# A Primer on the Validity of Assessment Instruments

Gail M. Sullivan, MD, MPH

### 1. What is reliability?[1]

Reliability refers to whether an assessment instrument gives the same results each time it is used in the same setting with the same type of subjects. Reliability essentially means *consistent* or *dependable* results. Reliability is a part of the assessment of validity.

### 2. What is validity?[1]

Validity in research refers to how accurately a study answers the study question or the strength of the study conclusions. For outcome measures such as surveys or tests, validity refers to the *accuracy* of measurement. Here validity refers to how well the assessment tool actually measures the underlying outcome of interest. Validity is not a property of the tool itself, but rather of the interpretation or specific purpose of the assessment tool with particular settings and learners.

Assessment instruments must be both reliable and valid for study results to be credible. Thus, reliability and validity must be examined and reported, or references cited, for each assessment instrument used to measure study outcomes. Examples of assessments include resident feedback survey, course evaluation, written test, clinical simulation observer ratings, needs assessment survey, and teacher evaluation. Using an instrument with high reliability is not sufficient; other measures of validity are needed to establish the credibility of your study.

### 3. How is reliability measured?[2–4]

Reliability can be estimated in several ways; the method will depend upon the type of assessment instrument. Sometimes reliability is referred to as internal validity or internal structure of the assessment tool.

For *internal consistency* 2 to 3 questions or items are created that measure the same concept, and the difference among the answers is calculated. That is, the correlation among the answers is measured.

Cronbach alpha is a test of internal consistency and frequently used to calculate the correlation values among the answers on your assessment tool.[5] Cronbach alpha calculates correlation among all the variables, in every combination; a high reliability estimate should be as close to 1 as possible.

**Gail M. Sullivan, MD, MPH,** is Editor-in-Chief, *Journal of Graduate Medical Education.*

Corresponding author: Gail M. Sullivan, MD, MPH, Editor-in-Chief, *Journal of Graduate Medical Education*, 515 N State St, Suite 2000, gsullivan@nso1.uchc.edu

**DOI: 10.4300/JGME-D-11-00075.1**

For *test/retest* the test should give the same results each time, assuming there are no interval changes in what you are measuring, and they are often measured as correlation, with Pearson *r*.

Test/retest is a more conservative estimate of reliability than Cronbach alpha, but it takes at least 2 administrations of the tool, whereas Cronbach alpha can be calculated after a single administration. To perform a test/retest, you must be able to minimize or eliminate any change (ie, learning) in the condition you are measuring, between the 2 measurement times. Administer the assessment instrument at 2 separate times for each subject and calculate the correlation between the 2 different measurements.

*Interrater reliability* is used to study the effect of different raters or observers using the same tool and is generally estimated by percent agreement, kappa (for binary outcomes), or Kendall tau.

Another method uses analysis of variance (ANOVA) to generate a *generalizability coefficient,* to quantify how much measurement error can be attributed to each potential factor, such as different test items, subjects, raters, dates of administration, and so forth. This model looks at the overall reliability of the results.[6]

### 5. How is the validity of an assessment instrument determined?[4,7,8]

Validity of assessment instruments requires several sources of evidence to build the case that the instrument measures what it is supposed to measure.[9,10] Determining validity can be viewed as constructing an evidence-based argument regarding how well a tool measures what it is supposed to do. Evidence can be assembled to support, or not support, a specific use of the assessment tool. Evidence can be found in *content, response process, relationships to other variables,* and *consequences.*

*Content* includes a description of the steps used to develop the instrument. Provide information such as who created the instrument (national experts would confer greater validity than local experts, who in turn would have more validity than nonexperts) and other steps that support the instrument has the appropriate content.

*Response process* includes information about whether the actions or thoughts of the subjects actually match the test and also information regarding training for the raters/observers, instructions for the test-takers, instructions for scoring, and clarity of these materials.

*Relationship to other variables* includes correlation of the new assessment instrument results with other performance outcomes that would likely be the same. If

there is a previously accepted "gold standard" of measurement, correlate the instrument results to the subject's performance on the "gold standard." In many cases, no "gold standard" exists and comparison is made to other assessments that appear reasonable (eg, in-training examinations, objective structured clinical examinations, rotation "grades," similar surveys).

*Consequences* means that if there are pass/fail or cut-off performance scores, those grouped in each category tend to perform the same in other settings. Also, if lower performers receive additional training and their scores improve, this would add to the validity of the instrument.

Different types of instruments need an emphasis on different sources of validity evidence.[7] For example, for observer ratings of resident performance, interrater agreement may be key, whereas for a survey measuring resident stress, relationship to other variables may be more important. For a multiple choice examination, content and consequences may be essential sources of validity evidence. For high-stakes assessments (eg, board examinations), substantial evidence to support the case for validity will be required.[9]

There are also other types of validity evidence, which are not discussed here.

### 6. How can researchers enhance the validity of their assessment instruments?

First, do a literature search and use previously developed outcome measures. If the instrument must be modified for use with your subjects or setting, modify and describe how, in a transparent way. Include sufficient detail to allow readers to understand the potential limitations of this approach.

If no assessment instruments are available, use content experts to create your own and pilot the instrument prior to using it in your study. Test reliability and include as many sources of validity evidence as are possible in your paper. Discuss the limitations of this approach openly.

### 7. What are the expectations of *JGME* editors regarding assessment instruments used in graduate medical education research?

*JGME* editors expect that discussions of the validity of your assessment tools will be explicitly mentioned in your manuscript, in the methods section. If you are using a previously studied tool in the same setting, with the same subjects, and for the same purpose, citing the reference(s) is sufficient. Additional discussion about your adaptation is needed if you (1) have modified previously studied instruments; (2) are using the instrument for different settings, subjects, or purposes; or (3) are using different interpretation or cut-off points. Discuss whether the changes are likely to affect the reliability or validity of the instrument.

Researchers who create novel assessment instruments need to state the development process, reliability measures, pilot results, and any other information that may lend credibility to the use of homegrown instruments. Transparency enhances credibility.

In general, little information can be gleaned from single-site studies using untested assessment instruments; these studies are unlikely to be accepted for publication.

### 8. What are useful resources for reliability and validity of assessment instruments?

The references for this editorial are a good starting point.

**References**

1 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association; 1999.
2 Downing SM. Reliability: on the reproducibility of assessment data. *Med Educ.* 2004;38(9):1006–1012.
3 Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Manderkar JN. How reliable are assessments of clinical teaching?: a review of the published instruments. *J Gen Intern Med.* 2004;19(9):971–977.
4 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments. *Am J Med.* 2006;119(2):166e7–166e16.
5 Bland JM, Altman DG. Statistics notes: Cronbach's alpha. *BMJ.* 1997;314:572.
6 Brennan RL. *Generalizability Theory.* New York, NY: Springer-Verlag; 2001.
7 Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ.* 2003;37(9):830–837.
8 Downing SM, Haldyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ.* 2004;38(3):327–333.
9 Kane M. Validating high-stakes testing programs. *Educ Meas Issues Pract.* 2002;1:31–41.
10 Kane M. The assessment of professional competence. *Eval Health Prof.* 1992;15(2):163–182.

The following are corrections to the June 2011 issue.

1. Sullivan GM. A primer on the validity of assessment instruments. *J Grad Med Educ*. 2011;3(2):119–120.

On p 119, the sentence should read: Cronbach alpha calculates correlation among all the variables, in every combination, and generates one number that the closer it is to 1, the higher the reliability estimate.

2. Salem JK, Jones RR, Sweet DB, Hasan S, Torregosa-Arcay H, Clough L. Improving care in a resident practice for patients with diabetes. *J Grad Med Educ*. 2011;3(2):196–202.

The Figure legends should read:

Figure 1 Description of Sample Selection for Outcomes Analysis

Figure 2 Timeline for Implementation of Interventions

3. Saeed F, Majeed MH, Kousar N. Easing international medical graduates entry into us training. *J Grad Med Educ*. 2011;3(2):269.

The lead author's name is Fahad Saeed, MD.

4. Sweeney A, Stephany A, Whicker S, Bookman J, Turner DA. Senior Pediatric Residents as Teachers for an Innovative Multidisciplinary Mock Code Curriculum. *J Grad Med Educ*. 2011;3(2):188–195.

The Figure 3 label for the seventh column is: Communicating Effectively.

5. Le-Bucklin KT, Hicks R, Wong A. Impact of a Teaching Rotation on Residents' Attitudes Toward Teaching: A 5-Year Study. *J Grad Med Educ*. 2011;3(2):253–255.

The Results section of the Abstract should read:

Results: Four categories showed significant improvement, including feeling prepared to teach ($P < .0001$), having confidence in their teaching ability ($P < .0001$), being aware of their expectations as a teacher ($P < .0001$), and feeling that their anxiety about teaching was at a healthy level ($P = .0037$). There was an increase in the level of enthusiasm, but the P value did not reach a significant range ($P = .12$). The level of enthusiasm started high and was significantly higher on the pretest than every other tested category ($P < .0001$).

Footnote c to Table 2 should read: *P* value as calculated using the Mann-Whitney U test.