Handout 5: Establishing the Validity of a Survey Instrument

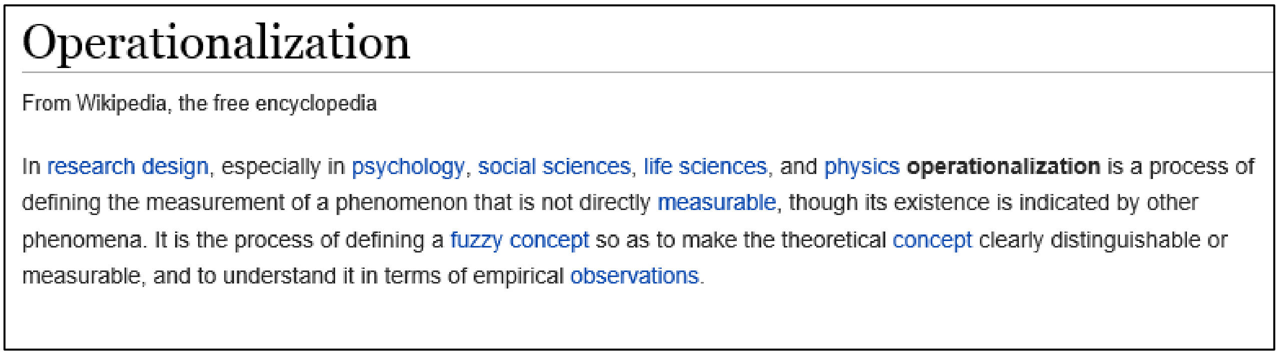
STAT 335 – Fall 2016

In this handout, we will discuss different types of and methods for establishing *validity*. Recall that this concept was defined in Handout 3 as follows.

Definition

Validity – This is the extent to which survey questions measure what they are supposed to measure.

In order for survey results to be useful, the survey must demonstrate **validity**. To better understand this concept, it may help to also consider the concept of **operationalization**. Wikipedia.org defines this as follows:

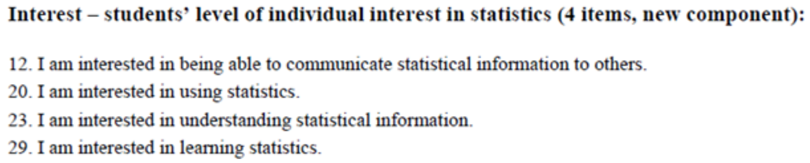


Operationalization

From Wikipedia, the free encyclopedia

In research design, especially in psychology, social sciences, life sciences, and physics **operationalization** is a process of defining the measurement of a phenomenon that is not directly measurable, though its existence is indicated by other phenomena. It is the process of defining a fuzzy concept so as to make the theoretical concept clearly distinguishable or measurable, and to understand it in terms of empirical observations.

For example, in the previous handout we considered measuring students' interest in statistics using the SATS (Survey of Attitudes Toward Statistics). Note that the construct of *Interest* is a theoretical and rather vague concept – there is no clear or obvious way to measure this. So, the creators of this survey defined their own measure of *Interest* using these four questions:



Interest – students' level of individual interest in statistics (4 items, new component):

12. I am interested in being able to communicate statistical information to others.
20. I am interested in using statistics.
23. I am interested in understanding statistical information.
29. I am interested in learning statistics.

The resulting sub-score for *Interest* is their operationalization of this construct. Will the resulting score really measure *Interest*? Will the resulting scores for the other constructs on the SATS really measure what the researchers intend? These are the questions we seek to answer when establishing the **construct validity** of this survey.

In the remainder of this handout, we will introduce various types of construct validity and briefly discuss how survey instruments are shown to be valid.

TYPES OF CONSTRUCT VALIDITY

When designing survey questionnaires, researchers may consider one or more of the following types of **construct validity**.

Types of Construct Validity

Face Validity – An operationalization has face validity when it appears to observers that it truly measures the construct it is intended to measure.

Content Validity – An operationalization has construct validity when it adequately covers the range of meanings included in the construct it is intended to measure.

Criterion-Related Validity – This is assessed by investigating the relationship between the operationalization and other variables. The following are all specific types of criterion-related validity:

- **Predictive Validity** – This is established if the operationalization is able to predict another variable that it should theoretically be able to predict.
- **Concurrent Validity** – This is established if the operationalization is well correlated with other variables measured at the same time to which it should theoretically be related.
- **Convergent Validity** – This is established if the operationalization is well correlated with other variables to which it should theoretically be related.
- **Discriminant Validity** – This is established if the operationalization is shown to be dissimilar from other variables that it theoretically should not be related to.

Ways to establish these types of validity are discussed in more detail below.

Face Validity

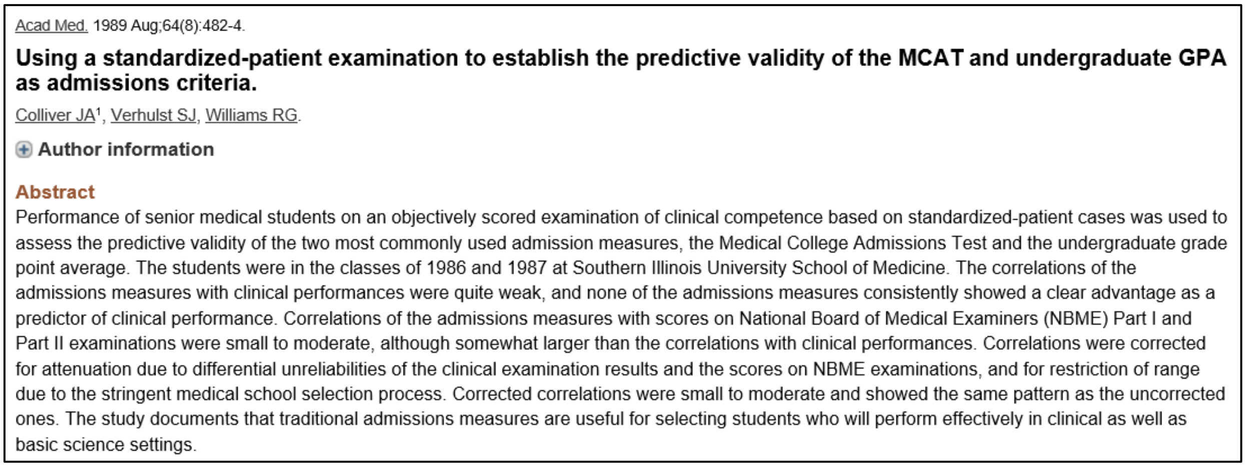
Face Validity can't be established with any sort of statistical analysis. Instead, it's based on a subjective judgment call (which makes it one of the weaker ways to establish construct validity). The best approach for establishing face validity is to assemble a panel of experts to report on whether or not they feel an operationalization appears to be a good measure of the construct of interest.

Content Validity

Like face validity, this is also best established by assembling a panel of experts. The researcher will describe the content domain for their construct (e.g., "My goal is to measure students' interest in statistics.") The experts will then be asked to judge how well the operationalization covers all of the criteria that constitute the content domain.

Predictive Validity

As mentioned above, this involves assessing the ability of the operationalization to predict something it should theoretically be able to predict. Typically, this involves the computation of correlation coefficients. For example, the abstract below describes a study conducted to assess the predictive validity of the Medical College Admissions Test (MCAT).



Acad Med. 1989 Aug;64(8):482-4.

Using a standardized-patient examination to establish the predictive validity of the MCAT and undergraduate GPA as admissions criteria.

Colliver JA[1], Verhulst SJ, Williams RG.

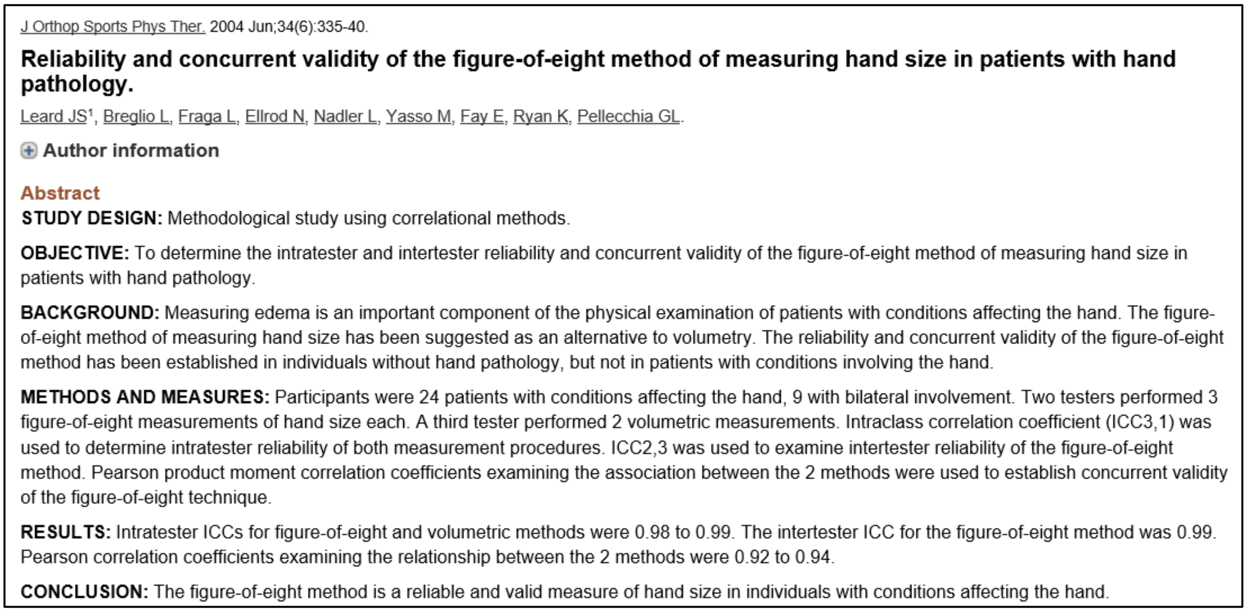
⊕ **Author information**

Abstract

Performance of senior medical students on an objectively scored examination of clinical competence based on standardized-patient cases was used to assess the predictive validity of the two most commonly used admission measures, the Medical College Admissions Test and the undergraduate grade point average. The students were in the classes of 1986 and 1987 at Southern Illinois University School of Medicine. The correlations of the admissions measures with clinical performances were quite weak, and none of the admissions measures consistently showed a clear advantage as a predictor of clinical performance. Correlations of the admissions measures with scores on National Board of Medical Examiners (NBME) Part I and Part II examinations were small to moderate, although somewhat larger than the correlations with clinical performances. Correlations were corrected for attenuation due to differential unreliabilities of the clinical examination results and the scores on NBME examinations, and for restriction of range due to the stringent medical school selection process. Corrected correlations were small to moderate and showed the same pattern as the uncorrected ones. The study documents that traditional admissions measures are useful for selecting students who will perform effectively in clinical as well as basic science settings.

Concurrent Validity

This involves assessing the strength of the relationship between the operationalization of interest and other variables measured *at the same time* to which the operationalization should theoretically be related. Once again, this typically this involves the computation of correlation coefficients. For example, consider the following abstract.



J Orthop Sports Phys Ther. 2004 Jun;34(6):335-40.

Reliability and concurrent validity of the figure-of-eight method of measuring hand size in patients with hand pathology.

Leard JS[1], Breglio L, Fraga L, Ellrod N, Nadler L, Yasso M, Fay E, Ryan K, Pellecchia GL.

⊕ **Author information**

Abstract

STUDY DESIGN: Methodological study using correlational methods.

OBJECTIVE: To determine the intratester and intertester reliability and concurrent validity of the figure-of-eight method of measuring hand size in patients with hand pathology.

BACKGROUND: Measuring edema is an important component of the physical examination of patients with conditions affecting the hand. The figure-of-eight method of measuring hand size has been suggested as an alternative to volumetry. The reliability and concurrent validity of the figure-of-eight method has been established in individuals without hand pathology, but not in patients with conditions involving the hand.

METHODS AND MEASURES: Participants were 24 patients with conditions affecting the hand, 9 with bilateral involvement. Two testers performed 3 figure-of-eight measurements of hand size each. A third tester performed 2 volumetric measurements. Intraclass correlation coefficient (ICC3,1) was used to determine intratester reliability of both measurement procedures. ICC2,3 was used to examine intertester reliability of the figure-of-eight method. Pearson product moment correlation coefficients examining the association between the 2 methods were used to establish concurrent validity of the figure-of-eight technique.

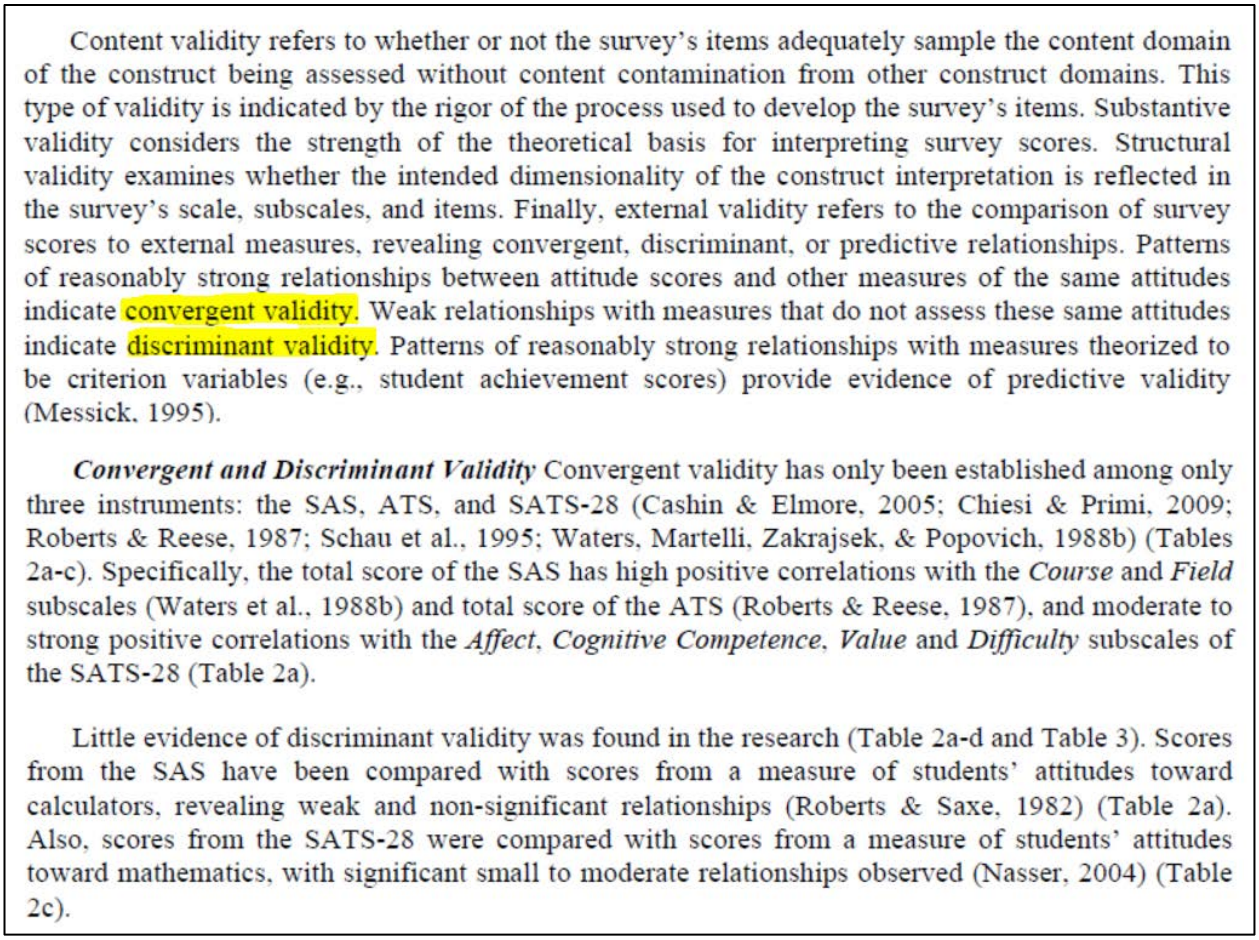
RESULTS: Intratester ICCs for figure-of-eight and volumetric methods were 0.98 to 0.99. The intertester ICC for the figure-of-eight method was 0.99. Pearson correlation coefficients examining the relationship between the 2 methods were 0.92 to 0.94.

CONCLUSION: The figure-of-eight method is a reliable and valid measure of hand size in individuals with conditions affecting the hand.

Convergent and Divergent Validity

As stated above, we establish **convergent** validity if the operationalization of interest is well correlated with other variables to which it theoretically should be related; alternatively, we establish **divergent** validity if the operationalization of interest is *not* well correlated with other variables to which it theoretically should not be related.

For example, consider the following excerpt from a paper titled "Surveys Assessing Students' Attitudes Toward Statistics: A Systematic Review of Validity and Reliability" written by Meaghan M. Nolan et al. This was published in the *Statistics Education Research Journal*, 11 (2), pp. 103-123 (November 2012).



Content validity refers to whether or not the survey's items adequately sample the content domain of the construct being assessed without content contamination from other construct domains. This type of validity is indicated by the rigor of the process used to develop the survey's items. Substantive validity considers the strength of the theoretical basis for interpreting survey scores. Structural validity examines whether the intended dimensionality of the construct interpretation is reflected in the survey's scale, subscales, and items. Finally, external validity refers to the comparison of survey scores to external measures, revealing convergent, discriminant, or predictive relationships. Patterns of reasonably strong relationships between attitude scores and other measures of the same attitudes indicate convergent validity. Weak relationships with measures that do not assess these same attitudes indicate discriminant validity. Patterns of reasonably strong relationships with measures theorized to be criterion variables (e.g., student achievement scores) provide evidence of predictive validity (Messick, 1995).

Convergent and Discriminant Validity Convergent validity has only been established among only three instruments: the SAS, ATS, and SATS-28 (Cashin & Elmore, 2005; Chiesi & Primi, 2009; Roberts & Reese, 1987; Schau et al., 1995; Waters, Martelli, Zakrajsek, & Popovich, 1988b) (Tables 2a-c). Specifically, the total score of the SAS has high positive correlations with the *Course* and *Field* subscales (Waters et al., 1988b) and total score of the ATS (Roberts & Reese, 1987), and moderate to strong positive correlations with the *Affect*, *Cognitive Competence*, *Value* and *Difficulty* subscales of the SATS-28 (Table 2a).

Little evidence of discriminant validity was found in the research (Table 2a-d and Table 3). Scores from the SAS have been compared with scores from a measure of students' attitudes toward calculators, revealing weak and non-significant relationships (Roberts & Saxe, 1982) (Table 2a). Also, scores from the SATS-28 were compared with scores from a measure of students' attitudes toward mathematics, with significant small to moderate relationships observed (Nasser, 2004) (Table 2c).

Note that to estimate the degree to which any two measures are related to each other, we typically use the correlation coefficient.

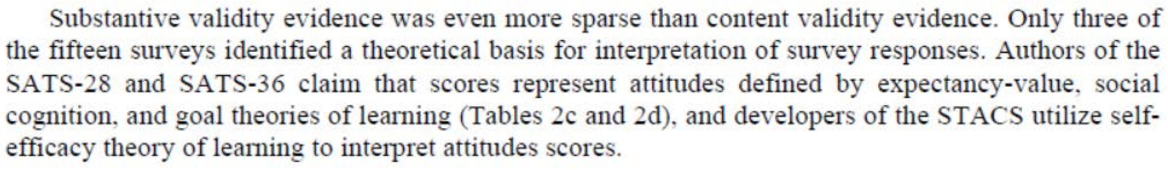
A MORE MODERN CONCEPT OF CONSTRUCT VALIDITY

In the 1990s, Samuel Messick proposed a new, more unified framework for the concept of validity. This framework describes six distinguishable aspects of validity. He argues that these six aspects should be viewed as interdependent and complementary to one another.

Modern Concept of Construct Validity

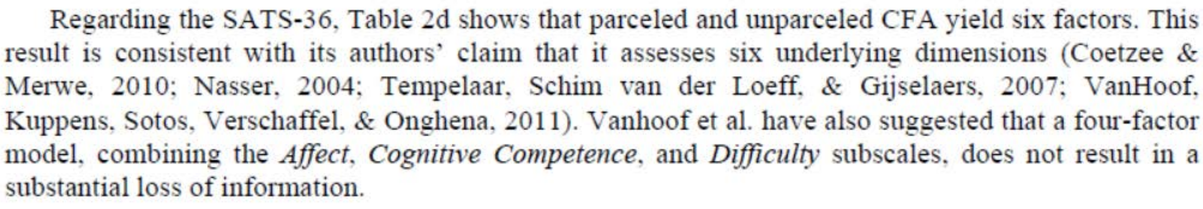
Content Validity – An operationalization has content validity when it adequately covers the range of meanings included in the construct it is intended to measure (i.e., it is representative of the construct). When assessing this, one should also consider the content relevance of each item and its technical quality.

Substantive Validity – This considers the strength of the theoretical rationales for interpreting the survey scores. For example, consider the following excerpt from the aforementioned paper by Nolan et al.



> Substantive validity evidence was even more sparse than content validity evidence. Only three of the fifteen surveys identified a theoretical basis for interpretation of survey responses. Authors of the SATS-28 and SATS-36 claim that scores represent attitudes defined by expectancy-value, social cognition, and goal theories of learning (Tables 2c and 2d), and developers of the STACS utilize self-efficacy theory of learning to interpret attitudes scores.

Structural Validity – This is assessed by investigating the degree to which the operationalization adequately reflects the dimensionality of the construct to be measured. For example, a researcher may conduct a factor analysis using the observed scores for a survey. If the factor analysis reveals the same number of factors as there were constructs measured on the survey (and the factor loadings show that the right questions are grouped together within factors), then structural validity is established. Such an analysis is discussed in the following excerpt from the Nolan et al. paper.



> Regarding the SATS-36, Table 2d shows that parceled and unparceled CFA yield six factors. This result is consistent with its authors' claim that it assesses six underlying dimensions (Coetzee & Merwe, 2010; Nasser, 2004; Tempelaar, Schim van der Loeff, & Gijselaers, 2007; VanHoof, Kuppens, Sotos, Verschaffel, & Onghena, 2011). Vanhoof et al. have also suggested that a four-factor model, combining the *Affect*, *Cognitive Competence*, and *Difficulty* subscales, does not result in a substantial loss of information.

Modern Concept of Construct Validity
Generalizability – This examines the extent to which scores generalize across different population groups, different situations or settings, different time periods, and/or to other operationalizations representative of the construct domain.
External Validity – This refers to the comparison of the operationalization of interest to external measures (see the earlier discussions of convergent, divergent, and predictive validity).
Consequential Validity – This includes gathering evidence and rationales for evaluating the consequences of score interpretations from a survey. Researchers should accrue evidence of positive consequences and evidence that adverse consequences are minimal.

With this more modern framework, the process of establishing construct validity could potentially involve the accumulation of all six of the aforementioned forms of validity evidence. Note, however, that a "compelling argument" for validity can still be made even if some of these aren't addressed.

Source: Messick, Samuel. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

CONSIDERING VALIDITY AND RELIABILITY SIMULTANEOUSLY

Often, validity and reliability are viewed as completely separate ideas. To think about how the two are related, we can use a "target" analogy. Let the center of the target represent the construct you intend to measure. For each subject that responds to your survey questionnaire, you take a shot at the target. If you measure the concept perfectly for a person, you hit the center of that target. The figure below shows four possible situations.

Reliable, but not Valid	Valid, but not Reliable	Neither Reliable nor Valid	Both Reliable and Valid