure.
- Don’t assume that you can “purchase” validity and reliability from a commercial vendor—May have different interpretations than your intended use.
- The higher the stakes, the more important that inferences are supported with strong evidence of technical quality (e.g., using a test score for entrance into a degree program).

Issues in Validity

FAIL: Assessment team selects a standardized instrument from a commercial vendor because it is inexpensive and used by many other institutions.

FIX: The instrument is selected because it is a good match for what the team wants to measure and has evidence to support the inferences the team wants to draw.

Activity

Issues in reliability and validity (handout)
- With members of your team, discuss the assigned scenario. Identify the issue and recommend a fix.
Qualitative Studies

• Use words, text, small numbers of participants in interviews, focus groups, discussion, often in a “natural” setting
• Purposeful sampling instead of random
• Authenticity, goodness, trustworthiness, credibility
  – Parallel but non-equivalent to validity / reliability in QUAN studies

“The basic issue in relation to trustworthiness is simple: How can an inquirer persuade his or her audiences (including self) that the findings of an inquiry are worth paying attention to, worth taking account of? What arguments can be mounted, what criteria invoked, what questions asked, that would be persuasive on this issue?”
(Lincoln and Guba, 1985, p. 290)

Lincoln and Guba

Identify four criteria:
• Credibility
  – “Truth value,” make sense to “constructors of multiple realities”
• Transferability
  – Knowing sending and receiving contexts (burden lies with person seeking to make application elsewhere)
• Dependability
  – Similar to reliability, or consistency
  – If similar study done under same circumstances, likely to find same results?

• Confirmability
  – Characteristic of the data
  – Member checking (“is this what you meant?”) and triangulation from multiple sources

Strategies
• Member checking (“is this what you meant?”)
• Triangulation (multiple sources of confirmation)
• Thick and rich description
• Clarifying the bias of the researcher
• Audit process
• Prolonged time in the field

Worksheets
Module 3 worksheet will get your team started on this portion of the task

Team presentations on this topic will be held in April
Choose
- 3a: Survey Methods
- 3b: Focus Groups
- 3c: Data Mining and Institutional Data
- 3d: Rubrics

Required
- At least 1 Team Work Time
- Module 3 Team Presentations
- Kendra will be contacting to schedule a date
- Scheduled work time sessions mid- to late-March
- Presentations mid-April

Presentations
- Each team will prepare a short (5-7 minute) presentation for the group scheduled tentatively for mid-April
- Please use PowerPoint