

## FROM RESEARCH TO PRACTICE

## 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.  
**J GEN INTERN MED 1998;13:757-761.**

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.<sup>1</sup> 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.<sup>2</sup> 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.<sup>3-6</sup> 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.<sup>7</sup> Validity is often defined as the extent to which a measurement actually measures what it purports to measure.<sup>7</sup> Introductory texts in clinical and social research depict “shots” at a target to convey the meaning of reliability and validity (Fig. 1).<sup>8,9</sup> 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 sphygmomanometer measure of blood pressure, these definitions and illustrations are useful.<sup>10</sup> This conceptualization of reliability and validity is less accurate when applied to subjective outcomes such as quality of life. Figure 2 shows that hitting a bull’s-eye accomplishes little if the wrong target is hit. Figure 3 demonstrates that consistently hitting the same area may still be relatively unreliable if the size of the target is smaller. Furthermore, for the purpose of establishing validity of abstract outcomes, Figure 1D is flawed as it is not possible to obtain an adequate validity coefficient if the scales are unreliable. Reliability is a necessary, albeit insufficient, requirement for validity. We describe a more comprehensive picture of reliability and validity coefficients.

## 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

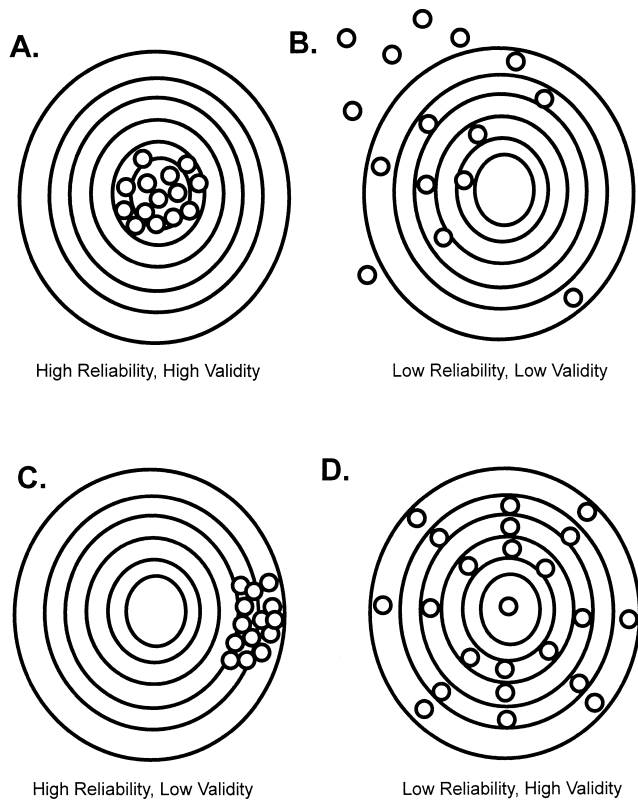
---

From Research to Practice, a Journal series, presents articles to heighten the clinician’s awareness of research and methodology issues that have direct relevance to practice. If you wish to submit a manuscript for consideration for this series, please contact Cynthia D. Mulrow, MD, MSc, Associate Editor, at mulrowc@uthscsa.edu, or contact the Journal of General Internal Medicine at (215) 823-4471 to receive the appropriate guidelines.

---

Received from the Division of Internal Medicine, Department of Medicine, Vanderbilt University, Nashville, Tenn. (TAE), and Institute for Research in Social Science, University of North Carolina, Chapel Hill, N.C. (GG).

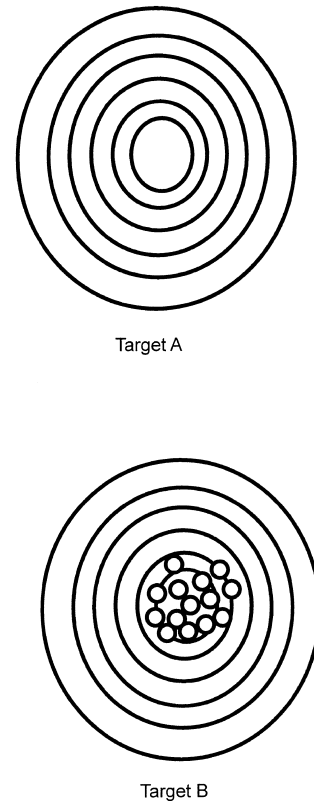
Address correspondence and reprint requests to Dr. Elasy: 7th Floor Medical Center East, Vanderbilt University Medical Center, Nashville, TN 37232-8550.



**FIGURE 1.** Typical illustrations depicting the meaning of reliability and validity.

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).<sup>11</sup> 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  $\pm 3^\circ\text{F}$  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  $\pm 3^\circ\text{F}$  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  $\pm 3^\circ$ ) 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  $\pm 3^\circ$  variation (error variation) is modest *relative* to the true variation seen in the weather. Because the true variation of human temperature (say  $95^\circ$  to  $105^\circ$ , a  $10^\circ$  range) is far smaller than the outside temperature, the same error variation is far more profound.

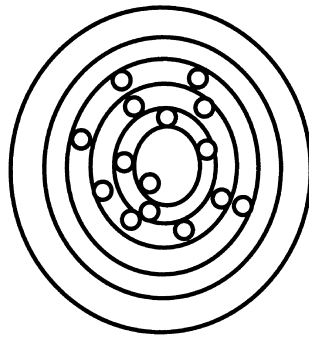


**FIGURE 2.** If A is the intended target, then the validity of shots at target B are irrelevant.

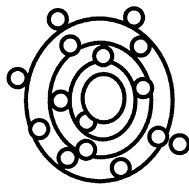
Experience with natural variation in temperature permits a critical evaluation of a particular error variation. Consider, however, a scale measuring health-related quality of life. Suppose the scale varied by  $\pm 5$  units on two separate occasions (error variation) when given to 10 stable, white elderly 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  $1^\circ\text{F}$  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.<sup>12</sup>

$$\text{Reliability coefficient} = \frac{\text{True variance}}{\text{True variance} + \text{Error variance}}$$

This is a simplified version of a correlation coefficient that quantifies reliability. Some form of this basic construction is found in all reliability coefficients.<sup>13</sup> Several observations merit comment given the basic structure of reliability coefficients. These comments explain the importance of using instruments on similar populations and



Moderate Reliability



Very Low Reliability

**FIGURE 3.** Targets of different size illustrate the relative reliability of the same results.

the advantages of increasing both the number of questions and the number of response options. Furthermore, it becomes clear that true variation, as opposed to error variation (i.e., consistency), may have a more profound influence on coefficient values.

A reliability coefficient may be improved in two ways: through greater true variation and through diminished error variation. Developing an instrument with a good reliability coefficient is difficult if there is little true variation of the characteristic between subjects. In other words, a group that is extremely homogeneous will not easily lend itself to the development of a measure that yields an acceptable reliability coefficient. Consider an attempt to measure self-efficacy in All-Star-caliber professional athletes. Because confidence is uniformly high (i.e., there is very little true variation) in this population, it is unlikely that any measure will yield a reasonable reliability coefficient even if results are fairly consistent (i.e., little error variation). A measure of a characteristic that had no variation across time or people could be consistent yet yield a poor reliability coefficient.

Conversely, measuring a population that is heterogeneous with regard to a particular attribute will, all else being equal, result in a measure that has a higher reliability coefficient when compared with its application in a more homogeneous population. Figure 3 captures this point by demonstrating that the same dispersion of shots

(error variation) will result in a different reliability coefficient depending on the size of the target (true variation). The same measure of health-related quality of life in the general population is likely to yield a lower reliability coefficient when given to patients with terminal cancer. Greater heterogeneity of a characteristic increases true variance. This fact mandates that testing and use of instruments be on similar populations.<sup>14</sup>

In addition to testing a scale on a heterogeneous population, the potential for detecting true variation is improved by increasing the number of response options for a particular question. For example, true variance is more likely to be detected if the response options are continuous rather than dichotomous ("She has a fever of 103" provides more real information than "She has a fever"). This, in part, explains why researchers prefer more response options—seven response options are preferred over five. This, however, holds only as long as the distinctions among the response options are real ("I am very much not pregnant," "Pretty much not pregnant," "Hardly pregnant," "Neither pregnant or not pregnant," etc., provides no more information, and maybe less, than a simple yes or no answer to "Are you pregnant?").

Increasing the number of questions on a scale may also enhance detection of true variance. If one imagines a scale constructed by summing the statements with which an individual agrees, the more statements given, the greater the potential variation. If there are three statements, scores can range from 0 to 3. If there are 10 statements, they can range from 0 to 10. Moreover, increasing the number of questions minimizes random error. For a scale, each question theoretically represents a parallel measure of the attribute. Increasing the number of questions may minimize random error because random error tends to cancel out. The pragmatic appeal of questionnaire brevity may be at the expense of reliability.

A reliability coefficient represents both a measure's inherent ability to give consistent results (i.e., free of error variation) and true variation. True variation reflects how much the characteristic of concern varies within the study population, length of the questionnaire, and the response scaling. In all circumstances, the amount of error variation *relative* to true variation "drives" the reliability coefficient.

## MEASURING VALIDITY

In Figure 2 a target may be hit perfectly but the measure may be invalid if the wrong target was hit. For instance, a thermometer is valid for measuring temperature but not blood glucose level. Although we often consider a measure as valid or invalid, it would be more accurate to say validity is the *extent* to which an instrument measures what it purports to measure. This is a fair definition, but fails to convey the means by which this relation is established or the nature of the correlation specified by validity coefficients. Classic techniques used to establish

instrument validity, with a focus on construct validity, are described. We will draw from an experience developing a health-related quality-of-life measure for older, southern African-American women with type 2 diabetes.

Face validity is commonly considered a minimum validation requirement. Questions are reviewed to see if they seem to relate to the measured attribute. Though frequently present, face validity is neither necessary nor sufficient.<sup>15</sup> Of particular concern are questions to which individuals are reticent to give truthful responses. Questions dealing with sensitive subject matter are frequently posed in a manner that leaves the investigator's objective unclear. Although reviewers may consider a question void of face validity, the response elicited may approximate the desired attribute. Conversely, a question may seem to have face validity yet elicit a false response.<sup>16</sup> Although these exceptions are important, most "valid" questions will meet the requirement of face validity.

A closely related method is determining content validity. Here the issue is *adequacy of sampling*. "My diabetes has interfered with my ability to participate in church activities" has face validity for many older, southern African-American women. However, if this were the only question asked, some would argue that the instrument did not adequately sample health-related quality-of-life issues (other aspects of social well-being in addition to mental and physical well-being) and was therefore void of content validity. A multidisciplinary research group guided by theory and directed primarily by patient input (e.g., focus groups and cognitive response interviews) is more likely to adequately sample the range of important issues. The central point is how well the content of a scale is sampling the content of an attribute.

Another method, criterion validity, consists of validating an instrument by correlation to a "gold standard." This is commonly done when an investigator attempts to develop a shorter or more convenient version of an established instrument.<sup>17</sup> The process is a fairly straightforward one akin to establishing the sensitivity and specificity of a new test. The criterion provides an accurate (gold standard) measure of the attribute. A word of caution regarding the criterion method as a means of validating new instruments. A health-related quality-of-life measure for older, southern African-American women with type 2 diabetes was developed because investigators concluded current instruments were insufficiently focused on issues germane to this population. If investigators conclude that current measures are inadequate, then criterion validity should not be an option for establishing validity. For example, if the investigators feel that the Medical Outcomes Study 36-Item Short Form (a commonly used generic health status measure) is inadequate for this population, then it should not be used as a criterion with which to correlate their scale.

The most common way to establish validity is quite indirect. In this setting, no standard exists for measuring the attribute. In using construct validity, investigators

construct hypothetical relations between the attribute and other attributes. Table 1 summarizes three steps investigators may use to establish a construct validity of a questionnaire (*a*) that seeks to measure a particular attribute (*A*). Again, consider the attempt to measure health-related quality of life in African-American women with type 2 diabetes. The attribute is health-related quality of life (*A*). The investigator hypothesizes that health-related quality of life (*A*) should correlate moderately well with social support (*B*). An instrument (*b*) has been validated for social support in this population. Both instruments are administered. The correlation between instrument *a* and *b* is computed. The original relation between the instrument *a* and health-related quality of life (*A*) is reported as valid to the extent that the two questionnaires (*a* and *b*) correlate. Notice that the validity coefficient of *a* represents the relation between *a* and *b* and only inferentially the relation between *a* and *A*. Validity reflects the adequacy of the hypothetical construction! If the investigator can construct several scenarios (i.e., correlation with other attributes such as coping, self-efficacy, etc.)<sup>18</sup> that yield hypothesized correlations, then confidence in the questionnaire's validity is enhanced. Unfortunately, it is still possible that the measure (*a*) is not measuring health-related quality of life (*A*) but something else that correlates well with social support, coping, and self-efficacy.

## SUMMARY

Reliability and validity are two criteria used to assess the adequacy of a scale. We have demonstrated that the term consistency does not adequately capture what a reliability coefficient conveys. We have highlighted the impact of true variation on reliability coefficients and demonstrated that error variation relative to true variation determines the coefficient value. Traditional means of accumulating validity evidence were reviewed; these approaches are complementary and may overlap in establishing validity. Theory plays a large role in establishing construct validity. Hypothesis quality "drives" the validity coefficient more than the extent to which the instrument measures what it purports to measure. Greater familiarity with the

**Table 1. Steps in Construct Validity**

1. What hypothesis may be drawn regarding individuals with high and low scores of *A*; i.e., Can I expect them to have high and low scores of another attribute (*B*); or are they likely to have some future performance?
2. Find a measure (*b*) of the other attribute (*B*) in *this population*; alternatively, observe and quantitate the future performance.
3. Correlate measure *a* with measure *b*; or the future performance. This represents the validity coefficient of measure *a*.\*

\*The maximum possible validity coefficient of *a* is the square root of the product of reliability of *a* and *b*.<sup>15</sup>

intricacies of establishing reliability and validity will enable clinicians to better evaluate the adequacy of subjective outcome measures.

*The authors wish to acknowledge the guidance of Robert F. DeVellis, PhD.*

## REFERENCES

1. McDowell I, Newell C. *Measuring Health: A Guide to Rating Scales and Questionnaires*. 2nd ed. Oxford, England: Oxford University Press; 1996.
2. Nunnally JC, Bernstein IH. *Psychometric Theory*. 3rd ed. New York, NY: McGraw-Hill; 1994.
3. Bergner M. Measurement of health status. *Med Care*. 1985;23:696-704.
4. Quality of life and clinical trials. *Lancet*. 1995;346:1-2. Editorial.
5. Stewart AL, Greenfield S, Hays RD, et al. Functional status and well-being of patients with chronic conditions: results from the Medical Outcomes Study. *JAMA*. 1989;262:907-13.
6. Kessler RC, Mroczek DK. Measuring the effects of medical interventions. *Med Care*. 1995;33:AS109-19.
7. Testa MA, Simpson DC. Assessment of quality-of-life outcomes. *N Engl J Med*. 1994;334:835-40.
8. Hulley SB, Cummings SR. *Designing Clinical Research*. Baltimore, Md: Williams & Wilkins; 1988:31-41.
9. Royce A, Singleton J, Straits BC, Staitis MM. *Approaches to Social Research*. 2nd ed. New York, NY: Oxford University Press; 1993:114-30.
10. Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology: The Essentials*. 3rd ed. Baltimore, Md: Williams & Wilkins; 1996:22-4.
11. DeVellis RF. *Scale Development: Theory and Applications*. Applied Social Research Methods Series. Newbury Park, Calif: Sage; 1991.
12. McCall R. *Fundamental Statistics for Behavioral Sciences*. 7th ed. Pacific Grove, Calif: Brooks/Cole; 1988:151-75.
13. Isaac S, Michael W. *Handbook in Research and Evaluation for Education and the Behavioral Sciences*. 3rd ed. San Diego, Calif: EDITS; 1995:174-80.
14. Perrin EB, Aaronson NK, Lohr KN, et al. Instrument review criteria. *Med Outcomes Trust*. 1997:1-5.
15. Sudman S, Bradburn NM. Asking questions: a practical guide to questionnaire design. In: Fiske DW, ed. *Series in Social and Behavioral Sciences*. San Francisco, Calif: Jossey-Bass; 1982.
16. Striener D, Norman G. Health measurement scales: a practical guide to their development and use. In: Striener D, Norman G, eds. 2nd ed. New York, NY: Oxford University Press; 1995:150-7.
17. Stewart AL, Hays RD, Ware JE. The MOS Short-Form General Health Survey: reliability and validity in a patient population. *Med Care*. 1988;26:724-32.
18. Rosenstock I, Strecher V, Becker M. Social learning theory and the health belief model. *Health Educ Q*. 1988;15:175-83.



## ANNOUNCEMENT

### American Board of Internal Medicine

#### *1999 ABIM Certification Examination in Internal Medicine*

Registration Period: September 1, 1998 – December 1, 1998  
 Examination Dates: August 24-25, 1999

#### *1999 ABIM Certification Examination in Sports Medicine*

Registration Period: July 1, 1998 – November 1, 1998  
 Examination Dates: April 16, 1999

Important Note: The 1999 Sports Medicine Examination is the last one for which Diplomates may qualify through a practice pathway.

For more information and application forms, please contact:

Registration Section  
 American Board of Internal Medicine  
 510 Walnut Street, Suite 1700  
 Philadelphia, PA 19106-3699  
 Telephone: (800) 441-2246 or (215) 446-3500 Fax: (215) 446-3590 E-mail: request@abim.org