Validity According to *Standards for Educational and Psychological Testing*

According to the *Standards* (1999), validity is "the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests" (p. 9).

For example, how does one know that scores from a scale designed to measure test anxiety provide scores that reflect test anxiety? More broadly, how can one demonstrate that scores derived from an instrument are valid?

Validity is established not by a single method, but by providing multiple examples of evidence. Below are some of the more commonly employed means for providing evidence for validity.

I. Evidence based on Test Content (EDUR 9131 relevant)

Logical validity, or content validity, stems from the logical/judgmental analysis of items and instrument format. As Goodwin and Leech (2003) explain:

> "...this type of validity evidence is based on logical analyses and experts' evaluations of the content of the measure, including items, tasks, formats, wording, and processes required of examinees. In general, it addresses questions about the extent to which content of a measure represents a specified content domain." (p. 183)

A basic, summary of steps one follows for judging content validity include the following:

1. Define construct
2. Identify domains of construct
3. Develop item pool to fit domains with adequate sampling of each domain
4. Expert item analysis, expert review of entire instrument
5. Pilot test with feedback

A more detailed approach to content validity is found in Holmbeck and Devine (2009). See also Table 1 of Goodwin and Leech (2003) for ideas about expert and participant reviews concerning instrument and items (e.g., item importance, clarity, relevance, bias).

II. Evidence based on Response Processes (EDUR 9131)

Whiston (2009) explains that evidence based upon response processes focuses on whether

> "…individuals either perform or respond in a manner that corresponds to the construct being measured. For example, if an assessment is attempting to measure vocational interest, then the instrument developers might examine whether people are answering the items based on what they like to do rather than what they can do." (pp. 70-71).

In short, this type of evidence addresses whether respondents view and understand items, instructions, and instrument in same way and respond using anticipated methods.

Examples:

- Teachers using rubric to assign scores (rating responses in appropriate manner)
- Checking for social acceptable responses on attitude measures

- Asking respondents how they derived their response choice
- Determining whether mathematic reasoning was used to derive answers (test-taker is responding to items using appropriate processes)
- "Talk-aloud" – asking respondents to explain their reasoning for the answers provided
- Show your work—illustrate how answers were determined
- Demographic item:

Sex ___

When responding to this item, do respondents think male vs. female, or do respondents think type or frequency? Responses may vary depending upon maturity of respondent. Revise item to make clearer:

What is your Sex: Female ____
Male ____

This type of validity evidence can overlap with content validity because it is partially concerned with how and why individuals respond the way they do. One reason for the field test in content validity is to determine whether individuals are reading and interpreting questionnaire items in a similar manner, and whether they attempt to address those items using a framework or schema that aligns with what the scale was designed to measure.

III. Evidence based on Internal Structure (EDUR 9131 relevant)

This type of evidence concerns whether data derived from a measure conform to theoretically expected patterns. Think in terms of construct and domains – do the domains show distinct response patterns; do different constructs show distinct response patterns?

Methods for addressing internal structure include:

- Correlations among items and scale summated scores
- Correlations among domains of a construct
- Exploratory factor analysis
- Confirmatory factor analysis
- Internal consistency is NOT a measure of internal structure (Floyd et al. 2005 mistakenly list internal consistency as a measure of internal structure)

Why is internal consistency not a measure of internal structure? Below is a discussion forum post I made about this issue:

> On Tue, 6 Oct 2009 16:55:36 -0500, Trudy Reynolds
> <<autumn2@SHAWNEELINK.NET> wrote:
>
> >Can anyone explain why alpha is unrelated to the internal structure of
> >the test? According to Traub (1997), "Kelly criticized the KR formulas
> >in a 1942 article in Psychometrika on the grounds that they are valid
> >only for tests 'with unity of purpose'-that is, for tests composed of
> >items that share just one factor in common." I would appreciate any
> >comments regarding this issue.

>
>Trudy Reynolds, Ph.D.

From: Bryan W. Griffin <bwgriffin@GEORGIASOUTHERN.EDU>
Subject: Re: Utility of Cronbach's alpha
To: EDRESMETH-L@LISTSERV.UCONN.EDU
Date: Wednesday, October 7, 2009, 1:55 PM

Hi Trudy –

I am not an expert in measurement, but I will take a stab at addressing your question.

Cronbach's alpha is not designed to measure internal structure (think in terms of factor analysis here), but can provide a measure of internal consistency (think in terms of mean inter-item correlations here), although that appears to be questionable too.

As noted by Bruce above alpha is a function of covariances (and correlations), and it is also a function of number of items. Here is a formula for Cronbach's alpha in terms of mean inter-item correlations (m[r]) and the number of items (k):

alpha = (k * m[r]) / (1 + (k - 1) * m[r])

where

k = number of items on instrument used to calculate alpha,
m[r] = mean correlation among the k items.

Given this formula, the following two scenarios are possible:

1. Researcher has instrument with 4 items designed to measure the same construct, so there should be one factor here. The mean correlation among items is m[r] = .5862. Using the formula above:

alpha = (k * m[r]) / (1 + (k - 1) * m[r])
alpha = (4 * .5862) / (1 + (4 -1) * .5862) =~ .85

2. Researcher has an instrument designed to measure 4 unrelated or weakly related constructs. Factor analysis reveals that the internal structure to contain four distinct factors. There are a total of 50 items on this instrument (10 items for factor A, 8 items for factor B, 17 items for factor C, and 15 items for factor D). The mean correlation among all 50 items is m[r] = .1018. If one erroneously applies the alpha reliability formula to these 50 items, the result would be:

alpha = (k * m[r]) / (1 + (k - 1) * m[r])
alpha = (50 * .1018) / (1 + (50 -1) * .1018) =~ .85

Note that Cronbach’s alpha is the same, within rounding error, in both situations, yet the internal structure is very different in both cases. These two examples demonstrate that Cronbach’s alpha is not designed to reveal internal structure of items. Better to use EFA or CFA to assess structure.

These examples also illustrate that Cronbach’s alpha also does not reveal much about the mean correlations among items because alpha is so influenced by the number of items.

Bryan

IV. Evidence based on Relations to Other Variables (EDUR 9131 relevant)

This type of evidence is used to demonstrate that measured scores behave in predictable patterns. Sometimes this type of validity is referenced as criterion-related and includes concurrent, predictive, convergent, and discriminate validity.

In essence researchers predict how obtained scores will relate to other variables. If the predictions are supported statistically, then this provides some evidence for validity of scores.

Examples of evidence based upon relations to other variables:

- Scores from new anxiety measure correlated with scores from previously established anxiety measure
- Scores from anxiety measure are predicted to correlate negatively with self-efficacy scores; correlate positively with number of items answered incorrectly
- Scores from anxiety measure are predicted to be higher for females than males; predicted to be higher for untreated group than group provided with anxiety reduction training
- Scores from anxiety measure are predicted to be unrelated to measures of subject interest or instructor ratings

IV. Evidence based on Consequences of Testing

What are the expected and unexpected results, or consequences, of measurement? This is especially relevant for diagnostic scales used to discriminate among individuals (e.g., reading readiness tests, graduation tests, etc.).

This type of evidence doesn't fit well with the purpose of EDUR 9131, but is an important component of validity especially for high-stakes assessments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Floyd et al. (2005). Measurement properties of indirect assessment methods for functional behavioral assessments: A review of research. School Psychology Review, 24, 58-73.

Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new Standards for Educational and Psychological Testing: Implications for measurement courses. Measurement and Evaluation in Counseling and Development, 36, 181-191.

Holmbeck, G. & Devine, K. (2009). Editorial: An author's checklist for measure development and validation manuscripts. Journal of Pediatric Psychology, 1-6.

Whiston, S. (2009). Principles and applications of assessment in counseling (3rd edition). Belmont, CA, USA: Brooks/Cole, Cengage Learning.