How to Measure Survey Reliability and Validity

Validity

Contributors: Mark S. Litwin
Book Title: How to Measure Survey Reliability and Validity
Chapter Title: "Validity"
Pub. Date: 1995
Access Date: November 04, 2015
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9780803957046
Online ISBN: 9781483348957
DOI: http://dx.doi.org/10.4135/9781483348957.n3
Print pages: 33-47

http://dx.doi.org/10.4135/9781483348957.n3

[p. 33 ↓]

Chapter 3: Validity

Besides determining a survey item's or scale's reliability, you must assess its **validity**, or how well it measures what it sets out to measure. An item that is supposed to measure pain should measure pain and not some related variable (e.g., anxiety). A scale that claims to measure emotional quality of life should not measure depression, a related but different variable. Reliability assessments are necessary, but they are not sufficient when examining the psychometric properties of a survey instrument. Once you document that a scale is reliable over time and in alternate forms, you must then make sure that it is reliably measuring the truth.

[p. 34 ↓] Example 3.1 takes another look at the problem of measuring fabric that was explored in Chapter 2's Example 2.1. Before, your concern was with reliability, but now you are concerned with validity.

Example 3.1: Fabric Measurement Revisited

Recall from Example 2.1 that Honor Guard Fabric Company uses expensive aluminum yardsticks to measure the length of each piece of fabric it sells. You previously determined that these metal yardsticks are very reliable because they measure out the exact same length of cloth every time. Suppose that during a store audit Honor Guard finds that each of its 1-yard measuring sticks is actually 40 inches long. Every single time the clerk sells a yard of fabric, the measuring sticks reliably count out 40 inches of fabric. Over the years, the company realizes that it has given away thousands of inches of extra fabric because its expensive aluminum measuring sticks are too long. The measurement instruments are reliable but not valid.

Validity must be documented when evaluating new survey instruments or when applying established survey instruments to new populations. It is an important measure of a survey instrument's accuracy.

Types of Validity

Several types of validity are typically measured when assessing the performance of a survey instrument: face, content, criterion, and construct.

[p. 35 ↓]

Face

Face validity is based on a cursory review of items by untrained judges, such as your sister, boyfriend, or squash partner. Assessing face validity might involve simply showing your survey to a few untrained individuals to see whether they think the items look OK to them. It is the least scientific measure of all the validity measures and is often confused with content validity. Although the two are similar, face validity is a much more casual assessment of item appropriateness. In fact, many researchers do not consider face validity a measure of validity at all.

Content

Content validity is a subjective measure of how appropriate the items seem to a set of reviewers who have some knowledge of the subject matter. The assessment of content validity typically involves an organized review of the survey's contents to ensure that it includes everything it should and does not include anything it shouldn't. When examining the content validity of medical scales, for example, it is important that actual patients and their families be included in the evaluation process. Clinicians may be unaware of the subtle nuances experienced by patients who live day-to-day with a medical condition. Families also may provide helpful insights into dimensions that might otherwise be overlooked by “experts.” This said, it remains important for clinicians to review the items for relevance to and focus on the variables of interest. Content validity is not quantified with statistics. Rather, it is presented as an overall opinion of a group of trained judges. Strictly speaking, it is not a scientific measure of a survey instrument's accuracy. Nevertheless, it provides a good foundation on which to build a methodologically rigorous assessment of a survey instrument's validity.

[p. 36 ↓] In Example 3.2, content validity is assessed for a sociological survey on interactions between spouses.

Example 3.2: Content Validity of a Marital Interaction Scale

Josh designs a new scale to collect data on marital interaction as a dimension of health-related quality of life. He develops a series of 16 items about spousal communication, interpersonal confidence, and discussions within the marriage. He plans to use his new scale to assess the impact of social support on a large population of married cancer patients undergoing a difficult and stressful chemotherapy protocol

Before administering his new scale, Josh asks 15 individuals – 3 oncologists, 3 psychologists, 2 social workers, 1 oncology nurse practitioner, 4 cancer patients, and 2 spouses of cancer patients—to review each of the items. He asks these reviewers to rate each item and the scale as a whole for appropriateness and relevance to the issue of marital interaction. He also asks each reviewer to list any areas that are pertinent to marital interaction but not covered in the 16 items. Once all the reviews are complete, Josh studies them to determine whether his new survey instrument has content validity.

If he had wished to assess face validity, Josh might have asked his college roommate or his mother to take a look at the survey and tell him whether the items seemed appropriate. Josh decides to bypass face validity because he has chosen to look more carefully at content validity and knows that face validity is basically worthless.

[p. 37 ↓]

Criterion

Criterion validity is a measure of how well one instrument stacks up against another instrument or predictor. It provides much more quantitative evidence on the accuracy of a survey instrument. It may be measured differently, depending on how much published literature is available in the area of study. Criterion validity may be broken down into two components: concurrent and predictive.

Concurrent

Concurrent validity requires that the survey instrument in question be judged against some other method that is acknowledged as a “gold standard” for assessing the same variable. It may be a published psychometric index, a scientific measurement of some factor, or another generally accepted test. The fundamental requirement is that it be regarded as a good way to measure the same concept. The statistic is calculated as a correlation coefficient with that test. A high correlation suggests good concurrent validity. Alternatively, a test may be selected for comparison that is expected to measure an attribute or behavior that is opposite to the dimension of interest. In this case, a low correlation indicates good concurrent validity. The reason why you would not use the established gold standard as your measure of choice is that it may be too cumbersome, expensive, or invasive to apply.

Example 3.3 demonstrates the use of an established scale to assess concurrent validity in a new scale.

[p. 38 ↓]

Example 3.3: Concurrent Validity of a Pain Tolerance Index

Alisha develops a new four-item index to assess pain tolerance in a group of patients scheduled for surgery. The items draw information from patients' memory of their past experiences with pain. The results from the four items are summed to form a Pain Tolerance Index score. The higher the score, the greater the tolerance for pain. Her index is self-administered and takes about 1 minute for patients to complete. To assess concurrent validity, Alisha administers her four items together with a published pain tolerance survey instrument that has been in use for more than a decade in anesthesiology research. It contains 45 items, requires an interviewer, and takes an average of 1 hour to complete. It is also scored as a sum of item responses. It is generally accepted as the gold standard in the field.

Alisha uses both survey instruments to gather data from a sample of 24 patients. Alisha calculates the correlation coefficient to be 0.92 between the two tests of pain tolerance. She concludes that her index has high concurrent validity with the gold standard. Because hers is much shorter and easier to administer, she convinces the principal investigator in a large national study of postoperative pain to use her more efficient index. Alisha publishes her findings and is awarded a generous academic scholarship as a result of her work.

In Example 3.4, a new survey instrument's validity in measuring water supply is assessed by comparing it with a more standard measure of water supply.

[p. 39 ↓]

Example 3.4: Concurrent Validity of a Water Supply Index

Luis develops an index of overall water supply in desert towns. It is a number calculated from a mathematical formula based on average monthly town rainfall, average monthly depth of the town reservoir, and average monthly water pressure in the kitchen of the local elementary school. The higher the index, the greater the water supply in the town. Luis collects data for 12 consecutive months and uses his formula to calculate a water supply index. During that year, he also records the number of days each month that the Department of Water declares as drought days. At the end of the year, Luis plots his index against the number of drought days for each month. He calculates the correlation coefficient between the two data sets to be -0.81. Because he reasons that these two variables should have an inverse relationship, Luis concludes that his index has good criterion validity.

Although it is important to evaluate concurrent validity, you must make sure to select a gold standard test that is truly a good criterion against which to judge your new survey instrument. It is not helpful to show good correlations with some obscure index just because it happens to be published in a journal or book. Always select gold standards that are relevant, well known, and accepted as being good measures of the variable of interest. When testing concurrent validity, select gold standards that have been demonstrated to have psychometric properties of their own. Otherwise, you will be comparing your new scales to a substandard criterion.

[p. 40 ↓]

Predictive

Predictive validity is the ability of a survey instrument to forecast future events, behaviors, attitudes, or outcomes. It may be used during the course of a study to predict response to a stimulus, election winners, success of an intervention, time to a clinical endpoint, or other objective criteria. Over a brief interval, predictive validity is similar to concurrent validity in that it involves correlating the results of one test with the results of another administered around the same time. If the time frame is longer and the second test occurs much later, then the assessment is of predictive validity. Like concurrent validity, predictive validity is calculated as a correlation coefficient between the initial test and the secondary outcome.

Example 3.5 demonstrates that the Pain Tolerance Index that was tested for concurrent validity in Example 3.3 may also be tested for predictive validity.

Example 3.5: Predictive Validity of the Pain Tolerance Index

Fourteen years after her initial success, Alisha from Example 3.3 becomes a professor of gynecology at a well-respected research university. She decides to use her Pain Tolerance Index to predict narcotic requirement in patients undergoing a hysterectomy. Having tested her index for reliability and concurrent validity, she now wants to test it for predictive validity. She administers the index to 24 of her preoperative patients and calculates an index score for each individual. Recall that a high score reflects a high tolerance for pain.

Once all the surgeries have been completed, Alisha reviews the medical records. She notes the number of doses of narcotic that **[p. 41 ↓]** were administered for postoperative pain control in each patient. She then calculates a correlation coefficient between the two data elements: index score and number of narcotic doses. When she finds that, the statistic is -0.84. As expected, there is a strong inverse correlation between the Pain Tolerance Index and the amount of narcotic required after surgery. Alisha is pleased to find that her index has high predictive validity in clinical practice. She publishes her results in a national medical journal and is later promoted to the position of chairperson of the gynecology department at the university.

Continuing with this theme, predictive validity can be used in a variety of settings to measure the accuracy of a survey instrument. One of the most well-known survey instruments is the Scholastic Aptitude Test (SAT). In Example 3.6, the SAT is used on a population of college students to predict academic success.

Example 3.6: Predictive Validity of SAT Scores

Bob is the dean of students at Brook College, a small liberal arts school in Arizona, and decides to look into whether the SAT scores of entering freshmen predict how well the students will perform during their first semester at Brook. The dean looks back into the registrar's records for the past 5 years and gathers two data elements for each freshman: SAT score and first-semester grade point average. The dean enters the two data sets into his laptop computer and calculates a correlation coefficient between the two. To his surprise, he finds the statistic to be 0.45. The students' SAT scores do not appear to have high predictive validity for early **[p. 42 ↓]** success at Brook College. He immediately writes a memo to the dean of admissions asking that evaluation policies be revised to reflect this important new information.

When Jackie, the dean of admissions at Brook, receives Bob's memo, she decides to do a little investigating of her own. She takes Bob's data and breaks them down year by year. Using the same formula, she calculates correlation coefficients with her own laptop computer and finds that over the past 5 years the predictive validity of SAT scores has been 0.21, 0.36, 0.39, 0.57, and 072. Jackie writes a memo back to Bob, suggesting that although SAT scores did not previously have much predictive validity, they have become increasingly more useful in recent years. Jackie proudly sends a copy of her memo to the chancellor for consideration in her upcoming decision on who should be promoted to provost.

Predictive validity is one of the most important ways to measure a test's accuracy in practical applications; however, it is seldom used in longitudinal medical experiments that rely on surveys. Because the time frames are often several years long in such trials, secondary interventions may be implemented during the trial to alter the course of a disease or medical condition. If the final outcomes were compared with a test score from the start of the study, their correlation may be diminished. This would falsely decrease the measured predictive validity of the test and perhaps call into question the statistical qualities of an otherwise valid survey instrument.

Example 3.6 about SAT scores demonstrates that predictive validity (or any other psychometric statistic) may be used in various ways to support different hypotheses. You must be careful with the conclusions you draw from any of these measured psychometric properties. A good exercise is to ask **[p. 43 ↓]** peers in your area who are unfamiliar with your hypothesis to look at a summary of your data and draw conclusions. If enough people draw the same conclusion, you may be somewhat reassured that your inferences are correct. You may also ask peers to take your data and statistics and try to support a point that is opposite to your conclusions. This may open up your mind to different interpretations. It may unhinge your argument, or it may guide your approach to collecting more irrefutable evidence.

Construct

Construct validity is the most valuable yet most difficult way of assessing a survey instrument. It is difficult to understand, to measure, and to report. This form of validity is often determined only after years of experience with a survey instrument. It is a measure of how meaningful the scale or survey instrument is when in practical use. Often, it is not calculated as a quantifiable statistic. Rather, it is frequently seen as a gestalt of how well a survey instrument performs in a multitude of settings and populations over a number of years. Construct validity is often thought to comprise two other forms of validity: convergent and divergent.

Convergent

Convergent validity implies that several different methods for obtaining the same information about a given trait or concept produce similar results. Evaluating convergent validity is analogous to measuring alternate-form reliability, except that the former is more theoretical and requires a great deal of work, usually by multiple investigators with different approaches.

[p. 44 ↓]

Divergent

Divergent (discriminant) validity is another theoretically based way of thinking about the ability of a measure to estimate the underlying truth in a given area. For a survey instrument to have divergent validity, it must be shown not to correlate too closely with similar but distinct concepts or traits. This, too, requires much effort over many years of evaluation.

Validity Recap

Testing a survey instrument for construct validity is more like hypothesis testing than like calculating correlation coefficients. Demonstrating construct validity is much more difficult and usually requires a great deal of effort in many different experiments. Construct validity may be said to result from the continued use of a survey instrument to measure some trait, quality, or “construct.” Indeed, over a period of years, the survey instrument itself may define the way we think about the variable. It is difficult to present a specific example of construct validity because its measurement and documentation require such an all-encompassing and multifaceted research strategy.

[p. 45 ↓] The following table summarizes the types of validity and their characteristics along with comments on their use.

Type of Validity	Characteristics	Comments
Face	Casual review of how good an item or group of items appear	Assessed by individuals with no formal training in the subject under study
Content	Formal expert review of how good an item or series of items appears	Usually assessed by individuals with expertise in some aspect of the subject under study

Criterion: Concurrent	Measures how well the item or scale correlates with "gold standard" measures of the same variable	Requires the identification of an established, generally accepted gold standard
Criterion: Predictive	Measures how well the item or scale predicts expected future observations	Used to predict outcomes or events of significance that the item or scale might subsequently be used to predict
Construct	Theoretical measure of how meaningful a survey instrument is	Determined usually after years of experience by numerous investigators

Validity is usually expressed as a correlation coefficient, or *r* value, between two sets of data. Levels of 0.70 or more are generally accepted as representing good validity.

[p. 46 ↓]

http://dx.doi.org/10.4135/9781483348957.n3