MEASURING HEALTH

A Guide To Rating Scales and Questionnaires

SECOND EDITION

Ian McDowell

Claire Newell

New York Oxford
OXFORD UNIVERSITY PRESS
1996

health, for there are external factors, such as an accident, which could drastically alter future health levels but which are not an aspect of a person's health.

The conceptual basis for a health index narrows the range of questions that could possibly be asked, but a relatively broad choice remains among, for example, the many questions that could be asked to measure "functional disability." Whether the index reflects a conceptual approach to health or whether it is developed on a purely empirical basis, procedures of item selection and item analysis are used to guide the final stages of selecting and validating the questions to be included in the measurement.

THE QUALITY OF A MEASUREMENT: VALIDITY AND RELIABILITY

Someone learning archery must first learn to hit the center of the target, and then learn to do this consistently. This is analogous to the validity and reliability of a measurement (77) The consistency (or reliability) of a measurement would be represented by how close successive shots fall to each other, wherever they land on the target Validity would be represented by the aim of the shooting—how close, on average, the shots come to the center of the target. Ideally, a close grouping of shots should lie at the center of the target (reliable and valid), but a close grouping of shots may strike away from the center, representing an archer who is consistently off target, or a test that is reliable but not valid, perhaps due to bias in the measurement.

The core idea in validity concerns the meaning, or interpretation, of scores on a measurement "Validity" is commonly defined as the extent to which a test measures what it is intended to measure. This approach is commonly used in epidemiology and underlies the notion of sensitivity. This is not a sufficient explanation of valid-

ity, however, for valuable interpretations of a measurement may be found that go beyond the original intent of the test. As an example, mortality rates were recorded to indicate levels of health or need for care, but the infant mortality rate may also serve as a convenient indicator of the socioeconomic development of a region or country. Hence, a more general definition holds that validity describes the range of interpretations that can be appropriately placed on a measurement score: What do the results mean? What can we conclude about the person who produced particular scores on the test? By focusing attention on the breadth of a measure, this approach also reflects its specificity (see page 31). The shift in definition is significant, for validity is no longer a property of the measurement, but rather of the interpretation we place on the results. This view holds advantages and disadvantages. It may stimulate a search for other interpretations of an indicator, as illustrated in the development of unobtrusive measures (78). This can lead to discovering links between constructs thought to be independent. An advantage of the broader definition is also seen in the dementia screening tests we review in Chapter 7. Most of these are valid in that they succeed in their purpose of identifying cognitive impairments (i.e., they are sensitive). However, several of the tests show low specificity in that people with low educational attainment also achieve positive scores, suggesting that the scales provide a more general indicator of cognitive functioning. A disadvantage of broadening the definition of validity is that it may foster sloppiness in defining the precise purpose of a measurement. Because of this, immense amounts of time have been wasted over arcane speculation on what certain scales of psychological well-being are actually supposed to measure. This can best be avoided by closely linking the validation process to a conceptual expression of the aims of the measurement and also linking that concept with other, related concepts to indicate alternative possible interpretations of scores.

The practical concern then becomes how we tell if a measurement is striking close to the center of the target and measuring what it is supposed to. There are many ways of testing validity. The choice depends on the purpose of the method (e.g., screening test or outcome measurement) and on the level of abstraction of the topic to be measured. The following section gives a brief introduction to the methods that will be mentioned in the reviews. More extensive discussions are given in many sources (e.g., references 32, 34, 79–81).

ASSESSING VALIDITY

Most validation studies begin by referring to content validity. Each health measurement represents a sampling of questions from a larger number that could have been included. Similarly, the selection of a particular instrument is a choice among alternatives, and the score obtained at the end of this multistage sampling process is of interest to the extent that it is representative of the universe of relevant questions that could have been asked. "Content validity" refers to comprehensiveness, or to how adequately the sampling of questions reflects the aims of the index that were specified in the conceptual definition of its scope. For example, in a patient satisfaction scale, do all the items appear relevant to the concept being measured, and are all aspects of satisfaction covered? If not, invalid conclusions may be drawn. Feinstein has proposed the notion of sensibility, which includes, but slightly extends, the idea of content validity (82, Chapter 10) "Sensibility" refers to the clinical appropriateness of the measure: is its design, content and ease of use appropriate to the measurement task? Feinstein offered a check list of 21 attributes to be used in judging sensibility; these involve subjective assessments rather than statistical analyses. Indeed, content validity is seldom tested

formally; instead, the face validity or clinical credibility of a measure is commonly inferred from the comments of experts who review its clarity and completeness. A common procedure to establish content validity is to ask patients and experts in the field to critically review the content of the scale. It is difficult, perhaps impossible, to prove formally that the items chosen are representative of all relevant items (83). Occasionally, tests of linguistic clarity are used to indicate whether the phrasing of the questions is clear.

Following content validation, more formal, statistical procedures are used to assess the validity of a measurement. These generally begin with "criterion validity," which considers whether the instrument correlates highly with a "gold standard" measure of the same theme. Criterion validity may test the accuracy of the complete measurement, or of each question. The latter involves "item analysis."

Several statistical procedures may be used in testing criterion validity. In the simplest case there is already an accepted way to measure the concept in question, a criterion or gold standard against which the new measurement may be compared. This typically occurs when the new instrument is being developed as a simpler, more convenient alternative to an accepted measurement. The new and the established tests are applied to a suitable sample of people and the results are compared using an appropriate indicator of agreement. Hence, this is sometimes known as correlational validity; typically, the correlation of each question with the criterion score is used to select the best questions and thereby refine draft versions of the questionnaire. Criterion validity is usually divided into "concurrent" and "predictive" validity, depending on whether the criterion refers to a current or future assessment. In the former, a questionnaire on hearing difficulties might be compared with the results of audiometric testing. In predictive validation, the test is applied in a prospective study in which the measurements made at the start of the study are compared to subsequent patient outcomes (mortality, discharge, etc.). Since this may demand a long investigation it is rarely used; there are also logical problems with the method. It is likely that during the course of a prospective study (and perhaps as a result of the prediction) interventions will be applied to selectively treat the individuals at highest risk. If successful, the treatment will alter the predicted outcome (contaminate the criterion) and wrongly make the test appear invalid. To avoid this predictive validity paradox, predictions are more commonly tested over a very brief time interval, making this approach equivalent to concurrent validity.

Our chapter on psychological measurements describes several screening tests, and their validation illustrates a major category of criterion validation studies. Here the criterion takes the form of a diagnosis made by a clinician who does not see the test result. The task is to show how well the test agrees with this classification and also to identify the threshold score on the test that most clearly distinguishes between well and sick respondents. The test is applied to contrasting groups of respondents and the patterns of their responses are compared. Two potential errors are recognized: a test may fail to identify people who have the disease or it may falsely classify people without the disease as being sick. The "sensitivity" of a test refers to the proportion of persons with a particular disease who are correctly classified as diseased by the test, while "specificity" refers to the proportion of people without the disease who are so classified by the test. These terms are quite logical: the sensitivity of the test indicates whether the test can identify the presence of a disease, while specificity indicates the degree to which it responds only to that disease. Specificity hence corresponds to "discriminal validity" in the language of psychometrics. Accordingly, tests of specificity may compare scores for peo-

ple with the disease to others who have different diseases, rather than to people who are completely healthy; as so often in research, the choice of comparison group is subtle but critical. Several extensions to sensitivity and specificity analyses exist, some of which will be encountered in this book. For example, Goldberg illustrates various ways of combining sensitivity and specificity estimates to give a single estimate of the adequacy of a test. He also gives formulae for calculating the prevalence of a disease using a screening test of imperfect, but known, sensitivity and specificity, correcting for the imperfections in the test (56, p93).

Sensitivity and specificity analyses require that the scores obtained using the test be divided into two categories, indicating presumed sick and presumed well. The threshold score that divides these two categories is known as the cutting point or cutting score; to avoid confusion it is often expressed as two numbers, such as 23/24, indicating that the threshold lies between these. While simple, the division into two categories fails to reflect the notion of disease as a continuum, and so results in a loss of information. Several ways to describe the sensitivity of the test over a range of possible cutting points have been proposed. Changing the cutting point alters the proportions of well and sick people correctly classified by the test such that an increase in sensitivity is generally associated with a decrease in specificity. Because of this, the two parameters may be plotted graphically at different cutting points for the test in a further application of the ROC analyses described earlier. The true-positive probability (sensitivity) is plotted against the false-positive probability (1 − specificity) for various cutting points. The resulting curve illustrates the tradeoff between sensitivity and specificity. The area under the curve (AUC) provides an indicator of the usefulness of the test, indicating the probability that a randomly chosen, healthy person will score higher than a randomly chosen person with the disease (67, 84). AUC values of 0.5 to 0.7 represent poor accuracy, 0.7 to 0.9 indicate "useful for some purposes," while values over 0.9 indicate high accuracy (85). Likelihood ratios offer another way of summarizing sensitivity and specificity; their use seeks to avoid the undue reliance on a single cutting point implicit in sensitivity analyses. The likelihood ratio for a particular value of a diagnostic test is the probability of the test result in people with the disease, divided by the probability of the same result in people without the disease (86). This is clinically useful, as it indicates how much more likely the test result is to occur in people with the disease than in those without.

Some caveats should be considered in interpreting sensitivity and specificity figures. Sensitivity may be influenced by factors such as age or educational level, and the dementia screening scales described in Chapter 7 illustrate this. It is convenient to report sensitivity for a single cutting point, but optimal detection may demand that the cutting point be varied for different age or educational groups. The reader should also critically review the study design in which sensitivity and specificity were estimated. The ideal is a study in which the gold standard and the test are applied to all persons in a selected population. This is rarely feasible, however, for it implies obtaining gold standard assessments from large numbers of people who do not have the condition; there are often cost or ethical constraints to this. Hence a common alternative is to take samples of those with and those without the condition, selected on the basis of the gold standard; typically these are people in whom the condition was suspected and who are attending hospital for assessment. Sensitivity and specificity are calculated in the normal manner. The problem is that the comparison group typically includes patients with other conditions; this may give misleading validity results if the test is to be used on an unselected population. An alternative approach is often used in validating a screening test in a population sample; here, the screening

test is applied, and those who score positively receive the gold standard assessment, along with a random subsample of those scoring negatively. This approach incurs a bias known as diagnostic workup bias, which inflates the estimate of sensitivity and reduces that of specificity. This bias must be corrected, although regrettably often it is not. There are several alternative formulae for this (87).

Construct Validity

For measurements of pain, quality of life, and depression, gold standards do not exist; validity testing is more challenging. It requires assembling multiple indicators of validity in a process known as construct validation. This is akin to assembling evidence to support or refute a complex scientific theory and to show under what circumstances it holds true. Construct validation begins with a conceptual definition of the topic or construct to be measured, indicating the internal structure of its components and the theoretical relationship of scale scores to external criteria. These may be expressed as hypotheses indicating, for example, what correlations should be obtained with other instruments, which respondents should score high or low, or what other findings would be predicted from the scores. None of these challenges alone proves validity and each suffers logical and practical limitations, although when carefully applied they build a composite picture of the adequacy of the measurement. A well-developed theory is required to specify such a detailed pattern of data, a requirement that is not easily met. The main types of evidence used to indicate construct validity are briefly described here; their practical application is illustrated in the reviews of individual measurements.

Correlational Evidence

Hypotheses are formulated which state that the measurement will correlate with other methods that measure the same concept; the hypotheses are tested in the normal way. This is known as a test of "convergent validity" and is equivalent to assessing sensitivity. Where no single criterion exists, the measurement is sometimes compared with several other indices using multivariate procedures. Hypotheses may also state that the measurement will *not* correlate with others which measure different themes. This is termed divergent validity, and is equivalent to the concept of specificity. For example, a test of "Type A behavior patterns" may be expected to measure something distinct from neurotic behavior. Accordingly, a low correlation would be hypothesized between the Type A scale and a neuroticism index and, if obtained, would lend reassurance that the test was not simply measuring neurotic behavior. Naturally, this provides little information on what it *does* measure.

Correlating one method with another would seem straightforward, but there are logical problems. Because a new measurement is often not designed to replicate precisely the existing method against which it is being compared (indeed, it may be intended to be superior), the expected correlation may not be perfect. But how high should it be, given that the two indices are inexact measurements of similar, but not identical, concepts? Here lies a common weakness in reports of construct validation: few studies declare what levels of correlation are to be taken as demonstrating adequate validity. The literature contains many illustrations of authors who seem pleased to arbitrarily interpret virtually any level of correlation as supporting the validity of their measure. Construct validation should begin with a reasoned statement of the types of variable with which the measure should logically be related; the recent studies of the EuroQol illustrate this (88). The expected strength of correlation coefficients (or of the variance to be explained) should be stated prior to the empirical test of validity.

Several guidelines may assist the reader in interpreting reported validity correlations. First, as there is always some error

of measurement, the maximum correlation can never reach 1.00. It can only rise to the square root of the reliability of the measurement. Where two measurements are compared, the maximum correlation between them is the square root of the product of their reliabilities (80, p84). Where the reliability coefficients are known, it is possible to compare the observed correlation to the maximum theoretically obtainable; this helps in interpreting the convergent validity coefficients between two scales. For example, a raw correlation between two scales of, say, 0.6 seems modest; but if their reliabilities are 0.7 and 0.75, the maximum correlation between them would only be 0.72, making the 0.6 seem quite high. By extension, we can estimate what the criterion validity correlation would be if both scales were perfectly reliable:

$$r_{xy}' = \frac{r_{xy}}{\sqrt{(r_{xx}r_{yy})}},$$

in our example 0.83. These corrections are only appropriate in large samples, of 300 or more; with smaller samples they can lead to validity coefficients that exceed 1.0; this seems over-optimistic for most of the measurements we review.

The second guideline involves translating a correlation coefficient into more interpretable terms. This assumes that the measurement is being used to predict a criterion, and the purpose is to assess how much the accuracy of prediction is increased by knowing the score on the measurement. The simplest approach is to square the correlation coefficient, showing the reduction in error of prediction that would be achieved by using the measurement compared to not using it. As an example, if a health measurement correlates 0.70 with a criterion, using the test will provide 0.7^2, or almost a 50% reduction in error compared with simply guessing. Applying this indicates that the value of a measurement declines rapidly below a validity coefficient of roughly 0.50. The derivation of this approach is given by Helmstadter (80, p119), who also describes other ways of interpreting validity coefficients.

The adequacy of validity (and reliability) coefficients may be interpreted in light of the values typically observed. The convergent validity correlations between the tests reviewed in this book are low, typically falling between 0.40 and 0.60, with only occasional correlations falling above 0.70 for very similar instruments (such as the Barthel Index and PULSES Profile). The attenuation of validity coefficients due to unreliability implies that a correlation of 0.60 between two measures represents an extremely strong association.

Pearson correlations are often used in reporting reliability and validity findings, but they should be interpreted with caution. If points are drawn on a graph to plot one rating against another for a series of patients, the correlation coefficient shows how closely the points lie to a straight line. This indicates how accurately one rating can be predicted from another. Coefficients range from −1.0 (indicating an inverse association) through 0.0 (indicating no association at all) to +1.0. Pearson correlations quantify the association between two measurement scales, but do not indicate agreement. This is because a perfect correlation requires only that the pairs of readings fall on a straight line but does not specify which line. For example, if one sphygmomanometer consistently gives blood pressure readings 10 mmHg higher than another, the correlation between them would be 1.0, but the agreement would be zero. Correlations are also influenced by the range of the scale: wider ranges tend to produce higher correlations, even though the agreement remains the same. Bland and Altman discuss these limitations and propose alternative approaches to expressing the agreement between two ratings (89).

Factorial Validity

Factor analysis can be used to describe the underlying conceptual structure of an

instrument; it examines how far the items accord in measuring one or more common themes It can therefore be used in studying content validity: do the items fall into the expected groupings? Using the pattern of intercorrelations among replies to questions, the analysis forms the questions into groups or factors that appear to measure common themes, each factor being distinct from the others. For example, Bradburn selected questions to measure two aspects of psychological well-being which he termed positive and negative affect Factor analysis of the questions confirmed that they fell into two distinct groups, which were homogeneous and unrelated to each other, and which by inspection appeared to represent positive and negative feelings. This analytic method is therefore commonly used to study the internal structure of a health index that contains separate components, each reflecting a different aspect of health. Typically, these components of the measurement will be scored separately. Factor analysis can also be used in construct validation by indicating the association among several measurements. Scales measuring the same topic would be expected to be grouped by the analysis onto the same factor (a test of convergent validity), while scales measuring different topics would group on different factors (divergent validity).

Factor analysis is widely used, and all too frequently misused, in the studies reported in this book; several principles guide its appropriate use (90, 91). The items to be analyzed should be measured at the interval-scale level; the response distributions should be approximately normal and there should be at least five (some authors say 20) times more respondents in the sample than there are variables to be analyzed. Because so many health indices use categorical response scales (such as "frequently," "sometimes," "rarely," and "never") the first and second of these principles are often contravened. In such cases latent trait analysis can be used to provide a factor analysis applicable to binary (yes/no) responses. In addition, for purposes of scale development, items can be selected on the basis of their association with the trait of interest.

Group Differences or Discriminant Evidence

An index that is intended to distinguish between categories of respondent (e.g., well and sick, before and after treatment) may be tested by applying it to samples of each group and by analyzing the scores for significant differences. Significant differences on the scores would disprove the null hypothesis that the method fails to differentiate between them. This approach is very frequently used, but like all other validation procedures, it suffers logical limitations. Screening tests for emotional disorder may compare psychiatric patients with the general population but they will underestimate the adequacy of the method if some members of the general population have undiagnosed emotional disorders. Many indices are developed for highly selected cases in clinical research centers. Just as it is risky to generalize the results of clinical trials from tertiary-care centers to other settings, so referral patients may score highly on an index, yet high scores in a community survey may not be diagnostic (83). New measurements should be retested in a variety of settings.

When the results of several drug trials are combined in a meta-analysis, the average impact of the treatment may be indicated using the effect size statistic to compare mean scores in treatment and control groups or before and after treatment. Results are expressed in standard deviation units: $(M_t - M_c)/SD_c$ The effect size is comparable to a z score where, assuming a normal distribution of scores, the effect size indicates how far along the percentile range of scores a patient will be expected to move following treatment. With an effect size of $+1.00$, a patient whose initial score lies at the mean of the pretreatment distribution will be expected to rise to about the 84th percentile of the control, or untreated, group after treatment. Meta-analyses are

now beginning to summarize the results of several trials in terms of the effect sizes achieved by selected treatments; an example is given by Felson et al. (92).

The idea of effect size for a treatment can be turned on its side and applied to measurement instruments to compare how sensitive they are in indicating change—an example is given by Liang et al. (93). For outcome measurements, sensitivity to change is a crucial characteristic, although there is no clear consensus as to how this should be assessed. Many indicators have been proposed, and careful reading of articles is often needed to identify which statistic has been used. Most indicators agree on the numerator, which is the raw score change; there is little agreement on the appropriate denominator. The basic measure of effect size is the raw score change on the measure divided by the standard deviation of the measure at time 1 (94). Alternatives include a t-test approach which divides the raw score change by its standard error (95), or a responsiveness statistic that divides the raw score change by the standard deviation among stable subjects (95); as an alternative, the standardized response mean divides the mean change in score by the standard deviation of individuals' changes in scores (96). Further alternatives include relative efficiency (97), measurement sensitivity, receiver operating characteristic analyses (in terms of the ratio of signal to noise) (98, 99) and an F-ratio comparison (100, 101). Refinements to the basic formulae correct for the level of test–retest reliability (102). The largest effect size that can be attained for a given measure in a given sample is the baseline mean score divided by its standard deviation. As a general convention, effect sizes of 0.2 to 0.49 are considered small; 0.5 to 0.79 are moderate, and 0.8 or above are large (94). These interpretations give a guide to assessing the clinical importance of an intervention: while large sample sizes may make the comparison between drug and placebo groups significant, the effect size can be used to indicate whether the difference is clinically important. Knowing the effect size of an instrument is also valuable in calculating the power of a study and the sample size required; as with all sample size calculations, this requires a prior estimate of the likely effect size for a particular comparison (say, of drug and placebo), with a particular measurement instrument. To illustrate, if the effect size is expected to be 0.2, about 400 subjects will be required per group to show the contrast as significant, setting alpha at 0.05 and power at 0.8. To detect a moderate effect size of 0.5, one would need about 64 patients per group (94, pS187). Effect sizes also help in comparing results from studies that used different measurements: an instrument with a large effect size will show a higher mean improvement following treatment than an instrument with a smaller effect size. Information on the effect sizes of various measurement instruments is starting to accumulate; one example is in the depression field (103).

Construct validation of a health measurement is part science and largely art form. Construct validity cannot be proved definitively, but it is a continuing process in which testing often contributes to our understanding of the construct, following which new predictions are made and tested. This is the ideal, but the literature contains examples of inadequate and seemingly arbitrary presentations of construct validity. We still see statements of the type, "Our instrument correlated 0.34 with the Serenity Scale, 0.21 with the Idiosyncrasy Index, and 0.55 with the Turpitude Test and this pattern of associations supports its construct validity." In the absence of *a priori* hypotheses of association, the reader may well wonder what pattern of correlations would have been interpreted as refuting validity. Mercifully, there are several examples of a more systematic approach to construct validation, such as those by McHorney et al. (104) and Kantz et al. (105). Good validation studies state clear hypotheses and test them, with justification for considering those hypotheses the most rele-

vant. Good studies will also try to disprove the hypothesis that the method measures something other than its stated purpose, rather than merely assembling information on what it does measure. A variety of approaches should be used in testing any index, rather than relying on a single validation procedure. Future developments should see greater use of recently developed structural modeling techniques that allow theoretical and unmeasured variables to enter into multivariate analyses of the links among a series of measurements. An example of this approach to construct validation is given by Andrews (106).

ASSESSING RELIABILITY