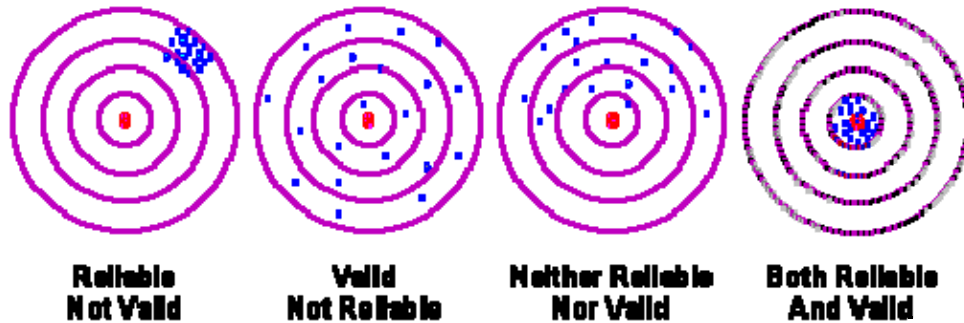


**How Do You Evaluate ACGME Competencies?
You Find a Valid, Reliable Instrument**
Su-Ting Li, MD, MPH and Daniel C. West, MD
University of California, Davis
APPD Workshop, May 2, 2008

Background Information for Small Group Exercises



Step I: Identify the Construct

- What area of knowledge, skill or behavior is being assessed?

Step II: Assess Importance

- Is what is being assessed important?

Step III: Assess validity:

- Does the assessment tool measure the characteristic, skill, or behavior that it is supposed to measure?
- Does it hit the target?

A. Internal validity:

1. *Face validity* – An individual’s perception of how valid the assessment is.
 - At first glance, does the instrument seem to assess what it is supposed to assess? (ie., if measuring math skills, does the instrument include math problems?)
 - Important for buy in or to get cooperation.
2. *Content validity* – Evidence (based on expert judgment, medical literature, etc) that items included in assessment match the competency to be assessed.
 - Does the instrument include all important aspects of the theoretical construct it is supposed to measure?
 - For example, if measuring math skills, does the instrument only measure addition, but does not measure other “math skills” like multiplication, etc.
3. *Criterion validity* – Is the instrument’s measurement consistent with a criterion-standard, or “gold-standard”?
 - If no gold-standard, then should use a combination of measures of construct and criterion validity to build a case for validity
 - a. *Construct validity* – Is there agreement between the theoretical construct and what you are measuring?

- i. Convergent validity – Are measurements of the same theoretical construct consistent?
 - If measuring addition skills, people who answer an addition problem correctly should tend to answer other addition problems correctly.
 - Do people who score high on your math assessment also score high on other math assessments thought to measure the same skills?
 - ii. Divergent validity – Are measurements of different theoretical constructs independent?
 - Whether a person answers addition problems correctly should be independent of whether they can correctly identify names of animals
- b. *Criterion validity* –
- i. Predictive validity: Does this measure predict performance in the future?
 - For example, do medical students who score high on USMLE Step 2 perform better as senior residents than students who score lower?
 - ii. Concurrent validity: Does this measure relate to current performance? Is performance consistent with other measures of the same theoretical construct?
 - For example, do more advanced learners score higher?
 - Do residents who score higher on an assessment of interpersonal skills completed by nurses also score higher on an assessment of interpersonal skills by patients and families?

B. External validity:

- How generalizable or transferable is the instrument to your situation?
- Did the study involve only a single institution or multiple?
- How much is the institution, residents, or others involved in the study similar to your situation?
- What is the evidence that populations that the assessment tools was test in is similar to the population in which you would use the tool?

Step IV: Assess Reliability

- Is the assessment dependable or reproducible?
- Does it hit the same spot on the target consistently?
- Example: If a washing machine runs every time you turn it on, it is reliable. If it gets clothes clean, it is valid.

Application of Measurement Theory:

- Goal is to measure behaviors or skills in people
- Measurement theory describes, categorizes, and assesses the quality of these measures
- All measurement tools have error
- 3 components of measurement theory.

A. Classical Test Theory:

General definition: Observation or measurement composed of two parts:

- The expected or *true score* (the person's real characteristic) and

- The difference between the true score and the measured score (that indicated by the instrument or tool)
- **Measured score = True score + Error** (can be + or -)

Appropriate use:

- Survey design: to ensure that questions are internally consistent
- Useful when you only want to understand how much error there is and do not care to understand where it coming from or how to modify your assessment to improve reliability.

Limitations:

- Measures only one source of error at a time
 - Does not allow measurement of multiple sources of error and their interactions
- Examinee and test characteristics cannot be separated
 - e.g. Did learners miss the question because the question was too hard or poorly written, or did they miss it because they just did not know the material?

Examples:

1. Test/retest reliability: When taking the same assessment, does the same person get the same score on two different occasions (assuming no learning or maturation between assessments)
2. Inter-rater reliability: Do multiple raters get the same results when rating the same resident at the same time?
3. Intra-rater reliability: Does the same rater get the same results when rating the same resident twice? How consistent are multiple ratings of the same resident under similar situations?
4. Internal consistency: How consistent are multiple ratings of the resident across different situations with the same theoretical construct?
 - a. Inter-item correlation; split-half reliability; Cronbach's alpha – How consistently do multiple items measure the same theoretical construct?

B. Generalizability Theory:

General definition: Statistical theory about dependability of behavioral measurement

- Accuracy of generalizing a person's observed score on a measurement to the average score that person would have achieved under all possible conditions of interest to the person administering the test.
- In other words, accuracy of generalizing a score over all possible occasions of testing, raters of performance, items that could be tested, or any other variable that could be a source of error (that the tester would care about).
- Concept of universe score
 - **Measured score = universe score + error**
 - *Key difference*: error can be decomposed, relative contributions of difference sources of error and their interactions to the total error can be determined.
- Allows for measurement of multiple sources of error and their interactions at the same time.
- Allows calculation of a generalizability coefficient, which is similar to reliability coefficient.

Appropriate use:

- Measurements of multiple sources of error allow for adjustments in assessment method to improve reliability.
- Likely, the most appropriate way to assess the reliability of most of the assessment tools a program director would want to use.

- Can run “what if” scenarios to figure out how to modify assessment in order to maximize reliability.
 - Can determine how many occasions to administer an assessment or how many raters are necessary in order to achieve adequate reliability.

Examples:

A. Structured Clinical Observation: Can use G-theory to determine how many different raters are necessary or on how many different occasions residents should be rated in order to achieve acceptable reliability.

B. 360 Degree Evaluation: Can use G-theory to determine how many different patients or nurses should evaluate a resident in order to have acceptable reliability to either rank order residents or determine whether they have achieved an adequate level of competency to be promoted.

C. Item Response Theory:

General definition: Modern approach to test construction

- Allows for determination of performance of individual test items, independent of examinee ability
- Allows for assessment of examinee performance, independent of test items, so called “sample-free” assessment (determine individual performance independent of what items are used in the assessment)

Appropriate use:

- Use to construct standardized written examinations
 - ABP and NBME uses this methodology

Limitations:

- Requires expertise not readily available
- Not useful for determining multiple sources of error

Examples:

A. High-Stakes Test: Creating new high stakes medical knowledge test that will be used to determine whether someone passes a rotation or gets to graduate from the training program.

Step V: Feasibility: Beware of making modifications—it could affect reliability or validity or both.

- A. Time – Time needed to implement?
- B. Training – Training of residents, faculty, others to implement?
- C. Equipment/technology – Special equipment/technology needed to implement?
- D. Cost – How expensive to implement?
- E. Other considerations