Three Criteria for Good Measurement

The two major criteria for evaluating measurements are reliability and validity.

Reliability

Reliability is an indicator of a measure’s internal consistency. Consistency is the key to understanding reliability. A measure is reliable when different attempts at measuring something converge on the same result. If a professor’s marketing research tests are reliable, a student should tend toward consistent scores on all tests. In other words, a student that makes an 80 on the first test should make scores close to 80 on all subsequent tests. Another way to look at this is that the student who makes the best score on one test will exhibit scores close to the best score in the class on the other tests. If it is difficult to predict what students would make on a test by examining their previous test scores, the tests probably lack reliability. When a measuring process provides reproducible results, the measuring instrument is reliable.

- **INTERNAL CONSISTENCY**

Internal consistency represents a measure’s homogeneity. An attempt to measure trust may require asking several similar but not identical questions, as shown above. The set of items that make up a measure are referred to as a battery of scale items. Internal consistency of a multiple-item measure can be measured by correlating scores on subsets of items making up a scale.

- **TEST-RETEST RELIABILITY**

The test-retest method of determining reliability involves administering the same scale or measure to the same respondents at two separate times to test for stability. If the measure is stable over time, the test, administered under the same conditions each time, should obtain similar results. Test-retest reliability represents a measure’s repeatability. Suppose a researcher at one time attempts to measure buying intentions and finds that 12 percent of the population is willing to purchase a product. If the study is repeated a few weeks later under similar conditions, and the researcher again finds that 12 percent of the population is willing to purchase the product, the measure appears to be reliable. High stability correlation or consistency between two measures at time 1 and time 2 indicates high reliability.
Validity

Good measures should be both precise and accurate. Reliability represents how precise a measure is in that the different attempts at measuring the same thing converge on the same point. Accuracy deals more with how a measure assesses the intended concept.

**Validity** is the accuracy of a measure or the extent to which a score truthfully represents a concept. Achieving validity is not a simple matter. The opening vignette describes this point. The job performance measure should truly reflect job performance. If a supervisor’s friendship affects the performance measure, then the scale’s validity is diminished. Likewise, if the performance scale is defined as effort, the result may well be a reliable scale but not one that maximizes validity. Effort may well lead to performance but effort probably does not equal performance.

Another example of a validity question might involve a media researcher who wonders what it means when respondents indicate they have been *exposed* to a magazine. The researcher wants to know if the measure is valid. The question of validity expresses the researcher’s concern with accurate measurement. Validity addresses the problem of whether a measure (for example, an attitude measure used in marketing) indeed measures what it is supposed to measure. When a measure lacks validity, any conclusions based on that measure are also likely to be faulty.

Students should be able to empathize with the following validity problem. Consider the controversy about highway patrol officers using radar guns to clock speeders. A driver is clocked at 75 mph in a 55 mph zone, but the same radar gun aimed at a house registers 28 mph. The error occurred because the radar gun had picked up impulses from the electrical system of the squad car’s idling engine. The house wasn’t speeding—and the test was not completely valid.

**ESTABLISHING VALIDITY**

Researchers have attempted to assess validity in many ways. They attempt to provide some evidence of a measure’s degree of validity by answering a variety of questions. Is there a consensus among my colleagues that my attitude scale measures what it is supposed to measure? Does my measure correlate with other measures of the same concept? Does the behavior expected from my measure predict actual observed behavior? The three basic approaches to establishing validity are *face* or *content validity*, *criterion validity*, and *construct validity*. 
Face (content) validity refers to the subjective agreement among professionals that a scale logically reflects the concept being measured. Simply, do the test items make sense given a concept’s definition? When an inspection of the test items convinces experts that the items match the definition, the scale is said to have face validity.

Clear, understandable questions such as “How many children do you have?” generally are agreed to have face validity. In scientific studies, however, researchers generally prefer stronger evidence because of the elusive nature of attitudes and other marketing phenomena. For example, the ACNielsen television rating system is based on the PeopleMeter system, which mechanically records whether a sample household’s television is turned on and records the channel selection. If one of the viewers leaves the room or falls asleep, the measure is not a valid measure of audience.

Criterion validity addresses the question, “Does my measure correlate with measures of the similar concepts or known quantities?” Criterion validity may be classified as either concurrent validity or predictive validity depending on the time sequence in which the new measurement scale and the criterion measure are correlated. If the new measure is taken at the same time as the criterion measure and is shown to be valid, then it has concurrent validity. Predictive validity is established when a new measure predicts a future event. The two measures differ only on the basis of a time dimension—that is, the criterion measure is separated in time from the predictor measure.

A practical example of predictive validity is illustrated by a commercial research firm’s test of the relationship between a rough commercial’s effectiveness (as determined, for example, by recall scores) and a finished commercial’s effectiveness (also by recall scores). Ad agencies often test animatic rough, or live-action rough commercials before developing actual finished commercials. One marketing research consulting firm suggests that this testing has high predictive validity. Rough commercial recall scores provide correct estimates of the final finished commercial recall scores more than 80 percent of the time.

Construct validity exists when a measure reliably measures and truthfully represents a unique concept. Construct validity consists of several components, including
Reliability versus Validity

Suppose an expert sharpshooter fires an equal number of rounds with a century-old rifle and a modern rifle.

The shots from the older gun are considerably scattered, but those from the newer gun are closely clustered. The variability of the old rifle compared with that of the new one indicates it is less reliable. The target on the right illustrates the concept of a systematic bias influencing validity. The new rifle is reliable (because it has little variance), but the sharpshooter’s vision is hampered by glare. Although shots are consistent, the sharpshooter is unable to hit the bull’s-eye.