Assessing Reliability and Validity Measures in Managed Care Studies

ISAAC D MONTOYA

BACKGROUND: To review the reliability and validity literature and develop an understanding of these concepts as applied to managed care studies.

RESULTS: Reliability is a test of how well an instrument measures the same input at varying times and under varying conditions. Validity is a test of how accurately an instrument measures what one believes is being measured.

METHODS: A review of reliability and validity instructional material was conducted.

CONCLUSIONS: Studies of managed care practices and programs abound. However, many of these studies utilize measurement instruments that were developed for other purposes or for a population other than the one being sampled. In other cases, instruments have been developed without any testing of the instrument's performance. The lack of reliability and validity information may limit the value of these studies. This is particularly true when data are collected for one purpose and used for another. The usefulness of certain studies without reliability and validity measures is questionable, especially in cases where the literature contradicts itself.

INDEX TERMS: Managed care; reliability; validity.

Clin Lab Sci 2002;16(3):153

Isaac D Montoya PhD CMC CLS(NCA) *is Clinical Professor, College of Pharmacy, University of Houston, Houston TX.*

The Reports and Reviews Section seeks to publish information on important clinical laboratory-related topics such as technological, clinical, and experimental advances and innovations. Case studies and literature reviews are also included. In addition, brief reviews of books, computer programs, audiovisual materials or other materials of interest to readers are appropriate for this section. Manuscripts and literature reviews published as a Report are peer reviewed. Direct all inquiries to Isaac Montoya PhD, Affiliated Systems Corporation, 3104 Edloe, Suite 330, Houston TX 77027-6022. (713)439-0210, (713)439-1924 (fax). imontoya@affiliatedsystems.com

Address for correspondence: *Isaac D Montoya, College of Pharmacy. University of Houston, 3104 Edloe, Suite 330, Houston TX 77027.*

Continuing Education Credit: *see pages 163 to 165 for learning objectives, test questions, and application form.*

LEARNING OBJECTIVES

After completing this article, the reader will be able to:

1. define measurement, describing sources and types of measurement errors.
2. define reliability and validity.
3. cite numerical ranges for reliability and validity.
4. describe how reliability and validity are each calculated.
5. explain how reliability and validity relate to measurement error(s).
6. apply reliability and validity concepts to the evaluation of managed care and similar studies.
7. describe the relationship between a theoretical concept and its operationalization.
8. list and define the types of validity that apply to evaluation of managed care and similar studies, including method(s) of measurement.

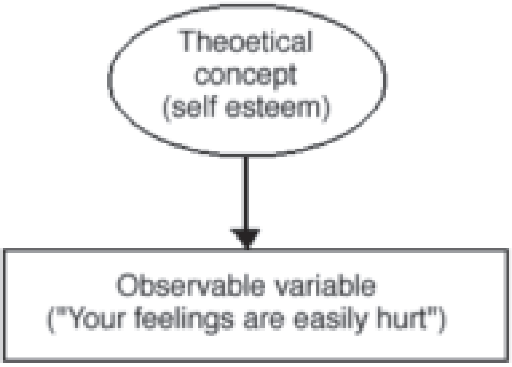
The introduction of the managed care paradigm has had a greater impact on healthcare delivery than any other single development. All aspects of healthcare have been substantially influenced by managed care. This paradigm has mandated that practitioners of both physical and behavioral medicine re-examine their clinical practices and protocols, focusing on efficiency and quality measures in an effort to control costs. The value of these measurements has been inconsistent due to the complex nature of measuring healthcare outcomes. The emphasis of the managed care paradigm is on efficiency without the sacrifice of quality. To accurately assess this, outcome measures must be employed to evaluate clinical improvements; such measures have been the subject of countless studies attempting to evaluate them. Some of these studies are complex and credible scientific research efforts, while others are simpler attempts merely to understand a particular intervention. The purpose of this paper is to examine those tools essential in high quality studies of managed care that measure actual changes resulting from this new paradigm.

It is often the case that scientists are interested in measuring concepts that are not directly observable. Concepts such as ‘kidney failure’, ‘cardiac disease’, ‘self-esteem’, or ‘depression’ must be estimated using such variables as clinical laboratory measurements or responses to a questionnaire. The process by which a theoretical concept, such as ‘self-esteem’, is measured is called the ‘operationalization’ of the concept. When a concept is operationalized, it is represented by an observable variable as illustrated in Figure 1.

As the theoretical concept and the observable variable are distinct from each other, the question always arises as to the adequacy of the observable variable as a direct and meaningful measure of the concept. Obviously, the utility of the observable variable for research depends on how well the variable captures the meaning of the concept. *Reliability* and *validity* are the two standard criteria by which the adequacy of the measurement can be assessed.

A note on terminology—when an observable variable is being used to estimate the value of an unobservable theoretical concept, the process is referred to as ‘measurement’. Thus, a questionnaire item such as “your feelings are easily hurt” is being used to measure the theoretical concept of self-esteem. When an attempt to assess the quality or adequacy of the process of measurement is made, we say that we are making an assessment of the measurement. When we ask how well the item “your feelings are easily hurt” measures self-esteem, we are assessing the validity of the item. When we ask how consistently respondents answer the item, we are assessing the reliability of the item. The meaning of reliability and validity are considered next.

Figure 1. A concept and its measurement


Theoetical concept (self esteem)
Observable variable ("Your feelings are easily hurt")

RELIABILITY

Reliability deals with the *consistency* of a measure.[1] Reliability assesses the extent to which an experiment, test, or any measurement yields the same results on repeated traits.[2] Consistency of an observable variable, as shown in Figure 2, can come from the validity of the variable as a measure of the target theoretical concept *(t).* Consistency can also come from other sources to the extent to which the observable variable measures some other concept, usually not intended *(b* in the figure). In Figure 2, *e* summarizes all the (random) sources of nonreliability.

VALIDITY

Validity is the extent to which an observable variable successfully measures (estimates the value of) a given theoretical concept.[1,2] Thus validity is always described in terms of an observable variable *and* a theoretical concept. In Figure 2, the validity of the item as a measure of self-esteem is given by the arrow *t.* If we can estimate the value of *t*, this estimate will be the assessment of the validity of the item as a measure of the concept. In the figure, *b* and *e* summarize all the sources of nonvalidity.

Thus validity is an assessment of how well one has done the measurement job from the perspective of theory. Validity is the purpose of measurement. Validity is, as we shall see, very difficult to assess because the very assessment process requires judgments about theory. Reliability is a more limited assessment of how well one has done the measurement job. However, it is easier to assess because the assessment can be carried out empirically without reference to the theory.

ASSESSMENT OF RELIABILITY

Classical Test Theory or True Score Theory (TST) is a theory about measurement.[3-6] According to TST, any measurement has two components: the true value of the theoretical concept, and an error component. Thus, any measure can be represented by:

(1) $x = X + \text{error}$

where x = observed score (value of the observable variable), X = true score (value of the theoretical variable), and error = the deviation of the value of the observed variable from the value of the theoretical concept.

Two types of errors exist; random error and systematic error. Random error is caused by any factor that randomly affects the measurement of a variable across the entire sample. For example, a person's mood at the time of completing a health questionnaire is a source of random error. Some individuals may be in a 'bad mood' as they are taking the test while others may be in a 'good mood'. Random errors do not have any systematic effects on the sample. They just push observed scores upward or downward randomly. Because there will be as many positives as negative errors, the expected mean of random errors will be zero. Random errors add variance to the data, but do not affect the average performance for the group. Because of this, random errors are sometimes called noise. Figure 3 illustrates the effect of random errors in the measurement.[7] In Figure 2, random error is indicated by the arrow *e*.

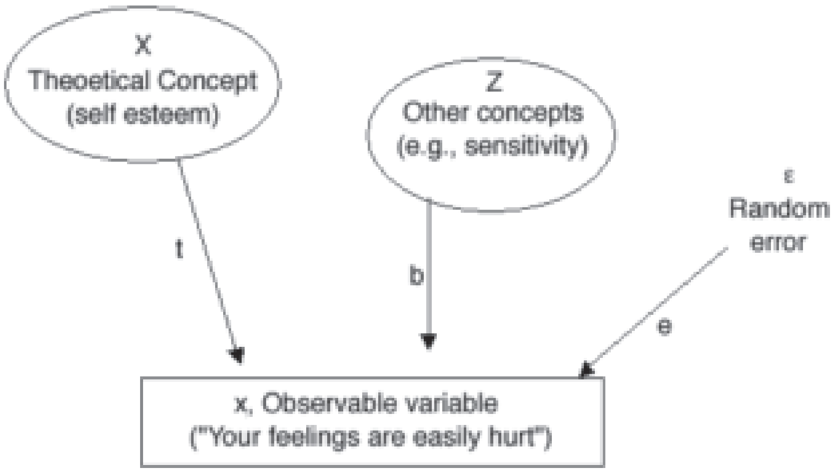
Systematic errors, on the other hand, are caused by factors that systematically affect the measurement of the variable across the entire sample. Unlike random errors, systematic errors can be consistently either positive or negative. Systematic errors are sometimes called *bias* in measurement. For example, if there is a loud noise outside the room where the patients are completing questionnaires, the noise will affect the responses for all patients, depending on the susceptibility of each patient to noise. Another example is the reading level of each patient: if the reading level of the measuring instrument is above the reading level of some patients, then the scores of those patients will be biased (usually lowered). Figure 4 illustrates the effect of systematic error *in* the measurement of a concept.[7] In Figure 2, bias (systematic error) is indicated by the systematic error term *b.*

Because the error term in equation 1 is composed of two parts, equation (1) can be rewritten as:

(2) $x = X + Z + \varepsilon,$

where x and X are as described above, Z is the unstandardized systematic error and ε is the unstandardized random error. In this equation, $X + Z$, the combined score, represents the constant part of the measure *x,* part of which represents the intended concept X and part of which represents other sources of stability (bias). This stable part can be represented as *W*.

Figure 2. Reliability and validity



(3) $x = W + \varepsilon$

When one is concerned about reliability, this equation is a useful transformation. When one is concerned about validity, this transformation is inappropriate because it confuses *X,* the stability from what one wants to measure with *Z,* the stability from unintended effects.

Another way to write equation (2) is with standardized variables. This shifts the focus from the descriptive view of *x* as composed of multiple scores. Instead it conceptualizes x in terms of the causes of its observed value. When the variables x, *X, Z,* and *e* are standardized, the equation becomes:

(4) $x = tX + bZ + e\varepsilon = rW + e\varepsilon,$

where *t* is the validity of concept X, *b* is the validity of all other systematic sources of error (concepts) not intended, *r* is the effect of the composite concept *W,* and *e* is the effect of random error.

The TST assumptions can be presented more formally as:

a) $E(\varepsilon) = 0;$
b) $\rho xz = 0;$
c) $\rho x\varepsilon = 0$ and $\rho_{w\varepsilon} = 0$*; and*
d) $\rho\varepsilon_1\varepsilon_2 = 0.$

Assumption (a) implies that the expected mean error score is zero. Assumption (b) states that the correlation between the true score and the bias is zero. Assumption (c) implies that the correlation between the error and the true score is zero, as well as the correlation between the error and the combined score. Assumption (d) states that the correlation between errors on different measurements is zero.

Reliability, however, is not meant to capture the consistency of a measure for an individual, but rather the consistency of a measure across individuals. Thus,

(5a) $Var(x) = Var(rW + e\varepsilon) = r^2 + e^2$,

in terms of effects, and

(5b) $Var\ (x) = Var(W + \varepsilon) = Var(\mathrm{W}) = Var(\varepsilon)$,

in terms of scores,

when *X, Z,* and ε are standardized and given the assumptions (a) through (c) above. Similar results to *(5a)* are given in Heise*,* and to (5b) in Pindyck and Rubinfeld.[4,8]

Equation (5a) shows the variance of *x*, a standardized variable, and *r* represents the stable, consistent factors determining *x*; the reliability (consistency) of *x* is given by the proportion of variance in *x* that is due to *W:*

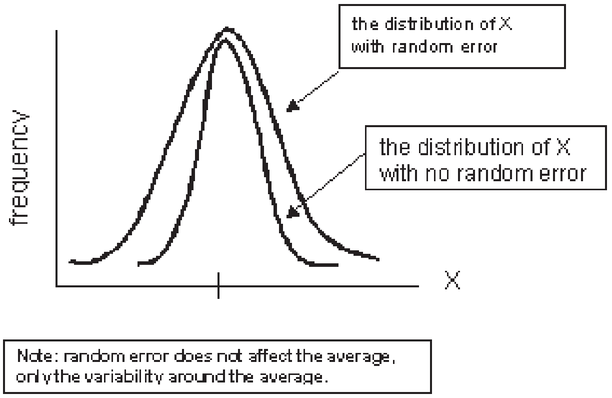
(6) $\rho_x = r^2 \div r^2 = e^2 = Var(W) \div Var(W) + Var(\varepsilon)$

Reliability then is a ratio or fraction of the combined score variance to observed variance. In other words, reliability may be defined as the proportion of consistency in the measure. Reliability thus varies between 1 and 0. Because x = rW + eå, by equation 5*,* equation 6 can be re-written as

(7) $\rho_x = r^2 \div r^2 + e^2 = Var(W) \div Var(W) + Var(\varepsilon)$

If a measure is perfectly reliable, then there is no random error in measurement. That is, all we observed is the combined score (that is, true score and bias or systemic error). Thus, for a perfectly reliable measure, ε will be zero, and $\varepsilon_x = 1$. On the other hand, if we have a perfectly unreliable measurement, there is no true score (*t* and *b* are zero). The measurement is entirely error. In this case, the above equation is reduced to:

Figure 3. Random error



(8) $\rho = 0 \div e^2 = 0 \div VAR(\varepsilon) = 0$

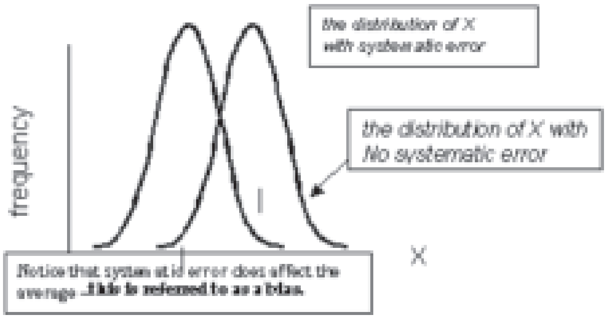
Thus, reliability will always vary between 1 and 0. The value of the reliability assessment tells us the proportion of variability in the measure attributable to the combined score. For example, a reliability of 0.5 tells us that about half of the variance of the observed score is attributable to the combined score and half is attributable to random error. A reliability of 0.8 means a variability is about 80% combined score and 20% random error and so on.

We have seen that the term reliability means 'repeatability' or 'consistency'. That is, a measure is considered reliable to the extent that it gives the same result repeatedly (assuming that what we are measuring isn't changing).

In the following section, X is represented by *x*, t is represented by *W,* and e is represented by epsilon (ε). Since reliability is based on consistency, the assessment of reliability is based on the comparison of two or more measurements. Such measurements are described as "parallel measurements." An assessment of a measure's reliability can be obtained by correlating parallel measurements. Two measurements are said to be parallel if they have identical composite scores and equal variances. Thus *x* and *x*' will be parallel if:

(9) $\mathrm{x} = \mathrm{W} + \varepsilon$

Figure 4. Bias (systematic error)



and

(10) $x' = W + \varepsilon'$

where $\mathrm{Var}(\varepsilon) = \mathrm{Var}(\varepsilon')$.

Thus, parallel measurements are distinct from one another, but similar and comparable in many ways. Parallel measurements have the same value of the theoretical composite variable *W*, and the differences between parallel measurements are the result of purely random errors. The correlation between parallel measurements can be expressed as:

(11) $\rho_{xx'} = \sigma_{xx'} \div \sigma_x \sigma_{x'}$

since X = t + e, then

(12) $\rho_{xx'} = \sigma_{(t+e)} \sigma_{(t+e')} \div \sigma_x \sigma_{x'}$

Distributing terms,

(13) $\rho_{xx'} = \sigma_t^2 + \sigma_{te} + \sigma_{et} + \sigma_{ee'} \div \sigma_x \sigma_{x'}$

Because the errors are uncorrelated with composite scores (assumption c) and uncorrelated with each other (assumption d), and because the variance, hence the standard deviations, of parallel measures are assumed to be equal, equation 13 reduces to:

(14) $\rho_{xx'} = \sigma_t^2 \div \sigma_x^2$

That is, the correlation between parallel measures is equal to the composite score variance divided by the observed variance. Rearranging terms, the unobservable composite score variance can be expressed as:

(15) $\sigma_t^2 = \sigma_x^2 \rho_{xx'}$

From equation 6 where reliability was computationally defined, and substituting, we have:

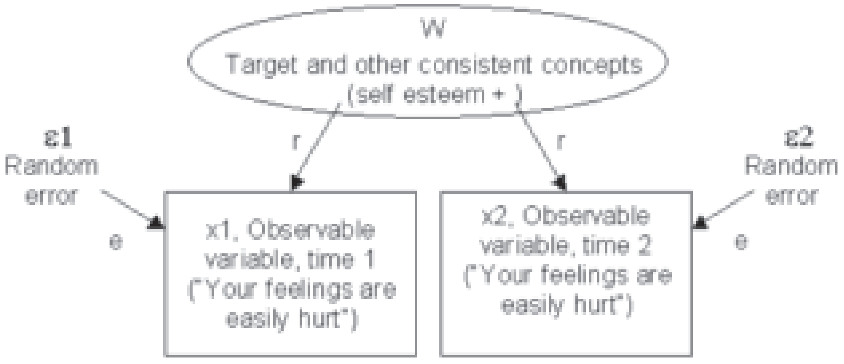
(16) $\rho_x = \sigma_t^2 \div \sigma_x^2 = \sigma_x^2 \rho_{xx'} \div \sigma_x^2 = \rho_{xx'}$

Thus, an assessment of reliability can be obtained by estimating the correlation between parallel measures. For example, if we have two or more items or a single item measured at two different times we can assess the reliability of the measurement.

Returning to the effects model of equations (4), (5a), and (6), reliability is equivalent to the following:

(17) $\rho_{xx'} = r^2 = t^2 + b^2$

Figure 5. Assessing test-retest reliability


W
Target and other consistent concepts
(self esteem +)
r
r
ε1
Random
error
ε2
Random
error
e
e
x1, Observable variable, time 1 ("Your feelings are easily hurt")
x2, Observable variable, time 2 ("Your feelings are easily hurt")

METHODS FOR ASSESSING RELIABILITY

The next question is how one can find parallel measures so that reliability can be assessed. There are four different methods of obtaining parallel measures: test-retest, alternative forms, split-half, and internal consistency methods[2]. Each of these methods is explained below.

Test-retest method

The test-retest method of establishing reliability entails administering the same instrument twice to the same group of individuals under the same conditions after some time interval has elapsed. The correlation coefficient between the first test and the retest is called *coefficient of stability*. As the name indicates, it gives an assessment of how stable the results are over a given time period. The shorter the period between tests, the higher the coefficient of stability. However, if the time interval between test and retest is very short, the participant is likely to remember how he/she answered the first time. This will give a misleadingly high coefficient of stability.

An example of an assessment of test-retest reliability is given in Figure 5. In this figure we see how the correlation between the item at two points in time is equal to r^2. This method assumes that the value of the theoretical concept does not change from time 1 to time 2. It assumes that the item reflects the theoretical concepts *(r)* at each time period equally well and that the random error effect is the same at each time period. It assumes that the sum of the effects of *X* and *Z* remains constant from time 1 to time 2 (there are not any different stable effects at the two time periods). Because the same item (or the same scale or the same instrument) is being administered at two times with the assumption that the underlying theoretical concept has not changed, time stability is taken as the causal effect of the theoretical concept on the observable variable.

Alternative-form method

The alternative or equivalent form method entails administering the research instrument to the same group of individuals at two different times using different, but equivalent forms.[6,10,11] The reliability coefficient is called the *coefficient of equivalence.* A high coefficient of equivalence indicates that both research instruments are assessing similar contents of the instrument. This method is illustrated in Figure 6, where two items are taken as equivalent measures of the self-esteem theoretical concept. This method assumes that each item (or each scale or each instrument) involves exactly the same causal effect from the theoretical variable and takes the consistency of responses as an assessment of the causal effect of the theoretical concept on each of the alternative forms of the item (scale, instrument).

Split-half method

The split-half method is the most frequent method used to assess reliability.[2,6,11] Under this method, the research instrument is administered to a group of respondents and then the items are split in half, for example odds and evens, for purposes of scoring. The results of the two halves are then compared. The association between the two halves is called the *coefficient of internal consistency* and measures the degree to which the two halves are equivalent. In Figure 6, this process is represented as the comparison of one item with another (although the comparison can be of several items with several other items). The split-half method offers a clear advantage in terms of time and resources over the test-retest and the alternative form methods in that it does not require the test to be administered twice to the same group of respondents.

The correlation between the two halves of the test, however, is a measure of the reliability for each half of the test rather than the total test as correlations on fewer items are usually less than on more items. Thus, a statistical correction should be made so that the researcher can get an assessment for the whole test, not just for the odd or even questions of the test.[2,6,11] This procedure is known as the Spearman-Brown prophecy formula. If the total test is twice as long as each half, the Spearman-Brown formula will be given by:

$$\text{(18)} \quad \rho_{xx''} = 2\rho_{xx'} \div 1 + \rho_{xx'}$$

where $\rho_{xx''}$ is the reliability coefficient for the entire test and $\rho_{xx'}$ *is* the split-half correlation. For example, if the split-half correlation is 0.75, the reliability for the whole test is (2 * 0.75)/(1 + 0.75) = 0.857.

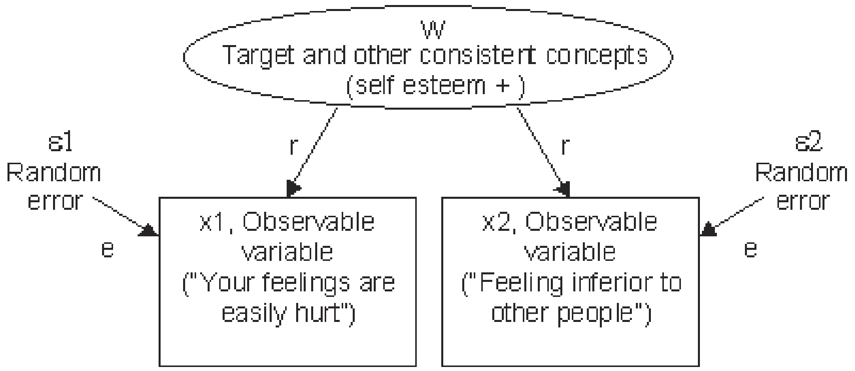
One disadvantage of the split-method, however, is that the reliability obtained may depend on the number of ways the instrument is subdivided. For example, it is possible that the correlation between the first and second halves of the test may be different than the correlation between odd and even items in the test. Thus using the split-half method, it is possible that reliability may differ even though the same items are administered to the same individuals at the same time.

Internal consistency method

As seen above, even the split-half method is not without its shortcomings. However, there are other methods of estimating reliability that do not require either the splitting of the items nor the repeating of items. The internal consistency method is one such method. Under the internal consistency method, a single test is administered. Within the test, questions are grouped together that measure the same concept and are then used to assess reliability of that portion of the test.

Another and more common way of computing correlation values among the questions in an instrument is the Cronbach's alpha. The Cronbach alpha splits all the questions on the instrument every possible way and computes

Figure 6. Assessing alternate forms, split-half, and internal consistency reliability



correlation values for all of them (using a statistical software program such as SPSS or SAS). At the end, the software program will generate a number for Cronbach alpha and just like a correlation coefficient, the closer it is to one, the higher the reliability assessment of the instrument. The Cronbach's alpha can be defined as:

$$(20) \quad \alpha = [N \div (N-1)][1-(\Sigma\sigma^2 Y_i \div \sigma_x^{\,2})]$$

where N = the number of items, $\Sigma\sigma^2 Y_i$ is the sum of the items variances, and $\sigma_x^{\,2}$ is the total variance. Alternatively, the Cronbach alpha may be defined as:

$$(21) \quad \alpha = N\bar{\rho} \div [1 + \bar{\rho}(N-1)]$$

where N is the number of items, and $\bar{\rho}$ is the mean of the inter-item correlation.

In both cases the alpha coefficient can be interpreted as the expected correlation of one test with an alternative form containing the same number of items.

If the responses are dichotomies (1 or 0), however, one can use the Kuder-Richardson formula given by:;

$$(22) \quad KR20 = N \div N\text{-}1[1 - \Sigma\rho(1-\rho) \div \sigma_x^{\,2}]$$

where N is the number of dichotomous items, p is the proportion responding 'yes' to the item, $\rho_x^{\,2}$ is the variance of the total test. KR2O has the same interpretation of the regular alpha.

RELIABILITY SUMMARY

Reliability is one important perspective from which to assess the operationalization of a concept. However, because all measurement involves uncertainty, not only because of random error, but also because of the involvement of unobservable theoretical concepts, reliability cannot be assessed with absolute accuracy. It must be estimated from fallible data. There are four methods to assess reliability: test-retest, alternative form, split-half, and internal consistency methods. As discussed above, both the test-retest and the split-half methods have shortcoming as estimators of reliability. The main shortcoming of the test-retest method is that experience of the first testing can influence the responses in the second testing. The shortcomings of the split-half method, on the other hand, are that correlation between halves will differ depending on how the total number of items in the instrument was divided. The internal consistency method is easy to use because it requires only a single test administration. Regardless of the method used to assess reliability, researchers agree that reliability should not be below 0.80.[2] Furthermore, it is not only important to achieve a high reliability level, but also it is important to report how reliability was assessed.

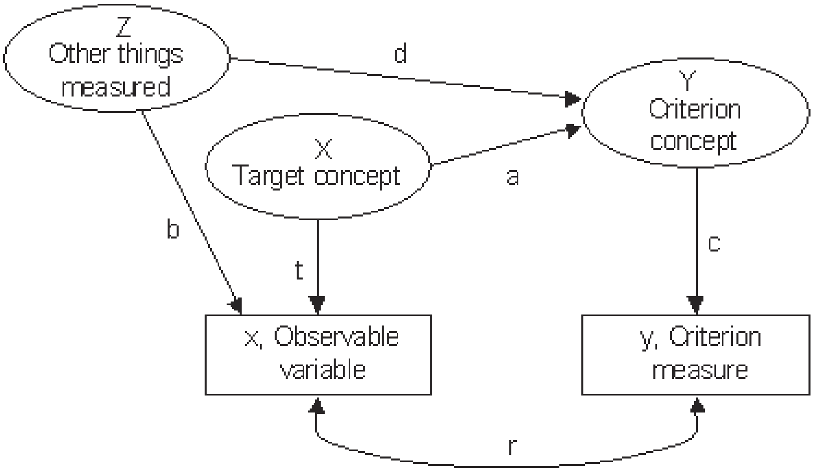
ASSESSMENT OF VALIDITY

The second tool in evaluating the operationalization of a concept is validity. Validity may be defined as the extent to which any measuring instrument measures what it is intended to measure.[2] There are three common types of validity common to the social sciences: content related, criterion-related, and construct validity.[2,6,10,11] Each of them is explained below.

Content validity

Measuring how well the operationalization of the concept compares to the relevant content domain for the concept assesses content validity. Thus, it is mostly applicable to concepts measured by multiple items. Content validity is assessed by the extent to which empirical measurement reflects the specific domain of the theoretical concept. For example, a test on mathematical ability will not be content-valid if it only includes summation problems and neglects subtraction, division, and others. Thus, content validity deals with the thoroughness or completeness of its observable variables. Content validity should answer the question of whether the assessment strategy covers the major dimensions or factors of the subject matter under assessment.

Figure 7. Assessment of validity



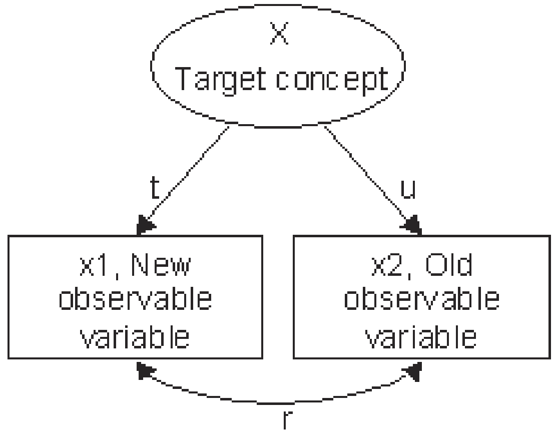
Content validity, unlike criterion and construct validity, is not assessed statistically. The assessment of content validity is a subjective judgment by the investigator, observer, or groups of subject matter experts. Like all validity issues, construct validity depends directly on theory. In this case, the theory involves the definition of the theoretical variable. Content validity can be neither achieved nor assessed unless the dimensions of the theoretical concept are clearly and explicitly defined. Content validity is directly and exclusively an assessment of measurement theory.

Criterion-related validity

All assessments of validity involve theory. All assessments of validity involve measurement theory (the effect of the theoretical concept on the observable variable) because it is the assertion that the observable variable measures the target concept that is being assessed. The most convincing assessments of validity also involve substantive theory, in which the target concept is tied to a related criterion concept.

In criterion-related validity one checks the performance of the operationalization against some criterion. For this to be done there must be an acknowledged and accepted theory that the target variable is causally related to the criterion concept. The assessment of criterion validity is depicted in Figure 7. We see, as in Figure 2, that a target concept is being measured (operationalized) by an observable variable. That is, the values on the observable variable are seen to depend causally on the theoretical variable. The observable variable is also seen to be dependent on other theoretical concepts as well, concepts that may be known or not known. The effect of the target concept *X* on the observable variable x is measured by *t.* In addition, the effect of the (unknown) other concept(s) *Z* on the observable variable is *b.* The coefficient *t* is the validity coefficient. The coefficient *b* is the bias in measurement.

Figure 8. Assessment of construct validity: convergence



The assessment of criterion validity involves computing a correlation coefficient *(r* in Figure 7) between the measure of the target concept and the measure of the criterion concept. Mathematically, in all assessments of criterion validity, it is assumed that *Z,* the unmeasured and unknown other things that affect the observable variable are not associated with the criterion concept, (i.e., $d = 0$), so that $r = tac$. In this analysis, *t, a,* and *c* are standardized effects, so that their product varies in absolute value between 0 and 1. If the

$$r = tac + bdc$$

criterion variable is considered to be a very good measure of the criterion concept *(c* is approximately equal to 1) and if the criterion theory is considered to be very strong *(a* is approximately equal to 1), then the correlation coefficient *r* is approximately equal to the validity coefficient *t.*

If the theory describes an effect in the future (or in the past), criterion validity is called "predictive validity".[5] If the theory describes an effect in the same time period, criterion validity is called "concurrent validity". In predictive validity, for example, a measurement of mathematical ability should be able to predict how well a person will do in an engineering profession. A high correlation between the observable measure of mathematical ability and the observable measure of engineering professional success is taken as an assessment of how well the mathematical ability observable variable measures the mathematical ability theoretical concept.

A limitation of using criteria-related validity is that one often lacks well established theory with which the measurement can be evaluated. Another limitation is that the validity assessment of the observable variable against its target concept is also an evaluation of the theory that links the target concept with the criterion concept and the measurement theory that links the criterion concept and the criterion measure.

Construct validity

Construct validity assesses the extent to which the observable variable measures the theoretical concept by comparing the observed variable(s) with observed variable(s) of related concepts. That is, construct validity is concerned with the extent to which a particular measure relates to measures of the same or different concepts. Construct validity is assessed by convergence and by discriminability.

Convergence

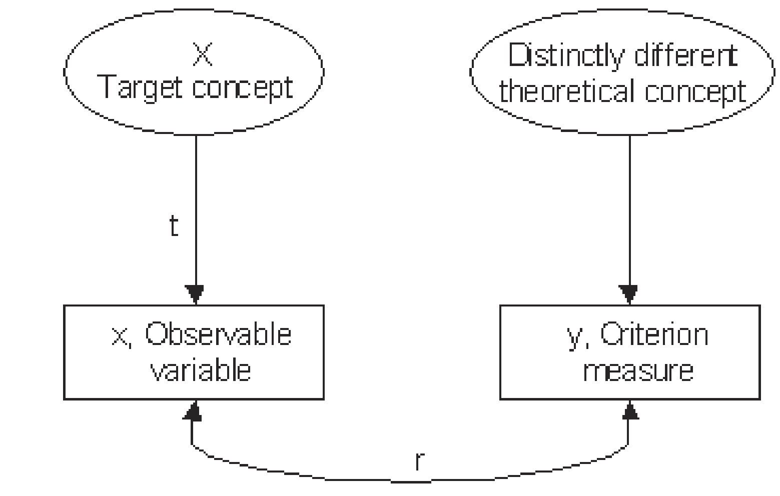
Convergence refers to direct attempts to measure the same concept by multiple methods or in multiple settings.[1] Convergence is depicted in Figure 8. Kerlinger describes how convergence is used to assess construct validity.[1] In Figure 8, a new observable variable (or set of variables) is being assessed as a measure of the target concept. This is accomplished by computing the association between the new observable variable and another observable variable that is recognized as already providing adequate measurement of the target concept. For example, computing the correlation of the new measure with the Beck Depression Scale assesses a new measure of depression. Thus the old observable variable is used to assess the adequacy of the new theoretical variable. A variation on convergence is to compare the observable variable with a measure of a closely related concept. It is, of course, a theoretical issue as to the closeness of two concepts.

Discriminability

Discriminability refers to assessments of an observable variable as a measure of a target concept by comparing the observable variable with observable variables of unrelated concepts. Discriminabilty is depicted in Figure 9.

The adequacy of the observable variable as a measure of the target concept is assessed by computing the association between the observable variable and another observable variable measuring a concept that is known or expected to have a very different meaning than the target concept. In this case a correlation near zero is considered to provide some evidence for the adequacy of the observable variable. Discriminability is always assessed alongside of convergence. It is not considered informative that an observable variable is not associated with unrelated concepts unless it is simultaneously shown that the observable variable is associated with other measures of the same concept or closely related concepts.

Figure 9. Assessment of construct validity: discriminability



Validity summary

As seen above, validity is more difficult to assess than reliability. Unlike reliability, the assessment of validity is more closely linked to the theory underlining the concept. There are three basic forms to assess validity. Content validity focuses on content relevance and coverage. It assesses whether the items on the test are part of the concept's domain and whether the items included represent the breadth of the concept. Criterion-related validity assesses the measurement according to the predictive or concurrent utility of the measurement. That is, criterion-related validity assesses the adequacy of the measurement at the same time that it assumes the predictive or concurrent adequacy of the theory. Finally, construct validity compares the observable variables of the target concept to observable variables of the same or different concepts.

RELIABILITY AND VALIDITY SUMMARY

Validity and reliability are both important tools to assess the operationalization of a concept. A measure can be very reliable, but not valid. That is, a measure can never be more valid than it is reliable. And of course a measure can be neither reliable nor valid. The goal of the managed care study should be to have a measurement that is both reliable and valid.

CONCLUSION

The use of reliability and validity measures in studying managed care greatly improves the scientific rigor and quality of research and evaluation studies. Only when it has been established that the instruments used in managed care studies are consistent and actually measure the parameters of the variable that they purport to measure, will it be possible to develop outcome measures that are of value in improving managed care practices. It is by means of such outcome measures that positive changes can be identified and implemented in order to achieve the goals of improved efficiency and effectiveness of, as well as access to, the contemporary managed healthcare system.

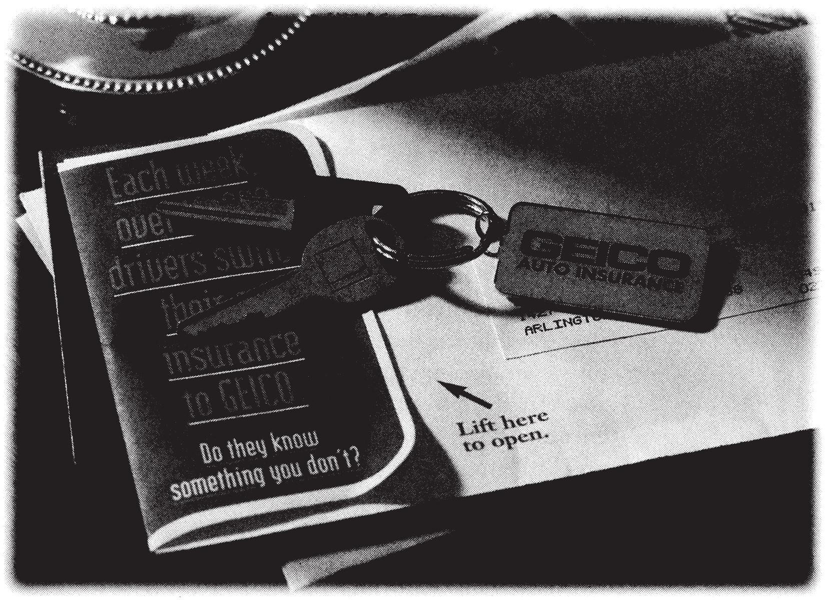
Health services researchers seeking to accurately measure outcomes should first consider creating an instrument that will measure exactly what the researcher intends to measure using language specific to the population that will be sampled. This instrument should then be pilot tested and reliability and validity measurements assessed. Only then should sampling for the intended study begin. These measurements should also be reported in all ensuing reports and publications. This will provide the reader with an increased comfort level and enhance the credibility of the study.

ACKNOWLEDGEMENTS

Support for this research was provided by grant DA 11414 to the Investigator from the National Institute on Drug Abuse. The author would like to thank Travis E Cal for his assistance in the preparation of this manuscript.

REFERENCES

1. American Psychological Association. Standards for educational and psychological tests. Washington DC: American Psychological Association; 1974.
2. Cannines EG, Zeller RA. Reliability and validity assessment. Beverly Hills CA: Sage; 1979.
3. Cronback FL, Gleser GL, Nanda H, Rajaratnan N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley; 1972.
4. Heise DR. Causal analysis. New York: Wiley; 1975.
5. Kerlinger FN. Foundations of behavioral research. 2nd ed. New York: Holt, Rinehart and Winston; 1973.
6. Lord FM, Novick MR. Statistical theories of mental scores. Reading, MA: Addison-Wesley; 1968.
7. Nunnally JC. Psychometric theory. New York: McGraw-Hill; 1978.
8. Pindyck R, Rubinfeld D. Econometric models and economic forecast. 3rd ed. New York: McGraw-Hill; 1991.
9. Stanley JC. In: Thorudike RL, editor. Educational measurement. Washington, DC: American Council on Education; 1971.
10. Steltiz C, Jahoda M, Deutsch M, Cook S. Research methods in social relations. Rev ed. New York: Holt Rinehart and Winston; 1959.
11. Trochim. Http:ITrochim.Human.Cornell.Edu. 1998 (cited. http:/trochim.human.cornell.edu)


Each week over drivers switch their insurance to GEICO
Do they know something you don't?
GEICO AUTO INSURANCE
ARLINGTON
Lift here to open.