

## RELIABILITY AND VALIDITY

The two most important and fundamental characteristics of any measurement procedure are reliability and validity. These two principles will be discussed in turn.

### Reliability

Reliability is defined as the extent to which a questionnaire, test, observation or any measurement procedure produces the same results on repeated trials. In short, it is the stability or consistency of scores over time or across raters. Keep in mind that reliability pertains to *scores* not people. Thus, in research we would never say that someone was reliable. As an example, consider judges in a platform diving competition. The extent to which they agree on the scores for each contestant is an indication of reliability. Similarly, the degree to which an individual's responses (i.e., their scores) on a survey would stay the same over time is also a sign of reliability.

An important point to understand is that a measure can be perfectly reliable and yet not be valid. Consider a bathroom scale that *always* weighs you as being 5 lbs. heavier than your true weight. This scale (though invalid as it incorrectly assesses weight) is perfectly reliable as it *consistently* weighs you as being 5 lbs. heavier than you truly are. A research example of this phenomenon would be a questionnaire designed to assess job satisfaction that asked questions such as, "Do you like to watch ice hockey games?", "What do you like to eat more, pizza or hamburgers?" and "What is your favorite movie?". As you can readily imagine, the responses to these questions would probably remain stable over time, thus, demonstrating highly reliable scores. However, are the questions *valid* when one is attempting to measure job satisfaction? Of course not, as they have nothing to do with an individual's level of job satisfaction. While this example may seem just a tad far-fetched I hope that you grasp the underlying difference between reliability and validity.

### Assessing the Three Aspects of Reliability

There are three aspects of reliability, namely: *equivalence*, *stability* and *internal consistency* (*homogeneity*). It is important to understand the distinction between these three as it will guide one in the proper assessment of reliability given the research protocol. The first aspect, equivalence, refers to the amount of agreement between two or more instruments that are administered at nearly the same point in time. Equivalence is measured through a *parallel forms* procedure in which one administers alternative forms of the same measure to either the same group or different group of respondents. This administration of the various forms occurs at the same time or following some time delay. The higher the degree of correlation between the two forms, the more equivalent they are. In practice the parallel forms procedure is seldom implemented, as it is difficult, if not impossible, to verify that two tests are indeed parallel (i.e., have equal means, variances, and correlations with other measures). Indeed, it is difficult enough to have one well-developed instrument to measure the construct of interest let alone two. Another situation in which equivalence will be important is when the measurement process entails subjective judgments or ratings being made by more than one person. Say, for example, that you are a part of a research team whose purpose is to interview people concerning their attitudes toward educational curriculum for children. It should be self-evident to you that each rater should apply the same standards toward the assessment of the responses. The same can be said for a situation in which multiple individuals are observing behavior. The observers should agree as to what constitutes the presence or absence of a behavior as well as the level to which the behavior is exhibited. In these scenarios equivalence is demonstrated by assessing *interrater reliability* which refers to the consistency with which observers or raters make judgments. The procedure for determining interrater reliability is:

$$\# \text{ of agreements} / \# \text{ of opportunities for agreement} \times 100.$$

Thus, a situation in which raters agree a total of 75 times in 90 opportunities (i.e., unique observations or ratings) produces 83% agreement. ( $75/90 = .83 \times 100 = 83\%$ .)

The second aspect of reliability, stability, is said to occur when the same or similar scores are obtained with repeated testing with the same group of respondents. In other words, the scores are consistent from one time to the next. Stability is assessed through a *test-retest* procedure that involves administering the

same measurement instrument to the same individuals under the same conditions after some period of time. Test-retest reliability is estimated with correlations between the scores at Time 1 and those at Time 2 (to Time x). Two assumptions underlie the use of the test-retest procedure. The first required assumption is that the characteristic that is measured does not change over the time period. The second assumption is that the time period is long enough that the respondents' memories of taking the test at Time 1 does not influence their scores at the second and subsequent test administrations.

The third and last aspect of reliability is internal consistency (or homogeneity). Internal consistency concerns the extent to which items on the test or instrument are measuring the same thing. If, for example, you are developing a test to measure organizational commitment you should determine the reliability of *each* item. If the individual items are highly correlated with each other you can be highly confident in the reliability of the entire scale. The appeal of an internal consistency index of reliability is that it is estimated after only one test administration and therefore avoids the problems associated with testing over multiple time periods. Internal consistency is estimated via the *split-half* reliability index, *coefficient alpha* (Cronbach, 1951) index or the *Kuder-Richardson formula 20* (KR-20)(Kuder & Richardson, 1937) index. The split-half estimate entails dividing up the test into two parts (e.g., odd/even items or first half of the items/second half of the items), administering the two forms to the same group of individuals and correlating the responses. Coefficient alpha and KR-20 both represent the average of all possible split-half estimates. The difference between the two is when they would be used to assess reliability. Specifically, coefficient alpha is typically used during scale development with items that have several response options (i.e., 1 = strongly disagree to 5 = strongly agree) whereas KR-20 is used to estimate reliability for dichotomous (i.e., Yes/No; True/False) response scales. The formula to compute KR-20 is:

$$KR-20 = N / (N - 1)[1 - \text{Sum}(p_i q_i) / \text{Var}(X)]$$

where  $\text{Sum}(p_i q_i)$  = sum of the product of the probability of alternative responses;

and to calculate coefficient alpha:

$$\alpha = N / (N - 1)[1 - \text{sum Var}(Y_i) / \text{Var}(X)]$$

where  $N$  = # items

$\text{sum Var}(Y_i)$  = sum of item variances

$\text{Var}(X)$  = composite variance (Allen & Yen, 1979)

Yes, I know, that's why computers were invented.

Granted, this is probably more than you would ever want to know about reliability but better I provide you with too much information than too little. A couple of questions that you may have at this point are: 1) What is considered a 'good' or 'adequate' reliability value? and 2) How do I improve the reliability of my survey instrument? With respect to the first question, obviously, the higher the reliability value the more reliable the measure. The general convention in research has been prescribed by Nunnally and Bernstein (1994) who state that one should strive for reliability values of .70 or higher. Regarding the second question, reliability values increase as test length increases (see Gulliksen, 1950 for a complete discussion of the relationship between test length and reliability). That is, the more items you have in your scale to measure the construct of interest the more reliable your scale will become. However, the problem with simply increasing the number of scale items when performing applied research is that respondents are less likely to participate and answer completely when confronted with the prospect of replying to a lengthy questionnaire. Therefore, the best approach is to develop a scale that completely measures the construct of interest and yet does so in as parsimonious or economical a manner as is possible. A well-developed yet brief scale may lead to higher levels of respondent participation and comprehensiveness of responses so that one acquires a rich pool of data with which to answer their research question.

## Validity

Validity is defined as the extent to which the instrument measures what it purports to measure. For example, a test that is used to screen applicants for a job is valid if its scores are directly related to future job performance. There are many different types of validity, including: *content validity*, *face validity*, *criterion-related validity* (or *predictive validity*), *construct validity*, *factorial validity*, *concurrent validity*, *convergent validity* and *divergent* (or *discriminant validity*). Not to worry, I will limit this discussion to the first four.

Content validity pertains to the degree to which the instrument fully assesses or measures the construct of interest. For example, say we are interested in evaluating employees' attitudes toward a training program within an organization. We would want to ensure that our questions fully represent the domain of attitudes toward the training program. The development of a content valid instrument is typically achieved by a rational analysis of the instrument by raters (ideally 3 to 5) familiar with the construct of interest. Specifically, raters will review all of the items for readability, clarity and comprehensiveness and come to some level of agreement as to which items should be included in the final instrument.

Face validity is a component of content validity and is established when an individual reviewing the instrument concludes that it measures the characteristic or trait of interest. For instance, if a quiz in this class comprised items that asked questions pertaining to research methods you would most likely conclude that it was face valid. In short, it looks as if it is indeed measuring what it is designed to measure.

Criterion-related validity is assessed when one is interested in determining the relationship of scores on a test to a specific criterion. An example is that scores on an admissions test for graduate school should be related to relevant criteria such as grade point average or completion of the program. Conversely, an instrument that measured your hat size would most assuredly demonstrate very poor criterion-related validity with respect to success in graduate school.

Construct validity is the degree to which an instrument measures the trait or theoretical construct that it is intended to measure. For example, if one were to develop an instrument to measure intelligence that does indeed measure IQ, then this test is construct valid. Construct validity is very much an ongoing process as one refines a theory, if necessary, in order to make predictions about test scores in various settings and situations.

## Conclusion

In conclusion, remember that your ability to answer your research question is only as good as the instruments you develop or your data collection procedure. Well-trained and motivated observers or a well-developed survey instrument will better provide you with quality data with which to answer a question or solve a problem. Finally, be aware that reliability is *necessary* but not *sufficient* for validity. That is, for something to be valid it must be reliable but it must also measure what it is intended to measure.

## REFERENCES

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, 151-160.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.