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

5. Methods of Constructing Health Measures

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

This chapter summarizes the MOS approach to writing questionnaire items, pretesting measures, constructing multi-item scales, evaluating scale variability, reliability, and stability, and labeling measures. Methods for validating health measures are presented in chapter 18.

In many instances, a single question might be all that is necessary to obtain the desired information. For example, to learn a person's age, sex, or weight, a single question suffices. In research on health and health-related matters, however, the concepts are complex and difficult if not impossible to define by a single item. Multiple items are needed to operationalize adequately each construct. The use of multiple items allows one to "cast light at different angles" (Converse and Presser, 1986). As one example, definitions of depression contain many components such as feeling blue, depressed, and hopeless. Thus, several questions are needed to represent all components of the definition.

Multi-item measures, which were constructed wherever possible, have several advantages over single-item measures: they reduce the number of final scores necessary to define each variable, assuming that separate scores would have been used from the separate items; they increase score reliability by pooling the information that items have in common; they increase validity by providing a more representative sample of information about the concept; they increase score variability and hence sensitivity; they minimize bias caused by individual tendencies to endorse items (acquiescence) or negate them (nay-saying) regardless of content (in cases where both favorably and unfavorably worded items are combined); and they provide the option, if item responses are missing, to estimate scores using other items in the measure, thus reducing missing scores on the multi-item scale.

Writing Items

After a content area was specified, items were written to operationalize each concept. For the MOS, closed-ended questions with a specific set of responses were used. In large-scale studies, open-ended questions are more burdensome and often yield uninterpretable answers, especially from the less educated. Open-ended questions are time-consuming to administer and require experienced coders; thus, they may be unsuitable for self-administered surveys (Dillman, 1978). However, during the pilot phase, some open-ended questions were used to identify any problems missed in structured questions.

Item Content The first step in writing items is to decide precisely what one wants to know. Part of operationalizing the concept being measured includes deciding whether to ask about frequency (how often a symptom occurred—never, occasionally, or often), intensity (how extreme it was—mild, moderate, or severe), or duration (how long it lasted—a few minutes, several hours). With respect to functioning, questions can compare a person's present level with their usual level or with other people their age. Questions about limitations in functioning can ask about limitations due to health, limitations due to a particular condition, or limitations regardless of cause. Questions can ask if people are having trouble doing certain things or need help. These decisions need to be consistent with the definition of the concept and thus warrant considerable preparatory thought to maximize content validity (Ware, 1987).

Item Stems Each item consists of the item stem and the response options. The item stem is the portion of the item that states the issue or question asked of the respondents. Consistent with traditional standards, item stems were short, simple, easy to understand, and restricted to one idea. Double negatives and ambiguous terms were avoided to maximize certainty of meaning. Information on how to write good items can be found in Fowler, 1984, or Converse and Presser, 1986.

Item Response Options The goal in the selection of item response choices was to pick a set of options that would provide approximately interval-level information. To achieve this, three important features of response options were considered: what type of response intervals to use, whether to offer a middle "neutral" category, and what the number of options should be.

Four basic types of response options—endorsement, frequency, intensity, and comparison—were used. Table 5–1 illustrates the three response options that were most frequently used. The endorsement option was used in assessing statements of perceived health such as "I have been feeling bad lately." The frequency option was used to assess various subjective states, such as energy/fatigue and anxiety. The intensity option illustrates a set of verbal choices for rating the severity of a symptom, such as pain. For intensity, a numbered response scale was sometimes used. For example, when asking patients to rate their pain by providing the numbers 1 to 20, endpoints were labeled "no pain at all" and "pain as bad as you can imagine." Whenever possible, similar response choices in different batteries were used so respondents would feel familiar with a limited set of choices.

Although most would agree that the various responses are ordered in terms of increasing levels (i.e., they yield at least an ordinal scale), it is never clear whether such "imprecise quantifiers actually have some common meaning" (Bradburn and Sudman, 1980). As Bradburn and Sudman point out, the meaning depends on the context in which the question is asked (e.g., which types of questions preceded it) and can vary across individuals. Standardized administration procedures minimize differences in interpretation by respondents.

When a final measure is a single item, it is important to understand the intervals between the response categories, because single items tend to have a "custom" set of response choices for which the intervals are likely to be uneven. For example, the difference in health from "excellent" to "very good" may be smaller than the difference between "good" and "fair" and should be reflected in the score. Thus, when a long-form measure of the same concept was available, the distance between categories was sometimes estimated empirically by calculating

Table 5–1 Response Choices Used to Measure Endorsement, Frequency, and Intensity

Endorsement	Frequency	Intensity
1–Definitely	1–All of the time	1–None
2–True	2–Most of the time	2–Very mild
3–Don't know	3–A good bit of the time	3–Mild
4–False	4–Some of the time	4–Moderate
5–Definitely false	5–A little of the time	5–Severe
	6–None of the time	6–Very severe

mean long-form scores for each response level of the single item. This enables the intermediate response levels (i.e., not the extreme levels) to be transformed using interpolation to reflect the intervals observed based on long-form mean values.

With endorsement scales, a neutral (e.g., "don't know") category is often offered to respondents to provide an option for people who have no opinion on a particular question or to provide an additional level of gradation. There is controversy about the usefulness of a neutral category; some argue it should not be offered because people choose it instead of being more committal (Converse and Presser, 1986). The MOS approach was to include it because it provides a valid response.

Regarding the number of item responses, several studies suggest that five to seven well-chosen response categories provide the lower bound necessary for optimal assessment of a measurement domain (Bollen and Barb, 1981; Johnson and Creech, 1983; Johnson and Dixon, 1984). Others note that most people cannot consistently discriminate their feelings beyond a 7-point classification (Osgood, Suci, and Tannenbaum, 1957, cited in Wells and Marwell, 1976), thus suggesting that more than seven categories are unnecessary. Items administered with five to seven response options have been shown to correlate strongly with corresponding items administered with a greater number of response options (Hays and Huba, 1988). Based on these guidelines, five or six options were selected most often.

Writing Multiple Items The optimal approach to measurement, combining several items into a single score, makes it possible to combine a variety of approaches to item content in a set of items. If frequency is chosen as the response option for a set of items, the item stems can reflect a range of severity (e.g., "feeling blue," "thinking about suicide"). If intensity is the response option, questions about usual intensity and intensity at its worst can be asked.

An advantage of combining items that have different response options (e.g., frequency and intensity), is that method effects are reduced. When this is done, it is helpful if the items have the same number of response options in order to enhance the likelihood that they will have equivalent variances and contribute equally to the combined score. A disadvantage of having different response options within one scale is that they cannot easily be administered over the telephone. In selecting psychological distress/well-being items for purposes of a telephone-administered version, many of the item response sets had to be revised to be consistent. If it is more important to retain the different

types of response options, respondents can be provided in advance with cards containing the various options that will be read during the interview.

Time Frame A time frame needs to be specified for most questions, especially those asking about intensity or frequency of various states. The time frame needs to be short enough to allow accurate memory for the event being asked about, yet long enough to represent the general time within which the event is likely to occur and to allow for variation across respondents (Fowler, 1984).

For most of the measures, questions were asked about the past four weeks, an interval in which most people could recall health events yet which would provide a reasonably stable sample of those events. This time frame helps to assure the assessment of the average occurrence of various health states rather than daily fluctuations. This interval is most important in the measures of psychological distress/well-being, where a smaller time frame assesses the "mood of the day" rather than the average levels of distress/well-being over a longer period of time.

The MOS rejected wording the items to say "during the past month," because of concern that respondents would think in terms of the last calendar month. Similarly, "the past thirty days" was rejected because thirty days also sounds like a calendar month. One exception to this nomenclature was the psychological distress/well-being items, which refer to the past month in order to maintain comparability with previous research. Because the cognitive functioning items were interspersed with the distress/well-being items, those items also refer to the past month.

Pretesting Items

In developing new items or in modifying existing measures, it is essential to pretest or to pilot test items to assure that they work. One researcher states, "Virtually every questionnaire could be changed in some way to make it easier for respondents . . . to meet the researcher's objectives" (Fowler, 1984). A preliminary pretest, which solves gross administration issues, might consist of a simple survey of a dozen people to judge the clarity of instructions, determine if the questions make sense, and estimate respondent burden. A full-scale pilot study might be conducted using 50–200 people similar to those in the main study. Others discuss what can be gained from adequate

pretesting and provide some guidelines (Converse and Presser, 1986).

In a preliminary pretest, the main source of information is a "debriefing" of the pretest subjects to determine which questions are confusing, difficult, or unclear and whether the subject understood the instructions or skip patterns. This method, where subjects are told that it is a pretest and understand that their task is to identify problems, is called a "participating pretest" (Converse and Presser, 1986). In a larger-scale pilot study, information can also be obtained from statistical evaluation of items.

The MOS conducted nine full-scale pilot studies of various measures: physical functioning, health perceptions, energy/fatigue, sleep, role functioning, pain, physical/psychophysiologic symptoms, family functioning, and sexual functioning. The purpose of the pilot studies varied according to the particular set of measures; however, they generally addressed administrative issues and tested a large item pool so the best items could be selected empirically. From a sample size of 50–100 patients, the MOS identified items with poor variability or a high percentage of missing responses as bad items and revised or eliminated them. Unclear instructions were clarified. By examining correlations among items, we identified items that did not converge (i.e., were not strongly related) with items intended to assess the same concept or that were too strongly correlated with items intended to measure distinct concepts. The importance of these pilot studies is exemplified by the expenditure of more than one year on this phase of the MOS. The results were documented in project memoranda and are summarized in corresponding chapters.

Techniques for Combining Items into Scales

Groups of items that could possibly be combined into a single score were first hypothesized. Hypotheses were based on logical combinations of items appearing from their content to measure the same construct. For many of the MOS measures, hypotheses were well grounded in prior work on the measures.

Multitrait scaling was used as the method for evaluating the hypothesized item groupings. The most commonly used method for evaluating the underlying structure of a set of related concepts in order to develop a set of measures of those concepts is exploratory factor analysis (Ford, MacCallum, and Tait, 1986; Montgomery, Shadish, Orwin, et al., 1987). This method is appropriate for exploring the relationships among a universe of items in the beginning phases of developing an understanding of the underlying dimensions of the items and was occasionally used during the pilot testing of some of the measures. Exploratory factor analysis, however, has a number of limitations. The resulting item structure depends on the choices regarding the factor model (principal components or common factor analysis), the number of factors that are appropriate, the rotation method selected, and the other items that are included in the analysis. "The decisions made at each choice point can have a substantial impact on the results of the factor analysis and on subsequent interpretation of these results" (Ford, MacCallum, and Tait, 1986). In addition, the interrelationship of variables is left unspecified, and it is impossible to test directly alternative theoretical structures underlying the data.

When theoretical progress on a concept has advanced beyond exploratory development and the hypotheses about its underlying structure are fairly good, confirmatory analysis is more informative than exploratory analysis. Multitrait scaling was selected as the primary scaling method for use in the MOS because it is a confirmatory approach that allows a direct test of a priori structures.

Multitrait Scaling

Multitrait scaling is based on the traditional Likert (1932) method of summated ratings. When responses to several questionnaire items are summed into a single scale score, it is generally termed a summated or a "Likert-type" scale. Summated scales are constructed by summing the items in each hypothesized scale, assigning equal weights to the items. A simple example of a summated ratings scale illustrates this method. Suppose you have two items—"How often during the last four weeks did you have a lot of energy?" and "How often during the last four weeks did you feel tired?"—each with the following response choices:

(1) none of the time
(2) a little of the time
(3) some of the time
(4) most of the time
(5) all of the time.

A summated rating (Likert) scale of these two items involves first reversing the scores on the second item (i.e., 1 = all of the time, 5 =

none of the time) to make a high score on both items refer to energy, and then adding the item scores. The lowest possible score on this summated scale is 2, obtained from two minimum item scores of 1 each, and the highest possible score is 10, obtained from two maximum item scores of 5 each.

A number of analyses must be performed to determine whether a set of items can be appropriately combined into a summated rating scale. In multitrait scaling several scaling criteria are added to those usually associated with Likert scaling. Multitrait scaling follows five steps to determine whether:

(1) each item in a hypothesized grouping is substantially linearly related to the total score computed from other items in that group (a traditional criterion of convergence usually expressed in terms of internal consistency);
(2) each item correlates significantly higher with the construct it is hypothesized to measure than with other constructs (item discrimination criterion);
(3) item groups not hypothesized a priori are not identifiable from the data (factor analytic test);
(4) items in the same scale contain the same proportion of information about the construct (test for approximately equal item-total correlations); and
(5) items measuring the same construct have equal variances and therefore do not need to be standardized before combining them in the same scale (equal variances criterion).

If items in each hypothesized grouping satisfy these criteria, simple summation (or averaging) of items to derive a scale score is appropriate. If the first two scaling criteria are not satisfied in a priori hypothesized groupings, item groupings should be revised. If unhypothesized groupings are identified using factor analysis, these should be evaluated according to the other four scaling criteria. The fourth criterion is often relaxed as long as each item contributes substantially to the total. However, if more stringent criteria are being applied and the fourth scaling criterion is not satisfied, unequal weights can be used for different items. Items should be standardized prior to combining whenever their variances differ significantly (fifth scaling criterion). Equality of variances can be assessed using multiple range tests (Levy, 1975).

Multitrait scaling involves examining item frequencies, means, standard deviations, item-scale correlations (corrected for overlap), scale

internal-consistency reliability estimates, and correlations among scales. Multitrait scaling goes beyond traditional tests of internal consistency primarily because it tests item discrimination across scales, as shown in step 2 of the 5-step scaling process. Thus, items are evaluated with respect to how well they represent a particular construct relative to other constructs.

All computations were performed using the Multitrait Analysis Program (MAP), which was derived from ANLITH (Analysis of Item-Test Homogeneity program), written by Thomas Gronek at IBM and Thomas Tyler at the Academic Computing Center at Southern Illinois University. The ANLITH program was first modified for use with SAS on the IBM at The RAND Corporation by William H. Rogers, Patti Camp, and John E. Ware, Jr. The MAP program represents a modification of ANLITH for use with SAS on the IBM PC and the VAX-780 at The RAND Corporation (Hays and Hayashi, 1990; Hays, Hayashi, Carson, and Ware, 1988).

Prior to the multitrait scaling analyses, it is necessary to examine item variability. We look for comparable item variances, roughly symmetrical item response distributions, and a standard deviation near 1.0 (for 5-point scales). If all of these conditions are met, then items can be combined into scales without weighting. There are exceptions to these criteria, however. In cases where an item reflects an uncommon but serious state, it might be preferable to retain a poorly distributed item in order to represent fully a range of health states. Because most of the items in the MOS measures are based on items from prior measures or from measures developed during pilot studies, the tendency was not to eliminate final items during scale construction because of poor variability. Poor item variability was a basis for eliminating items during pilot study analyses.

When items differed substantially in the number of response options, the items were standardized prior to combining in order to weight them equally. If the number of options was the same or differed only by one, the items were combined without standardizing unless their variances differed dramatically.

Item-scale correlations are the fundamental elements of multitrait scaling (Ware, Snyder, Wright, et al., 1983). The first two steps in multitrait scaling analysis involve examining a matrix in which items are rows and scales are columns. Each row contains correlations between one item and all scales, including the one it is hypothesized to be part of. Each column contains correlations between one scale and all

items in the analysis, including those hypothesized to be part of that scale and those hypothesized to be part of other scales. Correlations between items and the scale they are a part of are corrected for overlap (Howard and Forehand, 1962) so that estimates of the item-scale correlation are not spuriously inflated.

Item convergence is supported if an item correlates substantially (a corrected correlation of 0.30 or above was used as our standard) with the scale it is hypothesized to represent. Any item not meeting this criterion is eliminated from that scale. For MOS scales that had a previous history of development and for which analyses were intended to refine rather than develop anew, a more stringent criterion of 0.40 was used.

Satisfaction of the second multitrait scaling criterion is obtained if the correlation between the item and the scale it is hypothesized to measure is significantly higher than the correlation of that item with any other scale. This test of item discrimination (Campbell and Fiske, 1959; Jackson, 1970; Thorndike, 1967a) is satisfied and a scaling "success" counted whenever the correlation between an item and its hypothesized scale is substantially higher than other correlations in the same row. A "definite" scaling success is defined by a correlation between an item and its hypothesized scale that is more than two standard errors larger than another correlation in the same row. When a correlation between an item and its hypothesized scale is significantly lower than another correlation in the same row, a "definite" scaling error is counted. When the correlation between the item and other scales in the same row is within two standard errors of its correlation with its hypothesized scale, a "probable" scaling error is counted.

Items that consistently accounted for definite scaling errors were excluded from the scale in question. Inclusion or exclusion of items associated with probable scaling errors depends on several factors, including the number of subjects in the analysis, the number of items in the scale, the internal-consistency reliability, and the strength of associations between the constructs involved. If measures are being constructed that are known to be related theoretically (e.g., depression and anxiety), probable errors associated with items in the two scale groupings are more likely to be tolerated, at least early in the process of scale development. When scale development is in more advanced stages of refinement, like the MOS measures, we were less tolerant of probable errors.

Item-scale correlations uncorrected for overlapping items are

calculated using the following formula (Tyler and Fiske, 1968):

$$r_1 = \frac{n * X_J X_W - (\Sigma X_J)(\Sigma X_W)}{n^2 S_W S_J}$$

Where J = item
W = scale with J in it.

Corrected item-scale correlations are calculated using the following formula (Howard and Forehand, 1962):

$$r_2 = \frac{r_1 S_W - S_J}{[S_W^2 + S_J^2 - 2(r_1)(S_W)(S_J)]^{1/2}}.$$

Missing Data in Multitrait Scaling When summated ratings scales are scored, substitutions might need to be made for missing item responses. Four options are possible: (1) midpoint of the possible scale range, (2) sample central tendency statistics (mean, median, or modal score for the item in question), (3) regression estimate, and (4) respondent central tendency statistic (mean, median, or modal score for that respondent across either all items in the battery or other items in the same scale). When the range of response values differs for items used (e.g., one item with four possible responses and another with five), responses can be prorated to estimate the missing response.

In the multitrait analyses, the MAP program estimates missing values using option (4) above with the mean as the central tendency statistic. The MOS estimated missing values for respondents with data for at least half of the items in a scale. Because of this, sample sizes in each multitrait analysis varied. The MAP program provides two additional missing value options: (1) respondents can be excluded if they have missing data on any item in any item grouping, and (2) scores can be estimated if respondents answer at least one item for all scales.

Factor Analysis Exploratory factor analysis was sometimes used to test for unhypothesized item groupings, especially in the pilot studies. In exploratory factor analysis, the factors identified represent underlying dimensions that define the measured items. Factor analysis was also used to guide in the development of overall indexes (i.e., combining items across constructs). For this purpose, the size of the first factor was evaluated in terms of the variance accounted for and items were identified that correlated at least 0.30 with the first unrotated principal component or factor.

Both principal components analysis and common factor analysis

were performed, depending on the situation. Factors were extracted from a matrix of product-moment correlations among item scores with unities (principal components) or communality estimates (principal axes) in the matrix diagonal. When unities are used as communality estimates, all of the variance and covariance among items is explained by the components. In contrast, when communalities are inserted in the matrix diagonal, only the covariance among items and the portion of the total variance due to common factors are explained. In principal components analysis, no distinction is made among common, specific, and error variance. Common factor analysis provides a solution based on common variance among the items, excluding unique variance (Ford, MacCallum, and Tait, 1986).

In order to achieve simple structure, factor rotation was conducted. The decision regarding the number of factors to rotate is a central issue in exploratory factor analysis (Rummel, 1970). The initial unrotated solution was evaluated using various criteria for determining the number of factors to rotate:

(1) Guttman's (1954) weakest lower bound, in which the number of factors to rotate is indicated by the number of eigenvalues exceeding 1.0 when unities are inserted in the matrix diagonal (this is the most commonly used criterion);
(2) Cattell's (1966) scree test, which involves interpreting the eigenvalue plot across factors and identifying the point at which the negative slope of the curve levels off and begins the "scree";
(3) parallel analysis (Humphreys and Ilgen, 1969), in which actual data eigenvalues derived using unities (Allen, 1986; Holden, Longman, Cota, and Fekken, 1989) or squared multiple correlation (Montanelli and Humphreys, 1976) communality estimates are compared with random data eigenvalues derived from a correlation matrix produced from normally distributed random numbers. Parallel analysis sets an upper bound for the number of factors to rotate based on the number of actual data eigenvalues that exceed the corresponding random data eigenvalues (Hays, 1985, 1987; Montanelli and Humphreys, 1976; Silverstein, 1987); and
(4) use of trial rotations when the decision as to the best number of factors for final rotation is ambiguous according to the preceding criteria. Trial rotations are evaluated in terms of interpretability and the meaningfulness and desirability of alterations in major factors when additional factors are rotated.

Both orthogonal and oblique rotations were performed. Oblique rotations are generally preferred because "it is more sensible to rotate the factors obliquely and then determine the tenability of the orthogonality assumption" (Ford, MacCallum, and Tait, 1986). Oblique rotations generally produce a more realistic representation of the factors (Rummel, 1970).

Guttman Scalogram Analysis Scalogram Analysis (Guttman, 1944) or Guttman scaling is sometimes used to construct multi-item measures. Guttman scales were avoided for the final MOS measures for several reasons. Guttman scales require a dichotomous response format, limiting responses to two choices (e.g., yes or no), which considerably restricts the amount of information that can be obtained with any one item. Guttman scales also tend to be limited to a few items, because as the number of items exceeds five or six, it is hard to achieve a realistic ordering of difficulty. Thus the "richness" of a construct is hard to represent adequately and precision is limited. When Guttman scale levels are slightly ambiguous and variability in a certain item is skewed, a different ordering of items might be obtained in different samples of people. Guttman scaling ignores measurement error and is deterministic in the sense that all individuals are assumed to adhere to the same basic response model. The experience in the HIE with Guttman scales was that scoring people who do not fall into one of the perfect scale type categories is cumbersome and time-consuming.

Calculation of Scale Scores Scale scores were calculated by averaging item scores for all respondents that had nonmissing data for half or more of the items in the scale. Missing values were assigned for the scale to those who had missing data for more than half the items in the scale. Because of the distribution of missing values for items in the scales, the number of missing values for scales was similar to that obtained if we had used a more stringent criterion of 90% nonmissing items before allowing scale scores. Most respondents either answered half or more of the scale items or answered none of the items. Most scales were transformed so that the lowest possible score was 0 and the highest possible score 100, using the following formula:

$$100 \times \frac{\text{(observed score} - \text{minimum possible score)}}{\text{(maximum possible score} - \text{minimum possible score)}}$$

For many of the single items, however, the raw scale scores were retained.

Variability of Scales After multitrait scaling studies were completed, the variability of resulting score distributions was studied. Good variability means that the scores are spread over the full range of the measure and that the distributions are roughly normal. The distributions of each final measure, including single-item measures, were evaluated in a number of ways: by seeing whether the full range of scores is observed and, if not, what the range is; by analyzing the percentage who receive a perfect score (or the percentage with some limitations); and by looking at the shape of the frequency distribution. The shape of distributions was inspected to identify measures yielding nonnormal (i.e., skewed or kurtotic) score distributions. Skewed distributions are those that are bunched up at one end of the distribution. Positive skew means that the tail is to the right and the hump is to the left. Negative skew means that the tail is to the left and the hump is to the right. The skewness statistic ranges from negative to positive infinity. The closer the skewness statistic is to zero, the more normal the distribution. The highest skewness observed in the MOS was 3.3 for the mobility measure. Kurtosis refers to whether the distribution is more peaked (leptokurtic) or more flat (platykurtic) than normal. The kurtosis statistic ranges from -2 to positive infinity. A high positive statistic indicates a more peaked distribution, a high negative statistic a flatter one. Thus, the closer the score is to zero, the more normal the distribution.

Both skewness and kurtosis are interpreted in terms of whether they are statistically different from zero, using a *t*-test. This test uses a standard error based on the sample size. The standard error for skewness is the square root of $6/N$, and the standard error for kurtosis is the square root of $24/N$. In the large MOS sample, these standard errors are small and a significant statistic is easy to obtain. Thus, more weight was given to the more qualitative approaches evaluation of the score distributions.

Variability of scores on health status measures might be insufficient for many reasons. Assuming there is actual variability in the construct being measured, insufficient variability might indicate that the items do not adequately assess the particular health construct of interest; do not adequately detect differences in some range of values between the extremes (e.g., distances between the levels represented by the items might be too large, and scores might not reflect important differences

between the health states of respondents); and do not assess important differences in health states at one or the other end of the continuum (e.g., items assessing severe limitations only). Addition of items that more precisely assess clinically significant differences between scale levels or that increase the range of measurement should increase the variability of resulting score distributions and the usefulness of the scale in detecting actual differences in health status.

Reliability and Stability Reliability refers to the consistency of the score or to the extent to which a score is free of random error. Reliability of measurement refers to the extent to which measured variance reflects true score rather than random error. To the extent a score is unreliable, it becomes more difficult to observe or to measure the true situation. A reliability coefficient is an estimate of the proportion of total variance that is true score variance, as expressed in the following formula (Kerlinger and Pedhazur, 1973): reliability $= 1 - (V_e / V_t)$, where V_e equals the error variance and V_t is the total measured variance.

For multi-item scales constructed using multitrait scaling techniques, the internal consistency of the scale is the appropriate indicator of reliability. For single-item measures and multi-item measures constructed according to other techniques, test-retest reliability must be used. When reliability estimates are unavailable, an inference about reliability can be made based on studies of correlations between the measure and other variables.

Adequate reliability is a prerequisite for using a score for any purpose (Thorndike, 1967b). The criterion for "adequate" depends on the purpose of the measure. For purposes of group comparisons, including correlational studies, reliability need not be as high as it would have to be to make individual comparisons. Reliability of 0.50 or above is considered acceptable for group comparisons (Helmstadter, 1964). Coefficients of 0.90 or greater are acceptable for individual comparisons, including evaluation of changes in an individual over time (Helmstadter, 1964). Nunnally (1978) suggests that in the early stages of research, reliabilities of 0.50 or 0.60 suffice. As the theory and methods become more refined, additional resources can be allocated to improving the reliabilities of the more important concepts (e.g., by adding more items). Even then, in basic research, the burden imposed by the number of items necessary to obtain reliabilities of 0.80 might exceed the value of the increased reliability.

Internal-Consistency Reliability The internal-consistency approach

was used to estimate reliability for all multi-item scales. This approach considers common variance (shared by all items in a scale) to be true score (reliable variance) and unique item variance to be error. The reliability coefficient it yields, coefficient alpha (Cronbach, 1951), is a function of two properties of scale items: item homogeneity or the extent to which the items covary or have something in common and the number of items in the scale. Reliability is increased when either of these properties increases. The relationships among internal-consistency reliability, homogeneity, and scale length are shown in the following formula (Nunnally, 1978): $r_{tt} = kr_{ii}/(1 + (k - 1)r_{ii}$ where r_{tt} is the internal-consistency reliability of a score, k is the number of items used to compute the scale score, and r_{ii} is the estimated reliability for a single item and can be interpreted as the average inter-item correlation (Fiske, 1966; Tyler and Fiske, 1968).

Internal consistency estimates were made using the analysis-of-variance approach to reliability (Guilford, 1954). The analysis of variance is a one-way repeated measures design with items functioning as the repeated measures. R_{tt} (alpha) is calculated by using the Hoyt (1941) formula for reliability:

$$R_{tt} = 1 - \frac{MS_{within}}{MS_{respondents}} - \frac{MS_{respondents} - MS_{within}}{MS_{respondents}}$$

Calculation formulas for R_{ii} (intra-class correlation for items, average correlation between items), R_{gg} (coefficient of homogeneity of persons), and R_{pp} (intra-class correlation for persons, average correlation between persons) are provided below:

$$R_{ii} = R_{tt}/(k + R_{tt} - (k*R_{tt}))$$
$$R_{gg} = 1 - (MS_{within}/MS_{items})$$
$$R_{pp} = R_{gg}/(N + R_{gg} - (N*R_{gg}))$$

where k is the number of items in the scale, and N is the number of respondents.

Because the reliability coefficient is a function of differences between individuals, it will tend to be larger for samples that vary more on the trait being studied (Aiken, 1982; Nunnally, 1978). Indeed, for some measures the reliability increased in this patient sample over that observed in a general population. This is consistent with the fact that

scores on these health measures tend to be more skewed in general populations than in patient samples.

For most purposes of the MOS, the aim was to achieve an internal-consistency reliability of between 0.70 and 0.80, using a minimum of from two to four items per measure. However, for the short-form measures that might be used to assess patients routinely in office practice, the MOS attempted to achieve reliabilities approaching 0.90 in order to increase their usefulness in assessing individual patients over time. Measures that approach the 0.90 standard are potentially useful for individual assessment. Measures to be used for individual assessment can be supplemented by additional items to increase their reliability or used as they are with the knowledge that standard errors of measurement are somewhat larger than ideal.

Estimating Reliability For Single-Item Measures Test-retest reliability can be used for single items as well as for multi-item scales. In test-retest reliability, questions are administered to the same group of people at two points in time, and a correlation between the two times is computed. The time difference between the two administrations must be short enough so that substantial change in the attribute being measured is unlikely but long enough so respondents will not remember their previous responses. If change has occurred the true reliability will be underestimated. If respondents remember their previous responses, the true reliability will be overestimated (Anastasi, 1976; Nachmias and Nachmias, 1981).

As the time difference between the two measures increases, the coefficient can become an indicator of stability rather than reliability. This shift is more true for constructs that tend to change over time (e.g., health) than for those that tend to be relatively stable (e.g., attitudes, beliefs) (Nunnally, 1978) and for younger versus older adults (Finn, 1986). The degree to which stability and reliability are represented in a test-retest coefficient depends on the time interval, the nature of the measure, and the characteristics of the respondents. The relative amount of reliability and stability of measures can be estimated using sophisticated panel models (Werts, Linn, and Jöreskog, 1978; Wheaton, Muthen, Alwin, and Summers, 1977; Wiley and Wiley, 1970).

The MOS was able to evaluate the correlation of some single-item measures administered on an average of four months apart. Because this time difference is substantial enough to expect a real change in health, it provides a lower-bound estimate of reliability.

In addition to test-retest reliability, we were able to glean something about the reliability of the single-item measures based on their correlations with other health measures. The reliability of a measure limits the degree of validity that is possible. This means that the correlation between one measure and another can never exceed the square root of the product of the reliabilities of these measures (Nunnally, 1978). Thus, if both measures have reliabilities of 0.70, their intercorrelation, in theory, will not be any larger than 0.70. This principle allows us to evaluate the reliability of single-item measures. For example, if a single-item measure of health correlates 0.70 with another measure of health whose reliability is about 0.70, it can be inferred that the single-item measure is adequate for group comparison. This is referred to as alternate-forms reliability. Because of the noncomparability of the alternate forms, it is considered a lower-bound estimate of reliability.

Evaluating Reliability in Disadvantaged Groups Reliability tends to be poorer in more disadvantaged groups such as those with low education (Andrews, 1984; Ware, Brook, Davies-Avery, et al., 1980). The suggested cause of this is that less educated people have difficulty reading and understanding the questions or are less familiar with questionnaire procedures. Of direct relevance to the MOS were three additional groups of patients for whom reliability might be poor: the more severely ill, the very old, and the depressed. For comparisons to be made on outcome measures across groups diverse in these characteristics, it is important to test the reliability separately to assure that minimum standards are met for each of them.

The MOS tested the internal-consistency reliability of selected health scales constructed according to multitrait scaling techniques separately for four groups: (1) those with low education (less than a high school education); (2) the very old (over age 70); (3) those with a serious chronic condition; and (4) the depressed (those having depressive symptoms using an 8-item screener for depression [Burnam, Wells, Leake, et al., 1987]).

The reliability of the measures of psychological distress and well-being (chapter 7) and of the short-form measures (Stewart, Hays, and Ware, 1988) was evaluated for one or more of these subgroups. In all cases, reliabilities were comparable. Time constraints precluded the evaluation of all measures in these subgroups; however, we expect that similar findings would be observed for other measures.

Assigning Labels to Measures The importance of the process of labeling measures is often overlooked. When a measure is used later in various clinical studies, often the detailed content of the measure is abbreviated, and readers interpret the results based on the label. Because labels are often vague (e.g., functional status, emotional functioning; subjective well-being), more attention should be given to making this process more precise. Extended labels should be assigned that clearly indicate the nature of the measure, in order to avoid confusion and misinterpretation. The MOS provides a table of definitions of each of the MOS measures (chapter 20). These extended labels include information on the time frame of the questions and summarize the item content of multi-item measures. Information is also provided on the direction of scoring in the label of each variable. The sign (+) indicates that a high score represents better health (or is more positive), and the sign (−) indicates that a high score represents poorer health. This sign facilitates interpretation of tabulated results without having to refer to scoring information.