25

Reliability and Validity of Instruments

Thomas R. Knapp and Ralph O. Mueller

Both reliability and validity are essential parts of the psychometric properties of a measuring instrument.[1] The *reliability* of an instrument is concerned with the consistency of measurements: from time to time, from form to form, from item to item, or from one rater to another. On the other hand, the *validity* of an instrument is usually defined as the extent to which the instrument actually measures "what it is designed to measure" or "what it purports to measure." Validity is therefore concerned with the relevance of an instrument for addressing a study's purpose(s) and research question(s). Both reliability and validity are context-specific characteristics: for example, researchers are often interested in assessing if a measure remains reliable and valid for a specific culture, situation, or circumstance (e.g., a psychological test might be highly reliable and valid in a population of Caucasian adults but not in one of African American children). The conceptualization and specific definitions of reliability and validity have changed over time, as reflected in the various editions of *Educational Measurement* (Cronbach, 1971; Cureton, 1951; Feldt & Brennan, 1989; Haertel, 2006; Kane, 2006; Messick, 1989; Stanley, 1971; Thorndike, 1951). Table 25.1 contains a list of desiderata regarding reliability and validity of measuring instruments that should be followed in any empirical research report.

1. Instrument Description and Justification

Empirical data for analysis during research studies are collected with the aid of measuring instruments, be they laboratory equipment or, more common in the social and behavioral sciences, surveys, achievement batteries, or psychological tests. Because study results should only be trusted when investigators collect good data, authors should ensure that readers can judge the "goodness" of the data for themselves. Thus, at a minimum, a full description of the instrument(s) used is necessary (and should be followed by an assessment of reliability and validity; see Desiderata 2 and 3, respectively), including the purpose(s) and intended use(s), item format(s), and scales of measurement (i.e., nominal, ordinal, interval, or ratio). Obviously, a specific instrument is appropriate for use in some contexts, but not necessarily in others (e.g., a high school reading test is likely to be inappropriate to measure a middle schooler's intelligence). Authors must take care in justifying the choice of instrument(s) and making explicit the link to the study's purpose(s) and research question(s). Often, the description and justification for a particular instrument is presented in a manuscript's Instrumentation subsection of the Methods section but could also be accomplished in the Introduction.

Table 25.1 Desiderata for Reliability and Validity of Instruments

Desideratum	*Manuscript Section(s)**
1. Each instrument used in the study is described in sufficient detail. The appropriateness of the instrument to address the study's purpose(s) and research question(s) is made explicit.	I, M
2. Appropriate reliability indices are considered: The study's purpose(s) guide the choice of indices calculated from current data and/or examined from previous research.	M, R
3. Suitable validity evidence is gathered: The study's purpose(s) determine the type of validity support gathered from current data and/or consulted from related literature.	M, R
4. Applicable reliability and validity evidence is reported and interpreted. The study's conclusions are placed within the context of such evidence (or lack thereof).	D

* *Note*: I = Introduction, M = Methods, R = Results, D = Discussion

2. Reliability Indices

Several approaches exist to assess an instrument's reliability, whose appropriateness is dependent on a study's specific purpose. Four traditional strategies often found in the literature are briefly discussed below: test-retest, parallel forms, internal consistency, and rater-to-rater. All four are based on classical test theory, but alternative conceptualizations of reliability exist that are based on other analytical frameworks: generalizability theory (see Chapter 9, this volume), item response theory (see Chapter 12, this volume), and structural equation modeling (see Chapter 28, this volume).

Test-Retest Reliability. If a study's purpose is to assess measurement consistency of one instrument from one time point to another, a straightforward way to collect reliability evidence is to measure and then re-measure individuals and determine how closely the two sets of measurements are related (i.e., the *test-retest* method). In studies assessing psychological constructs such as attitude, a question with often serious ramifications is how much time should be allowed between the first and second testing. If the interval is too short, measurement consistency might only be due to the fact that individuals being tested "parrot back" the same responses at Time 2 that they gave at Time 1. If the interval is too long, some items might no longer be developmentally appropriate (e.g., academic achievement items on a middle school test administered to students entering high school) which could impact the validity of the instrument as well. Even in studies with physical variables, the length of time between measurements might be crucial for some (e.g., repeated weight measurements during a health-awareness program might fluctuate widely, depending on weight loss/gain), but not for other variables (e.g., repeated height measurements of adult participants are likely to remain consistent, irrespective of the time-lag between measurements). In general, authors should defend their choice of time intervals between measurements as the "correct" amount of time is situation specific and somewhat subjective. An assessment of test-retest reliability can be accomplished in either an absolute manner (e.g., the median difference between corresponding measurements) or relative manner (e.g., the correlation between the two sets of measurements) with the latter approach being more common than the former.

Parallel Forms Reliability. If a measuring instrument is available in two *parallel* (i.e., psychometrically equivalent and interchangeable) *forms*, say Form A and Form B, with measurements having been taken on both forms, reliability evidence can be obtained by comparing the scores on Form A with the scores on Form B, again either absolutely or relatively. The time between the administrations of the two forms is still an important consideration, but because the forms are not identical there is no longer the concern for "parroting back" if the time interval is short.

Internal Consistency Reliability. Given the disadvantage of multiple test administrations for test-retest and parallel forms reliability (e.g., increased costs, time lag between measurements, and missing data due to non-participation during the second testing), a commonly used alternative is the estimation of the *internal consistency* of an instrument. Here, an instrument consisting of multiple items measuring the same construct is administered only once, but now treating the items as forming two parallel halves of the instrument. The two half-forms are created after the actual measurement, traditionally by considering the odd-numbered items as one form and the even-numbered items as the other form (though other ways to split an instrument are certainly possible, e.g., random assignments of items to halves). The scores on the two forms are then compared, usually relatively by computing the correlation between the scores on the odd numbered items with the scores on the even numbered items. But the correlation must be "stepped up" by using the Spearman-Brown formula (Brown, 1910; Spearman, 1910), in order to estimate what the correlation might be between two full-forms as opposed to two half-forms. That estimate is obtained by multiplying the correlation coefficient by two and then dividing that product by one plus the correlation. The type of reliability evidence thus produced is strictly concerned with internal consistency (from half-form to half-form) since time has not passed between obtaining the first set of measurements and obtaining the second set of measurements.

Another type of internal consistency reliability is from item to item within a form. Such an approach was first advocated by Kuder and Richardson (1937) for dichotomously scored test items, and was subsequently extended to the more general interval measurement case by Cronbach (1951). Their formulae involve only the number of items, the mean and variance of each, and the covariances between all of the possible pairs of items. Cronbach called his reliability coefficient *alpha*. It is still known by that name and is by far the most commonly employed indicator of the reliability of a measuring instrument in the social sciences.[2]

Rater-to-Rater Reliability. When the data for a study take the form of ratings from scales, the type of reliability evidence that must be obtained before such a study is undertaken is an indication of the extent to which a rater agrees with him(her) self (*intra-rater reliability*) and/or the extent to which one rater agrees with another (*inter-rater reliability*). Several options exist to assess rater-to-rater consistency, with the *intraclass coefficient* and Cohen's *kappa* (1960) being among the most popular (see Chapter 11, this volume).

Norm- vs. Criterion-Referenced Settings. Literature devoted to reliability assessment within norm-referenced versus criterion-referenced frameworks is plentiful. Most users of *norm-referenced* tests—where scores are primarily interpreted in relation to those from an appropriate norm or comparison group—have adopted approaches to reliability assessment similar to those summarized thus far, with particular emphasis on correlations that are indicative of relative agreement between variables. *Criterion-referenced* (or *domain-referenced*) *measurement* is concerned with what proportion of a domain of items has been answered successfully and whether or not that portion constitutes a "passing" performance (e.g., "John spelled 82% of the words on a spelling test correctly, which was below the cut point for progressing to the next lesson."). Here, reliability assessment concentrates on measurement errors in the vicinity of the cut point with a particular interest on the reliability of the pass-or-fail *decision*. In parallel-form situations, for example, the matter of whether a person passes both Form A and Form B or fails both Form A and Form B takes precedence over how high the correlation is between the two forms.

3. Validity Evidence

In physical science research the usual evidence for the validity of measuring instruments is expert judgment and/or validity-by-definition with respect to a manufacturer's specifications. For example, "For the purpose of this study, body temperature is the number of degrees Fahrenheit that the Smith thermometer reads when inserted in the mouths of the persons on whom the measurements are being taken." As evidence of validity, the researcher might go on to explain that the Smith thermometer is regarded as the 'gold standard' of temperature measurement.

In the social and behavioral sciences, investigators are often urged to provide evidence for *content validity* (expert judgments of the representativeness of items with respect to the skills, knowledge, etc. domain to be covered), *criterion-related validity* (degree of agreement with a "gold standard"), and/or *construct validity* (degree of agreement with theoretical expectations) of the measuring instruments used in their substantive studies. More recently, all three validity types have been subsumed under an expanded concept of construct validity, but not without controversy. Whatever conceptualization is used, researchers must be clear that instrument validity is *not* context free: a measure might be valid in one situation or for one population, but not in or for another (e.g., the Scholastic Aptitude Test, SAT, is often argued as being valid to assess high school seniors' potential for success in undergraduate higher education, but not for measuring their intelligence or potential to succeed in vocational training).

Criterion-related Validity. When a measure is designed to relate to an external criterion, its validity is judged by either *concurrent* or *predictive* assessments (i.e., degrees to which test scores estimate a specified present or future performance). For example, a passing score on a driver's permit test with acceptable concurrent validity will allow the test taker to immediately drive motor vehicles, assuming an associated road test has been passed. On the other hand, evidence of predictive criterion-related validity is often helpful for judging instruments that are designed to measure aptitude, with passing achievement scores serving as the standards for whether or not the aptitude tests are predictive of achievement. But, herein also lies an interesting dilemma: How does one know that the achievement tests themselves are valid? Do the standards need to be validated against an even higher standard? Or, if the standards' validity is established by expert judgment, why not appeal to experts directly for validity assessments of the aptitude measure? Furthermore, if expert judgment is to be the ultimate arbiter, who are the experts and who selects them?

Construct Validity. In order to judge the degree to which a theoretical construct accounts for test performance, a researcher must assess the test's construct validity. Supportive evidence usually comes from exploratory or confirmatory factor analyses (see Chapter 8, this volume) in which the dimensionality and the degree of correlation of the variables comprising the instruments are investigated. The most popular approach is the *convergent/discriminant* strategy first recommended by Campbell and Fiske (1959): researchers determine the extent to which measurements obtained with the instruments in question correlate with variables with which they are theoretically expected to correlate (convergent) and the extent to which those measurements correlate with other variables with which they are theoretically *not* expected to correlate (discriminant). See also Chapter 22, this volume, on multitrait-multimethod analysis.

4. Reporting and Interpreting Reliability and Validity Results

Before reporting a study's main findings, investigators should discuss evidence of the reliability and validity of the instrument(s) used. Ideally, such evidence should come from both a thorough search of the related literature and an assessment based on current study participants. A comparison of present reliability and validity information with that gleaned from related literature is helpful to readers, especially when such information might be contradictory, as, for example, when earlier reliability/validity evidence could not be reproduced based on the current sample's data.

Certainly when no previous reliability and validity information is available—as is the case when investigators construct their own instruments— authors must report psychometric properties of the instrument(s) based on an analysis of the current data. But even if reliability/validity evidence is identified from previous studies, it is often the case that it does not generalize to the current population under study. Thus, it is incumbent upon each investigator to provide a thorough justification for why the instruments used are appropriate for the current sample of participants.

In the social and behavioral sciences, reliability and validity coefficients in the .70 or .80 or above range are often considered acceptable with values below these cut-offs being acknowledged as study limitations. However, the acceptability of coefficients should be judged with caution as value adequacy certainly depends on the particular phenomenon under study. Nevertheless, the interpretation of main results should commence from within the context of reliability and validity results as unreliability and/or invalidity usually attenuate the magnitudes of expected findings and lead to wider confidence intervals and less likelihood of the detection of effects and relationships in the data.

Notes

1 This chapter does not deal with the internal and external validity and reliability of a chosen research design (but see Chapter 26, this volume). Also, in fields outside the social/behavioral sciences, validity and reliability are sometimes known by different names. For example, in epidemiology, *reproducibility* is generally preferred over reliability. In engineering and related disciplines, equipment is said to be reliable if it does not, or is very unlikely to, break down. Also, the ambiguous term *accuracy* is sometimes used in lieu of either reliability or validity.

2 Kuder and Richardson actually derived several formulae for internal consistency by making successively relaxed assumptions, and they numbered them accordingly. The formula that is most frequently used to compute Cronbach's alpha is actually a direct extension of Kuder and Richardson's Formula #20 for dichotomous data.

References

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin*, *56*, 81–105.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 443–507). Washington, DC: American Council on Education.

Cureton, E. F. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). New York: Macmillan.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 65–110). Westport, CT: American Council on Education/Praeger.

Kane, M. L. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13–103). Washington, DC: American Council on Education.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*, 171–195.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 356–442). Washington, DC: American Council on Education.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 560–620). Washington, DC: American Council on Education.