Plenary Session, TH July 9, 2009
Dr. Terry Ackerman

Sharing tools and resources
Dr. Ackerman is happy to share online item tool, powerpoints, other resources via sign-up sheet

Origins of measurement
Lambruso – measure human behavior (described in book by Gould)
Classify criminals based on physical characteristics

Reliability
The general concept of reliability was discussed

- Consistency is key – examinee performance, across test, within test items
- Population of examinees changes → may change reliability since the population have different skill sets they bring to the situation
- Quality of your data is related to this: no sense in doing statistical analysis if the test is not reliable.

Next, a few specific issues impacting reliability were discussed:

Errors in Measurement

- Deviation from true performance
  - Usually assumed that +/- devations average to zero
  - Usually assumed errors are random
- DIF analysis – differential item functioning – a standard method, but need large populations to do this, of finding out of people perform differentially (say male vs. female have different performance). Goal of this → make sure tests are fair

True Scores

- % of correct answers from test of the entire universe
- Mean of infinite number of tests taken by same person is their true score

Observed score = true +error due to measurement

Item Response Theory

- Difficulty – need about 200 examinees to get good estimates of item difficulty so reliably measure performance.
- For an individual, a 50/50 chance of getting it correct would be the most reliable item for that individual

Raters of observed performances

- Scorer reliability – can the rater produce ratings of the same performance consistently
- Interrater reliability – do different judges agree? Everbody who uses the rubric needs to have thorough training to understand the meaning of the rubric

Test-retest

- Is the ranking of examinees the same in multiple uses of the same test.
- This issue is difficult if effective teaching produces different amounts of learning in different individuals.

Parallel forms – do different tests on same area give same rankings? If you teach a course multiple times and get better at teaching, do the students scores increase?

Internal consistency – do all items measure the same skill or composite set of skills?

- Split half – correlate one half to other half of each test to see if give same score. Can split test in various ways (first half second half, or odd even are most common).
- The Spearman Brown coeff rates how reliable the two halves are.
- KR20 coeff is average of all possible ways to split test IF you have only right/wrong scoring
- Coefficient alpha finds average of Spearman Brown coeff for all possible ways to split test in half and can use with scaled responses.

Validity

Judgment about evidence supporting actions taken based on test scores.

Content – expert opinion that the items match the learning objectives of the test

Criterion related – use empirical data from the test to look at relationship between those scores and some criterion

Construct and face validity also important

Standardized test

ACT more what did your learn, your achievement. SAT is more about what will you be able to learn, aptitude

Types of Items

Multiple Choice – stem needs to be complete information to define the question, then answer and distractors are possible choices.

- Hard to write good distractors because they need to be plausible.
- Good distractors indicate what the students are missing
- Underline key words or phrases
- Write correct answer and then distractors
- Avoid using negatives unless can't avoid doing so in order to test the concept
- Avoid using all of the above (OK to use sparingly) because they conceptually load the answer

T/F – just multiple choices with two options

Essay items

Issues to consider

- Focus on importance of objective to be tested,
- Focus on essential knowledge applied to a new area,
- be clear (about level of detail expected, what is being tested, etc)

Develop a rubric to assign ratings categories

Different levels of cognition – an update to Bloom's by Krafwall …
Factual
Perceptual
Procedural
Metacognitive – example of how to address this level: use questions asking student to explain results to different audiences

They have mapped Bloom to their six levels
See verbs for each set to help you craft items addressing the different levels

Software
Breaks scores into upper, middle, lower groups
Measures items for point-biserial and biserial measures of discrimination
Shows inter-item correlation
Gives several overall test data (mean, std dev, KR20, etc)