Reliability

Reliability and validity, the topics of this and the next chapter, are twins and cannot be completely separated. These two concepts comprise the dual "holy grail" of research, and outside of the central importance of theory, they are crucial to any sort of meaningful research. Without reliability and validity, research is nonsense. Many forms of reliability and of validity have been identified, perhaps to the point that the words themselves have been stretched thin.

Reliability refers to consistency of measurement and takes several forms: whether a construct is measured in a manner that is stable over time; are the items in a test congruent with each other; are two supposedly identical tests really the same; are people performing ratings that agree with each other. Validity is a deeper concept: does the instrument measure the right construct; is the experiment constructed so that the independent variable actually represents the theoretical construct; is it really the independent variable that caused changes in the dependent variable; and more.

Four kinds of reliability are usually identified in behavioral research: test-retest, parallel forms, internal consistency, and inter-rater.

Test-Retest Reliability

A good test (or "instrument") is stable over repeated administrations. An IQ test given to the same person three times in one month should reveal the same IQ score, more or less, each time. Likewise, any adequate measure of personality, attitudes, skills, values, or beliefs should come out the same, time after time, unless the person him or herself actually changes in either a short-term or long-term sense. Some degree of variability is expected due to temporary changes in the person, the situation, or just plain random factors like luck, but in general the scores should be consistent.

The test-retest reliability of a measure is estimated using a reliability coefficient. A reliability coefficient is often a correlation coefficient calculated between the administrations of the test. (Correlation coefficients are described in the SPSS-Basic Analyses chapter. See sidebar for a quick explanation of correlations.) A typical research study of test-retest reliability would administer a test to a sample of 100 people, wait a month, then readminister it to the same people. The correlation between the two administrations is an indicator of the instrument's reliability.

Parallel Forms Reliability

"Forms" are alternate versions of the same test. We use the terms "Form A" and "Form B" sometimes to identify such versions of a test. Parallel forms are forms that really do measure the same thing, that is, are equivalent. Parallel forms of a test are developed to be used in situations where we must obtain essentially the same information from people at several different but close-together times and we don't want their exposure to the test at one time to affect their responses other times. This "carry over" effect can occur due to practice (on a skills-oriented test) or remembering the items from a previous administration.

For example, the author's research on international student adjustment required obtaining psychological adjustment measures from students every week for 10 weeks. At some point, it was expected that the students would stop thinking about the test questions and just answer automatically. To try to avoid this, two forms of the adjustment measure were developed and administered on alternate weeks.

Parallel forms reliability is measured using the correlation coefficient between the forms. In a typical study, a group of 100 students would take all forms of the test. The correlations among the forms represent the extent of parallel forms reliability.

Correlation Coefficient

Correlation refers to the relationship or association between two variables within a sample. For example, we would expect to find a correlation between height and weight in a sample of people: generally, taller people weigh more. The relationship can range from very strong to nonexistent. Height and weight show a fairly strong relationship, although of course there are exceptions such as tall thin people and short wide people. On the other hand, weight and IQ show no relationship. Sometimes, as one variable (height) increases, another variable (weight) also increases, evidencing a "positive relationship." Other times, we see a negative relationship, such as between smoking cigarettes (packs per week) and life expectancy (years). All of these relationships are examined within groups of people and refer to the general trend present in the group, not to individuals.

The correlation coefficient is a statistical estimation of the strength of relationship. (A more formal, theoretical explanation of correlation is presented in the SPSS-Basic Analyses chapter.) Correlation coefficients are symbolized by the letter "r" and range from 1 to 0 to -1. At r=0, there is no relationship, as in weight and IQ. At r=1 there is a perfect positive relationship: as one variable increases, the other one increases in perfect lockstep. Such relationships are rare in the real world. At r=-1, there is a perfect negative relationship, also rare.

Correlation coefficients range anywhere along the scale from 1 to -1. For example, the relationship between people's personality traits and their behaviors tends to be around r=.30 (not so good). The correlation between the SAT-Math and overall university GPA among psychology undergraduates here is r=.53, a moderately good relationship.

Internal Consistency Reliability

Often it is important that the individual items in a test measure the same thing, or almost measure the same thing. If items are not consistent with each other, the overall test score will reflect several different underlying constructs and won't make any sense. For example, a measure of attitudes toward pizza might include several differ qualities of pizza. These items should be sufficiently similar to each other so that adding up the item responses produces a total score that focuses on the intended construct.

1. Pizza tastes good.
   Strongly Agree - Somewhat Agree - Neutral - Somewhat Disagree - Strongly Disagree
2. I feel hungry when I think of pizza.
3. Pizza is inexpensive.
4. Pizza is easy to get.
5. Pizza is popular in New York.

Items 1 and 2 would probably have high internal consistency reliability. Items 1, 2, 3, and 4 would probably have moderately high reliability. Item 5 really does not fit this test and would show little internal consistency reliability with the other items.

Internal consistency reliability is measured using special types of correlation coefficients termed Cronbach's Alpha and the Kuder-Richardson Coefficient. When researchers make new tests and examine aspects of the test such as the internal consistency reliability of the items, the procedure is termed "item analysis."

Inter-Rater Reliability

A somewhat different sort of reliability is at issue when the same stimulus (person, event, behavior, etc.) must be rated by more than one rater. For example, in studies of the relationship between physical attractiveness and social development, the researchers need to know how attractive the person is. (Research of this kind asks questions such as, "do prettier people develop better social skills?") How can this rating be done? Calculate the ratio of length of nose to distance between the ears? While some such physical indexes of attractiveness have been developed, the most common way is to assemble a panel of "judges" to rate the "stimuli." (Sounds like figure skating judging, but there is no bribery involved.) The researcher needs to look at the extent to which the raters agree on their ratings. When inter-rater reliability is low, the researcher has to wonder if it is possible to classify persons on a dimension such as attractiveness or whether his or her attempts to do so have failed.

Inter-rater reliability is measured in several ways, such as the percentage of agreement among the judges or a correlation-like coefficient called Kappa.

A Research Example

Here's a semi-hypothetical example that brings these reliability issues together. Of course, a complete example would require considering theory and reliability's twin, validity.

The researcher wants to know if personality traits are related to sexual activity. She cannot perform a true experiment because personality traits cannot be controlled or manipulated, so she must be content with simply measuring the traits and assessing sexual behavior. The trait of interest is Self-Monitoring, a blend of extraversion, willingness to self-present, self-presentation abilities, and low social anxiety. High Self-Monitors are outgoing people who know how to act in various situations to get what they want, and do so. A measure of sexual behavior is created just for this study, the "Sexual Activity Test" (SAT).

The Self-Monitoring Scale (SMS) already exists, so she will use it as it is. She must, however, construct the SAT from scratch. The details of how she would actually do this are complicated, but in the end she has a 20-item test that assesses various interpersonal sexual behaviors. The researcher theorizes that, overall, SAT is a "unitary construct." A unitary construct has one, central idea or dimension rather than several sub-dimensions. For example, some psychologists believe that IQ may not be a unitary construct, arguing that there are several different kinds of intelligence. If true, then a single IQ score is meaningless.

To determine if the SAT assesses a unitary construct, the researcher gives the initial version of the test to a large sample of people who will not be in the real study, then performs an item analysis. She looks at the internal consistency reli-

ability of the 20 items to see if they "hang together." (She also does some other things that are beyond our interest here.) Coefficient alpha, a common measure of internal consistency, turns out to be α=.45. This is too low, indicating that the items are not measuring the same thing. She has two choices: (1) give up on the idea that the SAT will assess a unitary construct and try to find the two or more sub-dimensions that represent interpersonal sexual activity; or (2) find the bad items and get rid of them. She chooses the latter. A bad item is one that is poorly related to the other items, sort of like a human who refuses to fit in to a social group. To find the "bad" items, she correlates each item with the total score (the average of all the items). These 20 correlations are termed "item-total correlations." She looks for items with a poor relationship to the total score and eliminates them from the SAT. Then she recalculates coefficient alpha on the new, smaller test and, hooray, it is now α=.85 (very good). As the Japanese say: "the nail that sticks up gets pounded down."

Next, she want to make sure that the SAT is stable over time. She finds still another sample of people and gives them the SAT twice, one month apart. The correlation between time 1 and time 2, the test-retest reliability, turns out to be r=.70. This is very good given the fact that people do change over time, and some instability in the test is expected for this reason rather than due to the test's qualities.

The researcher wants to do a longitudinal study of Self-Monitoring and sexual behavior so she can figure out whether these two variables change over time, if the relationship between them changes over time, and possibly which one causes the other. A longitudinal study measures the same thing over time. She has some fear that giving the same tests several times will reduce the quality of her results because her research subjects will remember the items from administration to administration and answer automatically without thinking. She needs at least two versions of each test. To produce these versions, she creates two SMSs by dividing up the items using the item analysis information published some time ago by the brilliant young psychologists Gabrenya and Arkin (1980). She does the same for the SAT. To make sure that she has parallel forms, she gives all four tests (two SMSs and two SATs) to another sample of people and calculates the correlation between the parallel forms. Because she is such a meticulous researcher and because the earlier work of Gabrenya and Arkin was so fine, she obtains parallel forms reliability coefficients that are good, r=.75 for SMS and r=.68 for SAT.

Now, finally, she is ready to perform her research. She selects a sample of 50 males and 50 females in the range 22-25 years old, all unmarried, and administers her two tests to them four times, once every six months. She gives Form A of the SMS and Form A of the SAT the first time, Form B the second time, Form A the third time, etc. Unfortunately, she gets very "noisy" results: the correlations between SAT and SMS are in the right direction, but low.

She concludes that something else is affecting sexual activity and, based on other social psychology research, theorizes that it is the physical attractiveness of the subjects. She must now evaluate each subject's physical attractiveness. She brings all 100 into her lab and takes professional quality photos of them from "a variety of angles." Then she assembles a panel of two men and two women in the 22-25

age range and has them rate each photo on attractiveness. To make sure they are producing a reliable measure, she looks at the agreement rates among the four raters. Coefficient Kappa comes out to be K=.65, which is good enough. Now she can use the attractiveness ratings to lower the noise (error variance) in her data.

How would such a study actually come out? This particular study has not been performed, but components of it have. Self Monitoring does predict higher sexual activity, and attractive people do get more dates. Based on other studies, we would predict that SMS would cause sexual activity, not the opposite. Adding attractiveness would undoubtedly strengthen the results of the study.

Reference

Gabrenya, W. K., Jr., & Arkin, R. M. (1980). Self-Monitoring Scale: Factor structure and correlates. *Personality and Social Psychology Bulletin, 6*, 13-22.