Reliability Theory

Reliability and validity are two major requirements for any measurement. Validity pertains to the correctness of the measure; a valid tool measures what it is supposed to measure. Reliability pertains to the consistency of the tool across different contexts. As a rule, an instrument's validity cannot exceed its reliability, although it is common to find reliable tools that have little validity.

There are three primary aspects to reliability: (a) A reliable tool will give similar results when applied by different users (such as technicians or psychologists). (b) It will also yield similar results when measuring the same object (or person) at different times. In psychometrics, reliability also implies a third feature, which is relevant to scales (measures that include various submeasures or items). Specifically it entails the requirement that (c) all parts of the instrument be interrelated.

1. *Interrater Reliability*: Some measurements require expertise and professional judgments in their use. The reliability of such a tool is contingent on the degree to which measurements of the same phenomenon by different professionals will yield identical results. What we want to avoid is a test that essentially relies on undefined judgments of the examiner, without concrete criteria that are clearly spelled out.

Statistically, this aspect of reliability is usually determined by having several raters measure the same phenomena, and them computing the correlations between the different raters. For typical measures that yield numerical data, a correlation index needs to be high (e.g., in the .90s) to demonstrate good interrater reliability.

2. *Test-Retest Reliability*: This feature, often referred to as *temporal stability,* reflects the expectancy that the measurement of a specific object will yield similar results when it is measured at different times. Clearly, this is based on an assumption that one does not expect the object to be changing between the two measurements. In fact, this may not be the case for many constructs. (Consider, for example, blood pressure or stress, both of which would be expected to vary from one time to another—even for the same person.) Specifically within psychology, it is important to understand that this aspect of reliability pertains only to measures that refer to traits (aspects of personality that are constant regardless of environmental events or context); it does not pertain to states (specific aspects of behavior or attitudes that vary based on situations and interactions at the moment).

Statistically, this aspect of reliability is usually determined by having a group of people measured twice with the instrument. The time interval can vary, based on specific studies, from several weeks to a year or two. Unfortunately, testing experts often choose a short duration between tests (there are published test-retest time periods of only 6 hours!), which make their claim of testing an actual trait equivocal. Correlations are computed between the two trials. For typical measures that yield numerical data, a correlation index needs to be high to demonstrate good interrater reliability (although .70 would be sufficient).

3. *Internal Consistency*: This usually entails the reliability of measures that have multiple items; such measures are known as scales.

Consider, for example, a scale of irritability. Such a scale might include 20 items that focus on different aspects of this construct. One might ask how often the person gets into arguments, another might look into how often the person changes doctors, whereas a third could ask about the person's rejection patterns of potential dates. Obviously, the experts who constructed this scale believed that all of these items are indicative of irritability. Indeed, it may well be true that irritable people would show high scores on all of these items. However, it is feasible that the major reason most people reject dates differs significantly from the reason people change doctors.

Another aspect of the problem of internal consistency might be understood from the perspective of the meaning of the score on a scale. Take, for example, the case of a firm that is interested in hiring interviewers

who will be pleasant to potential customers. Such a firm might decide to use a 100-item scale of pleasantness, which contains imagined scenarios that are presented to respondents, each having a specific response that is considered "pleasant." In the real world, the firm cannot expect to find sufficient interviews that will score 100, so the firm decides to hire those candidates who score highest. Say that two candidates scored 98; one of the respondents "failed" two items indicating that she would ask loud people to keep it down in a theater and that she would not allow someone to cut in line in front of her at a bank. The other respondent "failed" two items indicating that she has been a party to a lawsuit and that she would hang up on a telemarketer. By assigning these respondents equal scores, the instrument implies that they are at an equivalent level of pleasantness. But what evidence is there that the items they failed are just as meaningful in terms of the overall scale? Furthermore, is it possible that there are different factors in unpleasantness, and that these items don't really speak to the same construct?

There are several methods of establishing internal consistency, all of which are based on intercorrelations between items. One method is called split half, where the scale is divided into different sets, each containing half the items, and the halves are then tested for an acceptable relationship using the correlation statistic. Often, item-total correlations are used to establish internal consistency (that is, correlations between each of the items and the total scale score). Typically, an item must correlate .3 or higher with the total score in order to remain part of the scale. A popular overall statistic that takes into account all of the possible item intercorrelations is Cronbach's alpha (alternatively, the Kuder Richardson-20 formula for dichotomous items), where an alpha of .70 or higher is considered acceptable to establish this form of reliability.

Scaling issues that are related to internal consistency often are found in weighting procedures. Some tests contain items that are given more weight in the total score than others. These features introduce statistical complications in the process of establishing consistency, which are often solved through multiple regression methods.

All in all, reliability is a major part of the preparatory work that goes into scale construction, often in conjunction with factor analysis. It is not featured emphatically in test descriptions, but poor reliability will doom a test's validity and its usefulness to measure anything meaningful.

—Samuel Juni

Further Reading

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: Authors.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology, 78,* 98–104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85–98.

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement.* London: Pinter.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22,* 358–376.

Repeated Measures Analysis of Variance

In a repeated measures analysis of variance, we are faced with the task of comparing means of groups that are dependent. Unlike the usual analysis of variance (ANOVA), where the groups are independent, in repeated measures ANOVA, the groups and the group means are dependent. Because the group means are dependent, we must adjust the usual statistical and inferential processes to take the dependencies into account. Before going on with our discussion of repeated measures ANOVAs, let's consider two situations that are likely to yield correlated means.