Measuring Functioning and Well-Being

The Medical Outcomes Study Approach

With a foreword by Alvin R. Tarlov

Anita L. Stewart & John E. Ware, Jr., Editors

Duke University Press Durham and London 1992

18. Methods of Validating MOS Health Measures

Anita L. Stewart, Ron D. Hays, and John E. Ware, Jr.

Definition and Kinds of Validity

Validity refers to the extent to which a score measures what it is intended to measure and does not measure what it is not intended to measure (Anastasi, 1976) and to the extent to which a measure is useful scientifically (Nunnally, 1978). Validity studies increase understanding of the meaning of a score and the meaning of differences or changes in that score (Ware, 1984). Validating a health measure or a set of health measures is the process of accumulating many different kinds of evidence to determine the most appropriate interpretations(s) of a health score.

Since each piece of evidence only adds or takes away support for a particular interpretation, there is no clear point at which measures are considered valid. Therefore, the validation of health measures cannot be thoroughly assessed in a single study. Because health measures can be used for different purposes, their validity needs to be evaluated separately for each purpose. Health measures that are valid for one purpose (e.g., measuring outcomes) will not necessarily be valid for another (e.g., predicting demand for services) (Ware, Brook, Davies, et al., 1981).

A variety of different approaches help give understanding to the meaning of health measures. No standard guidelines are available for validating health measures. Thus, standards have been derived from those used to validate psychological and educational measures. Three categories of validity evidence are identified by the American Psychological Association, the American Educational Research Association,

and the National Council on Measurement in Education (American Psychological Association [APA], 1985): content related, criterion related, and construct related. Content-related validity is a more qualitative approach than the other two, which are based on studies of empirical relationships. Nunnally (1978) refers to the same three types, labeling the second one predictive validity. These categories are intended as conveniences rather than to suggest that there is a rigid distinction among them (APA, 1985). Evidence of one type of validity is often evidence of another. Empirical approaches to validity include known groups, predictive, convergent/discriminant, multitrait-multimethod, and factorial. Depending on the purpose of the measure, some of these are criterion-related approaches under some circumstances and construct-related approaches under other circumstances. Four additional approaches provide useful validity information: evaluation of the interpretability of scores, the need for additional measurement using incremental validity analysis (Sechrest, 1967), examination of whether measures contain response bias, and validity generalization. The MOS has adopted a definitional structure (see Table 18–1) that best serves the validation of health measures based on an evaluation of the many different ways of defining validity.

This chapter presents an overview of the essential steps necessary to validate health measures and indicates the types of validation studies that have been performed on the MOS measures.

Content validity pertains to how well items in a measure, or concepts in a set of measures, sample a specified universe or domain of content (APA, 1985; Nunnally, 1978; Ware, 1984, 1987). Evaluation of how well the important aspects of a concept are being measured can be assessed at the level of an entire battery of health measures, within the components of a particular health concept, or within a single scale.

To perform content validation requires a definitional standard against which the concepts or items are compared. Standards can be based on well-accepted theoretical definitions, on existing accepted standards, or on interviews with those experiencing the types of problems studied. In developing the Sickness Impact Profile, for example, patients were interviewed to determine the full range of impact of disease on behavior (Gilson, Gilson, Bergner, et al., 1975). Ware (1987) has provided a set of minimum standards by which to evaluate the content validity of a battery of health measures based on an extensive review of existing measures.

Content validity is important during the construction and develop-

Table 18–1 Overview of Approaches to Testing Validity

Type of Validity	Definition and Examples
Content validity	**Are all relevant concepts represented in the measure or set of measures?**
Content validity of a battery	Are all aspects of functioning and well-being represented in the set of health measures?
Content validity of a scale	Are all aspects of a definition of a concept represented in a scale?
Criterion validity	**Does the measure correlate highly with the "gold standard" measure of that concept?**
Criterion validity	Does a new measure of depression correlate with the "gold standard" measure (e.g., Diagnostic Interview Schedule of the DSM-III)?
Criterion-related validity	Does a short-form measure of physical functioning correlate highly with a validated long-form measure of physical functioning?
Predictive validity	Do scores on a measure of health perceptions predict whether or not people use any health services in the following year?
Construct validity	**Do the measures correlate with measures of other variables in hypothesized ways?**
Convergent validity	Does a measure of pain intensity correlate with a measure of the effects of pain?
Discriminant validity	Does a measure of physical functioning correlate lower with a measure of mental health than with a measure of mobility?
Multitrait-multimethod approach	Does a self-reported measure of depression correlate higher with an observer rating of depression than it does with a self-reported measure of anxiety?
Known groups validity	Is the mean health perceptions score significantly lower for a group of patients than for a general population sample, as hypothesized?
Factorial validity	Are two underlying constructs (factors) of physical and mental health defined by the MOS health measures, as hypothesized?
Interpretability of scale scores	**What is the meaning of a score or a change (or difference) in a score?**
	What is the difference in mean health perceptions scores as a function of having arthritis?
	What is the amount of change in a physical functioning score achieved by providing an effective pain relief medication for angina?

Table 18–1 (*Cont.*)

Type of Validity	Definition and Examples
Incremental validity	**Is a substantial gain in information achieved by adding items to a scale or scales to a battery of health measures?**
	Does a mobility measure add a substantial amount of information over and above a physical functioning measure?
Response bias	**Are scores systematically lower or higher due to response bias?**
	Are scores of social functioning significantly correlated with a measure of socially desirable responding?
Validity generalization	**Are the measures valid in different populations?**
	Are the correlations among the measures similar in a patient population and a general population?

ment of measures (APA, 1985; Nunnally, 1978) and is a prerequisite to empirical validity. If the content of a concept is not adequately reflected, it is difficult to establish the more empirical types of validity.

The content validity of the MOS health measures was evaluated in terms of individual measures, the more common approach, and in terms of the entire set of measures to determine how well the set represents a comprehensive definition of health and to assure that the measures are not confounded with each other (Ware, Brook, Davies, et al., 1981).

For the entire MOS battery, the literature was relied on for the definitional standard of content validity because most relevant health concepts have been identified. For particular concepts, such as psychological distress/well-being, content validity studies ascertained whether all components were represented (e.g., whether they included depression, anxiety, positive affect). Within each scale, content validity determined whether the items were comprehensive in terms of the definition of that scale (e.g., whether all aspects of physical functioning were represented). For new measures, a large item pool was first created from several existing key instruments from which pilot study items were selected. Pilot studies typically were of measures with a large number of items. In evaluating pilot study results, one of the MOS criteria for

retaining items was comprehensiveness (i.e., to represent the fullest range of components of each measure). Comprehensiveness at the item level also included attention to the range of levels tapped by items in the scale. For example, the MOS assured that the physical functioning scale reflected the performance of a full range of activities, from simple self-care to more difficult or vigorous ones.

Criterion and Construct Validity

Criterion validity demonstrates that test scores are systematically related to one or more outcome criteria (APA, 1985). One type of criterion validity involves the testing of a new measure in terms of how well it predicts an accepted "gold standard" measure. In the MOS, for example, a short screening measure of depression was developed by selecting items that best correlated with the gold standard measure of depression (Burnam, Wells, Leake, et al., 1988). Another type of criterion is some future event (e.g., poor health, death) that is predicted with the health measures.

Criterion validity is assessed by using scores on one measure to predict scores on the criterion (Anastasi, 1976; Cronbach, 1970). The higher the correlation between the measures, the stronger the evidence favoring criterion validity (Nunnally, 1978). MOS single-item measures were correlated with their respective long-form measure, which is considered a form of criterion-related validity.

When a measure is used to estimate some behavior external to the measure itself, the behavior is referred to as the "criterion," and the analysis is termed predictive validity (Nunnally, 1978). The predicted behavior can be concurrent in time or a future prediction. The latter is more common in health studies. If measures are to be used to identify patients likely to use a lot of health care services or patients at risk of poor health in the future (who thus may be in greater need of care), then the validity of the measures must be tested in terms of their ability to predict subsequent utilization and subsequent health. The prediction can be made in terms of group membership (e.g., those who have had any hospitalizations) or in terms of a continuous measure (amount of expenditures for health care). In general, the higher the correlation, the better. For the MOS measures, predictive validity tests involved examining the correlations between the measures and subsequent utilization (chapter 19).

The use of scores to predict some future event, such as utilization, can be considered criterion-related validity if the purpose of the score is to identify people who will be high utilizers. The same analysis can be considered evidence of construct validity if the purpose of the analysis is to confirm a hypothesis that a health measure is associated with subsequent utilization.

Construct Validity The basic issue in studies of construct validity is whether the health measure relates to other measures in ways consistent with plausible hypotheses. Patterns of relationships are hypothesized between the measure being validated and measures of other variables, and data are analyzed to see if the hypothesized patterns are confirmed empirically (Cronbach and Meehl, 1955; Nunnally, 1978). Hypotheses are usually stated regarding the direction and sometimes the strength of relationships that might be expected based on theory and literature. Validity is supported when the associations conform to the hypotheses. It is important that hypotheses be firmly grounded in theory. If the logic of the hypotheses is poorly thought through, a null finding might be taken as an indication of poor validity when in fact it reflects a true situation (McDowell and Newell, 1987).

Variables most often used to test patterns of relationships for health measures are utilization of health services, other general health measures, clinical measures, and mortality. Early studies of the validity of health measures included age as a validity variable, but because age does not consistently correlate with many health variables, we no longer consider it a good validity indicator.

The correlation of health measures with clinical status measures is a type of construct validity because certain measures of functioning or well-being can be hypothesized to be correlated with certain clinical measures. In this case, the clinical measures are not regarded as criteria but rather as constructs that should be related to the health measures. The best clinical measures for this purpose are those for which a strong hypothesis can be stated. The MOS examined the decrement in functioning and well-being associated with the presence of a variety of chronic medical conditions and depression to determine whether the decrement corresponded to the hypotheses about relative effects of those conditions (Stewart, Greenfield, Hays, et al., 1989; Wells, Stewart, Hays, et al., 1989). This process can also be considered a type of known-group validity.

For new measures, the evaluation of correlations of that measure with other measures provides the first step in beginning to understand

the meaning of measures. In such cases, relationships between the new measures and other health concepts were evaluated. In these cases, the MOS was less interested in testing hypotheses about the nature of the relationships than in building a knowledge base about the meaning of the measures. For example, because the sleep measures are new, the purpose of the MOS studies of the relationships between the sleep measures and a variety of other health measures was to understand better the meaning of the sleep measures.

Convergent Validity The most commonly used type of construct validation is convergent validity, which focuses on the extent to which several measures of the same concept correlate with each other. Convergent validity at the item level was an essential part of the MOS method of scale construction (chapter 5).

Most of the MOS measures of functioning and well-being were based on the same method (self-report), but because the MOS often developed several subscales pertaining to the same construct, the convergent validity of those subscales was evaluated. For example, measures of physical functioning, mobility, and satisfaction with physical abilities were expected to correlate at least moderately with one another because they all assess physical functioning. Product-moment correlation coefficients and factor analysis were used to test these relationships. In factor analysis, measures that in fact are assessing the same underlying construct should correlate with, or load on, the same factor.

Discriminant Validity It is also important to demonstrate that a measure does not correlate with other measures that are intended to be different. For example, a measure of physical functioning would not be expected to be highly related to a measure of depression or of loneliness. Because all health measures are somewhat related, significant correlations are often observed among nearly all measures, especially in large samples like the MOS. Thus, discriminant validity is usually tested in relation to the correlations observed in the convergent validity studies. That is, the appropriate test is whether measures correlate lower with measures to which they are not expected to be related than they do with measures to which they are expected to be related.

Because item convergence and discrimination were an essential part of scale construction in the MOS, the discriminant validity of many of the measures was enhanced as a result of this process. Additional discriminant validity studies were then conducted using the final measures.

Multitrait-Multimethod Approach When more than one method of

data collection or scale construction has been used, a variant of discriminant validity is possible. This elegant approach—the multitrait-multimethod (MTMM) procedure developed by Campbell and Fiske (1959)—blends convergent and discriminant validity. An MTMM matrix comprises intercorrelations of two or more health concepts (traits) measured by two or more methods. In the MTMM approach, a measure is expected to correlate significantly higher with other measures to which it should be (i.e., theoretically) related than with other measures to which it should not be related. Because all measures reported here are based on self-report, MTMM studies were not performed.

Known-Groups Validity Because one of the purposes of generic health measures is to detect the impact of disease and treatment on patient functioning and well-being, one test is how well the measures discriminate between groups known to differ in that health concept because of a particular disease. Known-groups validity involves comparisons of mean scores on a health measure across groups known to differ in the concept being validated (Kerlinger, 1973). For example, it was hypothesized that mean physical functioning would be lower in a group of patients than in a general population. Similarly, mean perceived health might be expected to be poorer in groups with a more severe disease (e.g., heart disease, cancer) than in those with a less severe disease (e.g., back problems, hypertension). Again, the evaluation of the generic measures in relation to clinical measures, such as the presence of various diseases, is especially useful in bridging the gap between these measures and those more familiar to clinicians. To represent a test of validity, it is important that the defined groups be clearly known to differ in the concept or construct being tested.

The MOS tested mean scores for the same measures administered to a patient sample and a general population sample for six short-form measures (Stewart, Hays, and Ware, 1988). Mean role limitations due to emotional problems scores were compared across groups of patients with and without depression. The MOS also determined the extent to which measures discriminated among nine chronic medical conditions defined in terms of physician report and patient report (Stewart, Greenfield, Hays, et al., 1989) as well as depression (Wells, Stewart, Hays, et al., 1989).

Factorial Validity When factor analysis is used to evaluate the structure of a set of scales, tests of the validity of the measures in relation to the underlying constructs can be performed. The same logic that leads to predictions about relationships among a set of variables can be applied to a particular concept. For example, if one underlying concept is expected to be represented by a set of measures (e.g., health perceptions), then one dimension or factor should be observed empirically using that set of measures.

The validity of a large set of measures was tested by evaluating their interrelationships in terms of their underlying structure, empirically testing the conceptual framework of health (chapter 2). Essentially, the MOS tested the presence of two underlying health dimensions: physical and mental health. Three patterns of correlations between measures and factors were hypothesized, namely, that some measured primarily physical health, others measured primarily mental health, and some measured both. Hypotheses were tested using confirmatory factor analysis, an analytic extension of exploratory factor analysis that allows for evaluation of a specific structure as defined by a pattern of factor loadings and factor intercorrelations (Long, 1983).

The Interpretability of Scale Scores

One aspect of validity pertains to how to interpret the score. Understanding what scale scores mean is confounded by two questions: the meaning of different scale levels (e.g., what does a score of 2 or of 71 mean?) and the meaning of differences or changes in scale units (e.g., is a 5-point difference meaningful?).

Meaning of Score Values For single-item measures with verbal response categories, the meaning of each score is quite clear. For example, the rating of health as excellent, very good, good, fair, or poor usually has scores ranging from 1 to 5; thus, each numeric score has a verbal interpretation. However, when multiple items are combined into a score, scores are possible over a broad range of numbers, and the score has no inherent meaning. The MOS provided several aids to understanding the meaning of scores. First, because the first clue to the meaning of the scale score itself (e.g., a group mean) comes from its relation to the endpoints, most measures were transformed to 0–100 scales, where the minimum possible score is 0 and the maximum 100. This process eliminates the problem of trying to interpret scores reported in many journal articles that all have different units.

Second, all measures were labeled to reflect the direction of scoring of the measure, and information on the direction of scoring is provided in all variable labels and tables so that it is immediately apparent

whether a high score means better health or poorer health. This facilitates interpretation of correlations and of differences in mean scores. For example, if a variable is labeled "pain severity," then a high score indicates more pain, and a high score on "physical functioning" indicates better functioning.

Third, the MOS used a normative approach. When scales are normed in a general population or in some representative sample, the mean and the standard deviation become the "norm" against which scores of various subgroups can be compared. For the MOS Short-Form 20-item General Health Survey, two such normative samples are provided: a general population (chapter 16), and a representative sample of patients visiting a variety of providers (Stewart, Hays, and Ware, 1988). These can thus be used as comparison groups against which to compare scores obtained in other populations and samples. The MOS sample on which the other measures were developed consists of patients selected because they have one or more of the MOS chronic conditions. This sample is thus less "normative" than those available for the short form. Nevertheless, some understanding of the meaning of scores in other samples can be obtained by comparing them to the overall MOS sample mean and standard deviation.

Finally, some single-item measures have verbal descriptive categories for the different response choices. By calculating mean multi-item scale scores for the various descriptive categories, the meaning of the multi-item scale scores is better understood (Deyo and Patrick, 1988). To be useful, the verbal descriptor scale must be one that is itself reliable and valid. For selected concepts, mean multi-item scale scores were calculated for each level of the parallel single-item measures (e.g., mean physical functioning scores for the five levels of the single-item physical functioning measure). Although this calculation was done to understand the properties of the single item, the same results can also facilitate interpretation of the meaning of the various multi-item scale scores.

Meaning of Score Differences Because the main purpose of using health measures is usually to determine whether various treatments or types of health care are effective, it is of great interest to understand the meaning of a difference in scores or a change over time. For example, what is the meaning of a 5-point or a 10-point change in that score? Or perhaps better, what is the meaning of a difference in the amount of 0.1 standard deviation units, or of 0.5 standard deviation units? Such knowledge provides very important information that can be used in designing subsequent studies with enough precision to test hypotheses. However, a 5-point difference at one end of the scale may not mean the same thing as a similar difference at the other end of the scale. That is, a 5-point improvement at the severe end of the scale may mean far more to patients than a 5-point improvement at the healthy end of the scale.

This evaluation can be referred to as the calibration of measures or effect size analysis (Ware, 1984). Calibration involves documenting the meaning of differences in health scores by expressing them in terms that are relevant to patients, clinicians, or society. An example of such an analysis is to determine the percentage of those who received mental health care whose scores on psychological distress improved, or to examine the change in a score resulting from the onset of an acute chronic condition.

Deyo and Patrick (1988) discuss several issues regarding the responsiveness of scales to clinically meaningful changes, which can be regarded as evidence of the meaning of score differences. Deyo and Inui (1984) and Deyo and Centor (1986) present some very useful methods for testing the sensitivity of scale scores to clinical changes. For example, Deyo and Inui (1984) compared scores on various dimensions of the Sickness Impact Profile to clinical judgments of improvement or deterioration.

The MOS has contributed some information on the meaning of some of the health scores by documenting the unique decrements in health due to each of nine chronic medical conditions (Stewart, Greenfield, Hays, et al., 1989). For example, the presence of hypertension was reflected in a 3-point decrement in the current health perceptions measure.

Incremental Validity

In selecting a battery of measures, the uniqueness of each measure is an important consideration. This is an issue of incremental validity. Empirical studies of incremental validity address two basic issues: whether a long-form measure contributes sufficiently more information over and above a short-form measure and whether the addition of another health concept to a comprehensive battery contributes sufficiently more information over and above the initial set of concepts to justify the additional respondent burden and data-collection costs.

Short-form measures of health are less expensive to administer and

are more popular with respondents; however, it is important to know that the use of short-form measures does not result in a serious loss of information compared to long-form measures. Thus, it is helpful to test the incremental validity of the long-form measures in relation to the short-form measures (Sechrest, 1967). In such tests, the increment in predictive power associated with the long-form measure is compared with the short-form one. For example, perceived current health could be predicted with a short-form physical functioning measure and then a long-form physical functioning measure added to the model to determine whether it significantly increases the variance explained. If it does, the long-form measure contributes a significant amount of additional information. The gain in predictive ability of the long-form measure must be weighed against its added cost or burden. In doing such tests, it is important to assure that any gains in prediction are not due simply to improved reliability (with additional items) rather than to the improved validity achieved by adding concepts or enriching a concept.

When the initial measures were being developed, the incremental validity of the MOS long-form measures was tested in several pilot studies using this method. The reasoning was that if an acceptable short-form measure contained as much information as the long form, then resources need not be expended on the long-form measure.

Response Bias

Two possible sources of bias that sometimes threaten the validity of health surveys and that often go along with self-reported methods (Bentler and Eichberg, 1975) are socially desirable response set (SDRS) and acquiescent response set.

Socially Desirable Response Set (SDRS) Socially desirable responding, a tendency to describe oneself in socially desirable or favorable terms, is among the most important sources of response bias in self-report research. Socially desirable responding affects the validity of self-reports because it results in underreporting of socially undesirable characteristics or overreporting of socially desirable behavior (Nunnally, 1978). There are two basic approaches to handling SDRS: minimizing its occurrence in the design of the question asked and measuring it so its effects can be evaluated and controlled for.

A number of instruments have been developed to measure the tendency to give socially desirable responses (Crowne and Marlowe, 1960; Jacobsen, Brown, and Ariza, 1983). These SDRS measures are typically included in the same questionnaires as the health measures. The correlations between the health measures and the SDRS measures are evaluated to determine the extent to which SDRS is present in other self-report measures (Crowne and Marlowe, 1960; Edwards, 1970). If SDRS is significantly correlated with a self-report measure, SDRS can be statistically controlled for in analyses involving that measure. Using such measures, SDRS has been found to be significantly correlated with depression, symptoms, life satisfaction, health-related behavior, problem drinking, and life events (Klassen, Hornstra, and Anderson, 1975; Kristiansen and Harding, 1984).

In the MOS, a brief SDRS measure was developed (Hays, Hayashi, and Stewart, 1989) and included in the validity studies combined with an evaluation of the relationship between SDRS and each of several health measures (chapter 19). Statistically significant correlations of SDRS with health measures are regarded as indicative of a problematic degree of SDRS in the measure.

By selecting appropriate methods for measuring those characteristics most likely to be susceptible to socially desirable responding, its occurrence in these measures can be minimized. For example, SDRS tends to be less problematic in mail surveys than it is in telephone and face-to-face interviews (Dillman, 1978). Because items differ in their susceptibility to social desirability responding (Edwards, 1970), its effects can be minimized when items are written. Thus, in developing item stems and response choices, the potential influence of SDRS was minimized wherever possible, following the recommendation of Smith (1967). Value-laden words were avoided and instructions were written to facilitate accurate, socially undesirable responses. Johnson's (1981) suggestion that "the best strategy for designing a valid scale is not to make lying or misrepresentation difficult, but to make self presentation as easy as possible" was adopted. Sometimes item response choices were reversed to put the one that was least "desirable" first. A response option that appears to the extreme left in a row of options tends to be selected more often than when it appears on the extreme right (Mathews, 1929).

Acquiescent Response Set Two types of acquiescent responding have been noted: agreement acquiescence and acceptance acquiescence (Bentler, Jackson, and Messick, 1971). Agreement acquiescence is a tendency of respondents to agree with statements regardless of content. Acceptance acquiescence refers to a tendency to accept characteristics

as descriptive of oneself. Thus, acceptance acquiescence is denoted by agreement with items that describe characteristics and disagreement with items that deny characteristics. Acquiescence tends to occur more often when questions are ambiguous, lengthy, complicated, or otherwise difficult to understand. It also occurs more frequently in people with less education (Converse and Presser, 1986).

Both forms of acquiescence can be minimized by keeping questions simple, clear, and short, which the MOS attempted to do wherever possible. Agreement acquiescence can also be minimized by using several items to measure each concept, some with favorable wording and some with unfavorable wording (e.g., "do you have energy?" "do you feel tired?"). When these items are summed into a multi-item scale, agreement acquiescence tends to cancel out. The latter has been demonstrated for measures of health care attitudes (Ware, 1978). In selecting a final set of items for each measure, the MOS attempted wherever possible to achieve a balance between positively and negatively worded items. In some cases, however, this was impossible because some concepts are by definition negative (e.g., psychological distress is measured using only negatively worded items).

Validity Generalization

Given that the MOS wanted these generic measures to be appropriate for both patient and general populations and for severely ill as well as mildly ill groups, an important question is whether the validity of the measure is the same across these different groups. That is, can the validity evidence obtained with one type of sample be generalized to other samples, or must additional evidence be obtained in the new sample (APA, 1985)? Average scores were expected to differ in different groups, but the meaning of the scores should be stable across the groups (as reflected in the patterns of correlations and other empirical validity evidence). As evidence accumulates of similar validities in samples other than the MOS, validity generalization is increasingly assured. To the extent that the measures have different meanings in different groups, validity studies in different samples need to be performed until the generalizability is assured.

For many of the measures, the MOS had considerable prior experience in a general population (in the Health Insurance Experiment). Wherever this was the case, correlations among measures were compared with those obtained in the HIE to assure generalizability. The MOS evaluations of the structure of the short-form health measures were compared between two populations (Stewart, Hays, and Ware, 1988).

Discussion

This chapter provides both a methodological guide to the assessment of the validity of health measures and a summary of the approaches to validation that were taken in the MOS. Because many of the standards for validating health measures have been derived from those for validating measures in education and psychology, specific standards for validating health measures need to be developed. The issues in health assessment are more specific, and the validation process depends more on the purpose of the measure. Such new guidelines could provide needed clarity and increase the quality of available health measures. Such standards could be published, as are those in psychology and education (APA, 1985). The methods outlined in this chapter are intended to provide a first step toward such guidelines. We agree with McDowell and Newell (1987) that the validation of health measures is to a large extent an art form. If specific standards were outlined for health measures in particular and contributions obtained from a variety of investigators in health measurement, the "art" could be reflected in those standards.

The approach to validity outlined in this chapter is comprehensive and represents an ideal toward which to strive. When a reliable and variable measure has been subjected to preliminary validity studies, it can be used in analytic studies. Such studies can become sources of new validity information at the same time as they answer important clinical and policy questions. Because of the importance of validity, and because relatively little attention is generally given to validating health measures, future studies should incorporate some validation strategies.

The validity studies presented in this book are based on self-report data, and most are based on cross-sectional analysis. Because of the importance of using multiple perspectives, additional validity studies currently underway include tests of associations of these health measures with other measures based on clinical assessment as well as on measures occurring at a later point in time. For example, the relationship of these measures to measures of disease severity defined in terms of

both physician-reported information and an independent clinical assessment will be tested. The MOS will evaluate the measures in relation to subsequent mortality and to changes in disease severity over time using prospective data.

Many of the health measures presented here are derived from measures that have been widely used and validated. The validity history of a measure is thus an important component. The prior validation evidence for those measures with a prior history (e.g., health perceptions and psychological distress/well-being) was reviewed.

The methods of validity presented in this chapter and applied in the MOS pertain primarily to tests of validity for measures to be used for research purposes. If measures are to be used for purposes of assessing individual patients in the offices of clinicians, additional standards are needed. For example, if a measure is to be used to identify patients in need of particular clinical interventions, then validation studies must be designed to assure that the measure is capable of doing so accurately. These types of studies are one of the next steps needed as the demand for measures that can be used by clinicians in everyday practice increases.