Measuring Subjective Outcomes
Rethinking Reliability and Validity
Tom A. Elasy, MD, MPH, Gary Gaddy, PhD

Reliability and validity are criteria used to assess metric adequacy and are typically quantified by correlation coefficients. Reliability is described as the extent to which repeated measurements yield consistent results. Validity is described as the extent to which a measure actually measures what it purports to measure. These conceptualizations are less useful when applied to measures of subjective outcomes because they do not convey other influences that “drive” correlation coefficients. Consistency is a manifestation of a reliable instrument but does not ensure that an instrument is reliable. Establishing the validity of an instrument is a complex process that is heavily dependent on an investigator’s hypothesis. Hence, validity coefficients may be more a reflection of hypothesis adequacy than of the extent to which instruments measure what they purport to measure. Appreciating how coefficients are influenced will better enable clinicians to assess the adequacy of subjective outcome measures.

KEY WORDS: outcome; measures; reliability; validity.

Traditionally, physicians have been called on to promote physical healing. Reducing tangible negative outcomes, especially mortality, has been a primary objective. Recently, clinicians have begun to focus on more subjective outcomes such as quality of life, social health, pain, and patient satisfaction. Measuring subjective outcomes is difficult.

Psychometrics, the process of quantifying subjective outcomes, has traditionally been the purview of psychologists and educators who wrestle with measuring concepts like personality and intelligence. While not attempting to be a primer in psychometrics, this article will explicate two criteria used to assess subjective outcome measures. Reliability and validity assess the adequacy of any metric. Understanding how these criteria are estimated will better enable clinicians to evaluate the quality of instruments reported in the medical literature.

Our goal is to demonstrate what “drives” the numbers used to convey reliability and validity.

RELIABILITY AND ACCURACY

Reliability is commonly defined as the extent to which repeated measurements yield consistent results. Validity is often defined as the extent to which a measurement actually measures what it purports to measure. Introductory texts in clinical and social research depict “shots” at a target to convey the meaning of reliability and validity (Fig. 1). Reliability is illustrated by the scatter of shots, with greater scatter indicating lesser reliability. Validity is illustrated by how well centered are the shots on the bull’s-eye—the truth. When considering physiologic outcome measures such as a sphygmonanometer measure of blood pressure, these definitions and illustrations are useful. This conceptualization of reliability and validity is less accurate when applied to subjective outcomes such as quality of life. Measuring subjective outcomes is difficult.

Psychometrics, the process of quantifying subjective outcomes, has traditionally been the purview of psychologists and educators who wrestle with measuring concepts like personality and intelligence. While not attempting to be a primer in psychometrics, this article will explicate two criteria used to assess subjective outcome measures. Reliability and validity assess the adequacy of any metric. Understanding how these criteria are estimated will better enable clinicians to evaluate the quality of instruments reported in the medical literature. Our goal is to demonstrate what “drives” the numbers used to convey reliability and validity.

MEASURING RELIABILITY

The reliability coefficient is a quantitative expression of an instrument’s reliability. There is, however, a disconnection between our traditional understanding of reliability and what coefficients represent. The consistency of...
scores (a common conceptualization of reliability) is only one component of what reliability coefficients represent. The reliability coefficient conveys the proportion of a scale’s total variance that is due to true variance (i.e., nonerror variance). There is good reason for expressing reliability in this manner and not simply as consistency.

Consider a porch thermometer. To convey reliability, variations from the “truth” when measuring outside temperature are stated. A ± 3°F variation would not be troublesome in that it would not cause one to modify attire or plans. One would consider the variation trivial and decide the porch thermometer was sufficiently reliable. Consider, on the other hand, a thermometer with ± 3°F variation used to measure neutropenic cancer patients’ temperatures. Such a thermometer is inadequate as it does not discriminate between those who need further evaluation or perhaps empiric antibiotics and those who do not. Hence, the amount of error inherent in a measure (the ± 3°F) is not sufficient to describe whether it is reliable. One must also have a sense of the normal variation of the characteristic being measured. As the outside temperature can have a wide range of values, a ± 3°F variation (error variation) is modest relative to the true variation seen in the weather. Because the true variation of human temperature (say 95°F to 105°F, a 10°F range) is far smaller than the outside temperature, the same error variation is far more profound.

Experience with natural variation in temperature permits a critical evaluation of a particular error variation. Consider, however, a scale measuring health-related quality of life. Suppose the scale varied by ± 5 units on two separate occasions (error variation) when given to 10 stable, white elderly women with arthritis who gave a range of responses (true variation) from 30 to 70 units. With many scales, a “feel” for a unit of measurement is absent even if the response range given by individuals is known. Unlike the example above in which one has a sense of what 1°F means (i.e., its discriminative and predictive utility), we have no idea what a unit of health-related quality of life means and therefore cannot make a gestalt assessment from error variation alone. To overcome this obstacle, true variation is expressed as a fraction of total variation (true variation + error variation). In statistical parlance, the amount of variation is represented by variance. This is a simplified version of a correlation coefficient that quantifies reliability. Some form of this basic construction is found in all reliability coefficients. Several observations merit comment given the basic structure of reliability coefficients. These comments explain the importance of using instruments on similar populations and