Developing a Valid, Reliable Instrument To Evaluate Learners and Programs

Su-Ting Li, MD, MPH¹ and Daniel C. West, MD²
University of California Davis¹ and University of California San Francisco²
APPD Workshops, April 17, 2010
PAS Workshop, May 4, 2010

Background Information for Small Group Exercises

Establishing Validity of a Measurement Tool

What are you trying to measure?
- What area of knowledge, skill or behavior is being assessed?

What are the theoretical constructs of what you are trying to measure?
- The broad categories of skills or behaviors that constitute what you are trying to measure.
- Often obtained by expert opinion, observation or information from published empirical studies.

What are the elements (definition) of each construct?
- Component parts of each construct.
- Usually multiple.
- Usually these will become the items in the assessment tool.
- Usually derived from expert opinion (e.g., expert panels or focus groups) and/or empiric studies.

How will you establish 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? (i.e. if measuring math skills, does the instrument include what both students and teachers think are relevant math problems?)
- Important for buy in or to get cooperation—can be very important in residency programs.
- Weakest evidence of validity.

Copyright © 2010; University of California Regents, S.T. Li, and D.C. West. All rights reserved.
2. **Content validity** – Identify evidence (i.e., empiric evidence) that items included in assessment match the competency to be assessed. The quality and rigor of the evidence is the key difference between content validity and face validity.

- Demonstrate that the instrument includes 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?
- Evidence is often expert judgment.
  - Expert panel that rates validity of items (usually iterative, Delphi method is good example).
- Can be based on other empirical evidence from the medical literature.

3. **Construct validity** – Is there agreement between the theoretical construct and the measurements obtained with your new tool?

   a. **Convergent validity** – Are measurements of the same theoretical construct consistent?
      - If measuring addition skills, people who answer one 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?
      - People who score high on one assessment of math skills (e.g. addition) are likely to score high on an assessment of different, but related math skills (e.g. subtraction).

   b. **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.

4. **Criterion validity** – Is there the expected relationship between measurements obtained by new tool and other established or agreed upon measures?

   a. **Criterion-standard validity** (“gold standard”)
      - Very few gold standards exist.
      - If there is a gold standard, how will you compare your measure to the gold standard?
        - Typically comparison is made by administering assessment and comparing that score to the gold standard score by…
          - Determining degree of correlation.
          - Similar to diagnostic test: sensitivity/specificity/likelihood ratios.
          - Regression modeling (preferred method because it provides more information about the strength of the relationship between the new assessment and the gold standard while potentially controlling for confounding factors).
      - If no gold-standard, then you will need to use a combination of measures of construct and criterion validity to build a case for validity.
b. Predictive validity – Does this measure predict performance or outcomes in the future?
  • For example, do medical students who score higher on USMLE Step 2 perform better as senior residents than students who score lower? Do patients who score high on respiratory distress observation scale go on to develop respiratory failure?

c. Concurrent validity – Does this measure relate to current performance or outcomes? In other words, is the measure 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?
  • Do patients with high respiratory distress score also have high PRISM score?

B. External validity
  • How generalizable or transferable is the instrument to other settings or situations?
    o How much is the institution, residents, or others where the tool is tested similar to those at other institutions?
    o 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?
    o How will your tool perform in conditions and settings outside of the controlled conditions of your study? How will your tool perform in “real-life” applications (i.e. efficacy versus effectiveness)?
  • Establishing this usually requires multi-institution studies.

Establishing Reliability of Measurement Tool

What is Reliability:
  • The dependability or reproducibility of a measurement tool?
  • Does it hit the same spot on the target consistently?
  • Example: If a washing machine runs through all the cycles every time you turn it on, it is reliable. If it gets clothes clean, it is valid.

Establishing Reliability Involves Application of Measurement Theory:
  • Goal is to measure characteristics, attributes, behaviors or skills, usually in people.
  • Measurement theory describes, categorizes, and assesses the quality of these measures.
  • All measurement tools have error.
  • Error in this setting is defined as variability in measured score due to sources other than the differences in the characteristic or attribute you are trying to measure.
  • The first step in assessing reliability is identifying potential sources of error.

What are the potential sources of error?
  • Should be sources of error that you care about.
  • Need to be able to measure the error.
  • Should be something that you can do something about.
  • Examples:
Raters: Differences between raters (e.g. hard raters versus easy raters, well trained raters versus poorly trained raters, raters with good eye sight versus poor eye sight, etc.)

Occasion: Differences from one occasion to another (e.g. resident performs better pre-call than post-call, etc.)

Items: Differences based on item assessed (e.g. some questions on a test are harder than others or a resident performs better on asthma-related problems versus development problems).

Location: Differences based on location (e.g. resident performs better in outpatient clinic versus ICU).

- Sources of error can interact with each other.
  - The effect of a particular source of error could depend on some other source of error or the individual being tested.
  - For example, a particular rater may be hard or easier depending on the occasion they have to do the rating (e.g. one rater might not like Mondays so always grades harder on Mondays than Friday. Another rater might not care about Mondays).

Minimizing error (undesirable sources of measurement variance)
- Need to understand sources of error in order to maximize reliability of measurement tool
- Examples of things you can do to reduce measurement error:
  - Train raters
  - Optimize assessment occasions (e.g. conditions)
  - Improve clarity of tool
  - Develop anchors for response scales
  - Enhance validity of tool

Approach to establishing reliability depends on how you plan to use the assessment tool.
- To establish reliability you will need to use some form of measurement theory. There are three components of measurement theory.
  - Classical Test Theory
  - Generalizability Theory (G-theory)
  - Item Response Theory (this is used for test construction and beyond the scope of this workshop)
- Choice of which component of measurement theory to use depends on what you care about.
  - Will this be a high stakes assessment?
    - Will you use the information to make decisions about advancement, whether someone meets competency standards, or whether someone will graduate or not?
    - Will you use the information for formative feedback only?
    - Will you use the information in a research study?
  - Do you care about determining how best to administer the instrument?
    - If it is a high stakes assessment, you should care about this.
    - If you do reliability assessment the right way, you can use the information to figure out how to minimize the effect of error.

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 (the score derived from 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 is coming from, the interactions between sources of error or how to modify and improve the reliability of an assessment tool.
- Knowing how to modify administration can be very important in research studies when resources are limited.
- Usually not a good choice for high stakes assessments.

**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.
  - Example: 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?
- Will not help you figure out how to reduce the effect of error.

**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?

**Generalizability Theory (G-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
  - Typical sources of error include:
    - Rater
    - Occasion
    - Items
    - Setting
  - 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.
• Use Analysis of Variance (ANOVA) to determine contribution of the subject (e.g. resident) and each source of error (called “facets”) to the variance of scores on a particular assessment tool.
• 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 where grades will be given.
• Usually the appropriate choice for high stakes assessments.

**Decision-Studies (D-Study) can be used to minimizing error and determine optimum administration of measurement tool**

• Use results of G-study (variance data) to figure out how to best administer the assessment tool to minimize error (i.e. achieve adequate reliability).
• Can run “what if” scenarios to figure out how to modify assessment in order to maximize reliability (called “Decision-study or D-study”).
  o For example: can determine how many occasions to administer an assessment or how many raters are necessary in order to achieve adequate reliability.

**Examples**

1. **Respiratory Distress Measurement Tool**: Can use this to determine how many raters you will need, how many items to include in your tool, or how often you need to do measurement.
2. **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.
3. **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.

**What is adequate reliability?**

**Reliability and Generalizability Coefficients:**

• Both Classical test and Generalizability theories produce coefficients (on scale of 0-1) that indicate reliability.
• Classical test theory – produces reliability coefficient.
• Generalizability theory – produces generalizability coefficient (G-coefficient).
• What is considered good?

\[
\begin{align*}
0.8 – 1.0 & \quad \text{Good reliability} \\
0.6 – 0.8 & \quad \text{Marginal reliability, not reliable enough for high stakes decision making} \\
< 0.6 & \quad \text{Poor reliability}
\end{align*}
\]
References

General References on Validity and Reliability


Example of application of Generalizability Theory

Acad Med. 2008 Dec;83(12):1196-203.

Measuring knowledge structure: reliability of concept mapping assessment in medical education. Srinivasan M, McElvany M, Shay JM, Shavelson RJ, West DC.

Abstract

PURPOSE: To test the reliability of concept map assessment, which can be used to assess an individual's "knowledge structure," in a medical education setting. METHOD: In 2004, 52 senior residents (pediatrics and internal medicine) and fourth-year medical students at the University of California-Davis School of Medicine created separate concept maps about two different subject domains (asthma and diabetes) on two separate occasions each (four total maps). Maps were rated using four different scoring systems: structural (S; counting propositions), quality (Q; rating the quality of propositions), importance/quality (I/Q; rating importance and quality of propositions), and a hybrid system (H; combining elements of S with I/Q). The authors used generalizability theory to determine reliability. RESULTS: Learners (universe score) contributed 40% to 44% to total score variation for the Q, I/Q, and H scoring systems, but only 10% for the S scoring system. There was a large learner-occasion-domain interaction effect (19%-23%). Subsequent analysis of each subject domain separately demonstrated a large learner-occasion interaction effect (31%-37%) and determined that administration on four to five occasions was necessary to achieve adequate reliability. Rater variation was uniformly low. CONCLUSIONS: The Q, I/Q, and H scoring systems demonstrated similar reliability and were all more reliable than the S system. The findings suggest that training and practice are required to perform the assessment task, and, as administered in this study, four to five testing occasions are required to achieve adequate reliability. Further research should focus on whether alterations in the concept mapping task could allow it to be administered over fewer occasions while maintaining adequate reliability.

Examples of published valid, reliable instruments


BACKGROUND. Competency in pediatric resuscitation is an essential goal of pediatric residency training. Both the exigencies of patient care and the Accreditation Council for Graduate Medical Education require assessment of this competency. Although there are standard courses in pediatric resuscitation, no published, validated assessment tool exists for pediatric resuscitation competency.
OBJECTIVE. The purpose of this work was to develop a simulation-based tool for the assessment of pediatric residents’ resuscitation competency and to evaluate the tool’s reliability and preliminarily its validity in a pilot study.

METHODS. We developed a 72-question yes-or-no questionnaire, the Tool for Resuscitation Assessment Using Computerized Simulation, representing 4 domains of resuscitation competency: basic resuscitation, airway support, circulation and arrhythmia management, and leadership behavior. We enrolled 25 subjects at each of 5 different training levels who all participated in 3 standardized code scenarios using the Laerdal SimMan universal patient simulator. Performances were videotaped and then reviewed by 2 independent expert raters.

RESULTS. The final version of the tool is presented. The intraclass correlation coefficient between the 2 raters ranged from 0.70 to 0.76 for the 4 domain scores and was 0.80 for the overall summary score. Between the 2 raters, the mean percent exact agreement across items in each domain ranged from 81.0% to 85.1% and averaged 82.1% across all of the items in the tool. Across subject groups, there was a trend toward increasing scores with increased training, which was statistically significant for the airway and summary scores.

CONCLUSIONS. In this pilot study, the Tool for Resuscitation Assessment Using Computerized Simulation demonstrated good interrater reliability within each domain and for summary scores. Performance analysis shows trends toward improvement with increasing years of training, providing preliminary construct validity.

Violato C, Lockyer JM, Fidler H. Assessment of Pediatricians by a Regulatory Authority. *Pediatrics* 2006;177;796-802.

OBJECTIVE. To determine whether it is possible to develop feasible, valid, and reliable multisource feedback data for pediatricians.

METHODS. Surveys with 40, 22, 38, and 37 items were developed for assessment of pediatricians by patients, co-workers, medical colleagues, and themselves, respectively, using 5-point scales with an “unable to assess” category. Items addressed key competencies related to communication skills, professionalism, collegiality, continuing professional development, and collaboration. Each pediatrician was assessed by 25 patients, 8 medical colleagues, and 8 co-workers. Feasibility was assessed with response rates for each instrument. Validity was assessed with rating profiles, the percentage of participants unable to assess the physician for each item, and exploratory factor analyses to determine which items grouped together into scales. Cronbach’s alpha and generalizability coefficient analyses assessed reliability.

RESULTS. One hundred pediatricians participated. The mean number of respondents per physician was 23.4 (93.6%) for patients, 7.6 (94.8%) for co-workers, and 7.6 (95.5%) for medical colleagues. The mean ratings ranged from 4 to 5 for each item on each scale. Few items had high percentages of “unable to assess” responses. The factor analyses revealed a 4-factor solution for the patient survey, a 3-factor solution for the co-worker survey, and a 4-factor solution for the medical colleague survey, accounting for at least 64% of the variance. All instruments had high internal consistency. The generalizability coefficients were .85 for patients, .87 for co-workers, and .78 for medical colleagues.

CONCLUSION. Surveys can be developed to provide feedback data on key competencies.