Quality of Life Research **11**: 193–205, 2002.

Assessing health status and quality-of-life instruments: Attributes and review criteria

Scientific Advisory Committee of the Medical Outcomes Trust[1] (E-mail: klohr@rti.org)

Accepted in revised form 8 January 2002

Abstract

The field of health status and quality of life (QoL) measurement – as a formal discipline with a cohesive theoretical framework, accepted methods, and diverse applications – has been evolving for the better part of 30 years. To identify health status and QoL instruments and review them against rigorous criteria as a precursor to creating an instrument library for later dissemination, the Medical Outcomes Trust in 1994 created an independently functioning Scientific Advisory Committee (SAC). In the mid-1990s, the SAC defined a set of attributes and criteria to carry out instrument assessments; 5 years later, it updated and revised these materials to take account of the expanding theories and technologies upon which such instruments were being developed. This paper offers the SAC's current conceptualization of eight key attributes of health status and QoL instruments (i.e., conceptual and measurement model; reliability; validity; responsiveness; interpretability; respondent and administrative burden; alternate forms; and cultural and language adaptations) and the criteria by which instruments would be reviewed on each of those attributes. These are suggested guidelines for the field to consider and debate; as measurement techniques become both more familiar and more sophisticated, we expect that experts will wish to update and refine these criteria accordingly.

Key words: Health status, Item response theory, Measurement, Quality of life, Reliability, Responsiveness, Validity

Introduction

The field of health assessment

The field of health status and quality of life (QoL) measurement – as a formal discipline with a cohesive theoretical framework, accepted methods, and diverse applications – has been evolving for the better part of 30 years. It has been characterized by proliferation of instruments that vary widely in their methods of development, content, breadth of use, and quality. It has grown and matured through the efforts of both individual developers of instruments and teams of researchers, supported at crucial junctures by private philanthropic organizations and public sector (government) agencies particularly in North America, the United Kingdom, and various European countries. A further important source of support for the field has been the pharmaceutical industry on both sides of the Atlantic.

Adding to this mix of research and application have been many constructive developments: creation of various institutes and centers that develop

[1] Neil Aaronson, PhD, The Netherlands Cancer Institute, Amsterdam; Jordi Alonso, MD, Institut Municipal d'Investigacio Medica (IMIS-IMAS), Barcelona, Spain; Audrey Burnam, PhD, The RAND Corporation, Santa Monica, CA; Kathleen N. Lohr, PhD, RTI International, Research Triangle Park, NC, and Program on Health Outcomes, University of North Carolina at Chapel Hill; Donald L. Patrick, PhD, MSPH, Department of Health Services, University of Washington, Seattle; Edward Perrin, PhD, Department of Health Services, University of Washington, Seattle, Ruth E.K. Stein, MD Albert Einstein College of Medicine/Children's Hospital at Montefiore, Bronx, NY.

new instruments, translate and culturally adapt existing instruments, and facilitate research among academicians, clinicians, and health care organizations; emergence of a professional society devoted explicitly to the furtherance of this field (the international society for quality of life research [ISOQoL]); convening of numerous international colloquia and conventions on methods and issues in assessing health-related quality of life; production of numerous compilations of rating instruments and questionnaires for measuring health status, functioning, and related concepts; and publication of at least one journal whose core content relates to QoL measurement (namely, Quality of Life Research).

Today, the field is now broadly international, and its leaders are at the forefront of applying both traditional and modern theories and methods from health, psychology, and related fields to the creation and validation of such instruments. This heterogeneity induces extremely productive and useful debate and methodologic advances, and for experts in the field, this diversity is acceptable and, indeed, welcome.

The Medical Outcomes Trust and its Scientific Advisory Committee (SAC)

This complex mix of nonprofit organizations, academic researchers, public sector agencies, and commercial firms has, of course, pursued no single set of objectives. In 1992, however, the Medical Outcomes Trust was incorporated with the mission of promoting the science and application of outcomes assessment, with a particular emphasis on expanding the availability and use of self- or interviewer-administered questionnaires designed to assess health and the outcomes of health care from the patients' point of view. To accomplish this mission, the trust undertook to identify such instruments, bring them into an instrument library, and disseminate them (together with appropriate users' guides and related materials) to all persons with an interest in and need for them.

In furtherance of this task, in 1994 the trust created a Scientific Advisory Committee (SAC) – an independently operating entity charged with reviewing instruments and assessing their suitability for broad distribution by the trust. The SAC determined that, to discharge its responsibilities, it would need to establish some principles and criteria, as well as procedures, by which it would acquire, review, and make assessments about instruments that came to its attention or were submitted to the trust.

Instrument review criteria

The SAC thus set about to define a set of attributes and criteria to carry out instrument assessments and, after external peer review and revisions, published and disseminated the first set of its 'instrument review criteria' in 1996 (Evaluating quality-of-life and health status assessment instruments: Development of scientific review criteria. Clin Ther 1996; 18(5): 979–992). They were re-published in the Monitor, a trust publication, in March 1997; a subsidiary set of criteria relating solely to the evaluation of translations and cultural adaptations of instruments was published in the Bulletin, a sister trust publication, in July 1997.

Within the SAC approach and criteria, the term instrument refers to the constellation of items contained in questionnaires and interview schedules along with their instructions to respondents, procedures for administration, scoring, interpretation of results, and other materials found in a users' manual. We use the term attributes to indicate categories of properties or characteristics of instruments that warrant separate, independent consideration in evaluation. Within the attributes, we specify what we denote as criteria, which are commonly understood to be conditions or facts used as a standard by which something can be judged or considered. We view the criteria as prescribing the specific information instrument developers should be prepared to provide about particular aspects of each attribute.

In general, we have used these criteria to review instruments developed in English and cultural and language adaptations based on the English language version of the given instrument, but they can and have been used to consider instruments developed in other languages as well. We apply them to instruments that measure domains of health status and quality of life (QoL) in both groups and individuals. Although we believe that the criteria apply to the 'individualized' class of measures

(such as the schedule for the evaluation of individual quality of life [SEIQoL]) that do not have standardized items across respondents, we have not yet had experience applying these criteria with such measures.

Revised instrument review criteria

There matters stood for about 2 years, as the SAC applied its original set of instrument review criteria to instruments submitted from the United States, the United Kingdom, Canada, and various European countries as part of the larger trust activities. Increasingly, however, the SAC encountered two problems. One was that developers sometimes found the documents describing the criteria difficult to apply to their particular situation; the other was that the criteria were less applicable to instruments developed in accordance with the principles of modern test theory than to instruments created in line with classical psychometric rules.

Thus, in the course of using the initial criteria set over several years, we determined that they required revision and expansion to address advances in the science of psychometrics and to apply to a broader range of instruments. Quite apart from the fact that more instruments are being developed on principles other than classic test theory, we recognized that being able to apply the same concepts of assessment to other types of instruments, such as screeners or instruments by which consumers might rate their satisfaction with health care and plans, is also desirable.

To address these concerns, we undertook to revise the criteria following the same process as used initially. Specifically, we determined that we would retain the basic structure of the criteria set but expand the definition and specific criteria to reflect modern test theory principles and methods. We also revamped the presentation of the criteria, primarily to make clear the distinction between the description or definition of a specific attribute (e.g., reliability or respondent burden) and the specific pieces of information that we believe developers should try to provide about that attribute. Before publishing these revised criteria, we solicited outside peer review from six reviewers in the United States, the United Kingdom, and Denmark (see Acknowledgments) and revised the document accordingly.

We have three goals in mind in disseminating these criteria. First, we hope to enhance the appreciation of health outcomes assessment among as wide an audience as possible and to prompt yet more discussion and debate about continuous improvement in this field. Second, we want to provide a template by which others setting out to assess materials or systems (e.g., performance measurement or monitoring systems in the quality-of-care arena) might similarly undertake to state their evaluation criteria clearly and openly. Third, we aim to document the process and criteria used by the SAC within the context of the trust's mission.

Attributes and criteria

Eight attributes have served as the principal foci for SAC instrument review and are the core of this paper. They are:

1 Conceptual and measurement model
2. Reliability
3. Validity
4. Responsiveness
5. Interpretability
6. Respondent and administrative burden
7. Alternative forms
8. Cultural and language adaptations (translations)

Within these attributes, we established specific review criteria that are based on existing standards and evolving practices in the behavioral science and health outcomes fields. These criteria, which are general guidelines, reflect principles and practices of both classical and modern test theory. Table 1 summarizes the attributes and main criteria for each attribute. In general, our review criteria have been designed primarily for health status and QoL profiles; we acknowledge that for various utility or preference measures, yet other attributes and criteria may be appropriate. (At the end of this paper, readers will find a selected bibliography of seminal texts and articles that provide the conceptual and empirical base for these attributes and criteria; we judged this approach to be simpler than trying to document the numerous sources that could be cited for this material within the text itself.)

We review instruments in the context of 11 documented applications:

Table 1. Attributes and criteria for reviewing instruments*

Attribute	Review criteria
1. Conceptual and measurement model The rationale for and description of the concept and the populations that a measure is intended to assess and the relationship between these concepts.	– Concept to be measured – Conceptual and empirical bases for item content and combinations – Target population involvement in content derivation – Information on dimensionality and distinctiveness of scales – Evidence of scale variability – Intended level of measurement – Rationale for deriving scale scores
2. Reliability The degree to which an instrument is free from random error.	*Internal consistency* – Methods to collect reliability data – Reliability estimates and standard errors for all score elements (classical test) or standard error of the mean over the range of scale and marginal reliability of each scale (modern IRT)
Internal consistency The precision of a scale, based on the homogeneity (intercorrelations) of the scale's items at one point in time.	– Data to calculate reliability coefficients or actual calculations of reliability coefficients – Above data for each major population of interest, if necessary
Reproducibility Stability of an instrument over time (test–retest) and inter-rater agreement at one point in time.	*Reproducibility* – Methods employed to collect reproducibility data – Well-argued rationale to support the design of the study and the interval between first and subsequent administration to support the assumption that the population is stable – Information on test-retest reliability and inter-rater reliability based on intraclass correlation coefficients – Information on the comparability of the item parameter estimates and on measurement precision over repeated administrations
3. Validity The degree to which the instrument measures what it purports to measure.	– Rationale supporting the particular mix of evidence presented for the intended uses – Clear description of the methods employed to collect validity data – Composition of the sample used to examine validity (in detail) – Above data for each major population of interest
Content-related: evidence that the domain of an instrument is appropriate relative to its intended use. *Construct-related:* evidence that supports a proposed interpretation of scores based on theoretical implications associated with the constructs being measured. *Criterion-related:* evidence that shows the extent to which scores of the instrument are related to a criterion measure.	– Hypotheses tested and data relating to the tests – Clear rationale and support for the choice of criteria measures
4. Responsiveness An instrument's ability to detect change over time.	– Evidence on the changes in scores of the instrument – Longitudinal data that compare a group that is expected to change with a group that is expected to remain stable – Population(s) on which responsiveness has been tested, including the time intervals of assessment, the interventions or measures involved in evaluating change, and the populations assumed to be stable
5. Interpretability The degree to which one can assign easily understood meaning to an instrument's quantitative scores.	– Rationale for selection of external criteria of populations for purposes of comparison and interpretability of data – Information regarding the ways in which data from the instrument should be reported and displayed – Meaningful 'benchmarks' to facilitate interpretation of the scores
6. Burden The time, effort, and other demands placed on those to whom the instrument is administered (respondent burden) or on those who administer the instrument (administrative burden).	*Respondent burden* – Information on: (a) average and range of the time needed to complete the instrument, (b) reading and comprehension level, and (c) any special requirements or requests made of respondent

Table 1. (*Continued*)

Attribute	Review criteria
	– Evidence that the instrument places no undue physical or emotional strain on the respondent – When or under what circumstances the instrument is not suitable for respondents *Administrative burden* – Information about any resources required for administration of the instrument – Average time and range of time required of a trained interviewer to administer the instrument in face-to-face interviews, by telephone, or with computer-assisted formats – Amount of training and level of education or professional expertise and experience needed by administrative staff
7. Alternatives modes of administration These include self-report, interviewer-administered, trained observer rating, computer-assisted interviewer-administered, performance-based measures.	– Evidence on reliability, validity, responsiveness, interpretability, and burden for each mode of administration – Information on the comparability of alternative modes
8. Cultural and language adaptations or translations Involves two primary steps: 1. Assessment of conceptual and linguistic equivalence 2. Evaluation of measurement properties	– Methods to achieve conceptual equivalence – Methods to achieve linguistic equivalence – Any significant differences between the original and translated versions – How inconsistencies were reconciled

* For all entries in this column, developers are expected to provide definitions, descriptions, explanations, or empirical information.

– assessing the health of general populations at a point in time,
– assessing the health of specific populations at a point in time,
– monitoring the health of general populations over time,
– monitoring the health of specific populations over time,
– evaluating the impact of broad-based or community-level interventions or policies,
– evaluating the efficacy and effectiveness of health care interventions,
– conducting economic evaluations of health interventions,
– using in quality improvement and quality assurance programs in health care delivery systems,
– screening for health conditions,
– diagnosing health conditions,
– monitoring the health status of individual patients.

An instrument that works well for one purpose or in one setting or population may not do so when applied for another purpose or in another setting or population. The relative importance of the eight attributes may differ depending on the intended uses and applications specified for the instrument. Instruments may, for instance, document the health status or attitudes of individuals at a point in time, distinguish between two or more groups, assess change over time among groups or individuals, predict future status, or some combinations of these. Hence, the weight placed on one or another set of criteria may differ according to the purposes claimed for the instrument.

In reviewing instruments, the SAC aimed to be thorough without holding instruments to unrealistically high standards. For example, we accepted some instruments even though their responsiveness to change over time (attribute 4) had not been evaluated at the time of submission. In a case such as this, we would note that the instrument had been approved for group comparisons but that no data were available regarding the instrument's responsiveness. In other cases, developers may provide support for content and construct validity but not criterion validity because true gold standards are often not available for evaluating the latter. In yet other cases, reliability may be judged sufficient

for comparing groups but not for evaluating individuals. In summary, we matched criteria to particular uses claimed for the instrument and accepted instruments for specific applications when evaluation of the instrument and its documentation supported these applications.

In the remainder of this paper, we present our definition of the attributes noted above and then give our current (i.e., now revised) review criteria. The criteria are offered in terms of our view of what instruments developers should 'do' (e.g., describe, provide, or discuss) in documenting the characteristics of their instruments, so the material appears largely in bulleted form. We emphasize here that our definitions and criteria are open to further discussion and evolution within the field of health status assessment, and we hope that experts around the world will be encouraged to engage in a dialogue about these issues in years to come.

Conceptual and measurement model

Definition

A conceptual model is a rationale for and description of the concepts and the populations that a measure is intended to assess and the relationship between those concepts. A measurement model operationalizes the conceptual model and is reflected in an instrument's scale and subscale structure and the procedures followed to create scale and subscale scores. The adequacy of the measurement model can be evaluated by examining evidence that: (1) a scale measures a single conceptual domain or construct; (2) multiple scales measure distinct domains; (3) the scale adequately represents variability in the domain; and (4) the intended level of measurement of the scale (e.g., ordinal, interval, or ratio) and its scoring procedures are justified.

Classical test theory approaches may employ, for example, principal components analysis, factor analyses, and related techniques for evaluating the empirical measurement model underlying an instrument and for examining dimensionality. Methods based on modern test theory may use approaches including confirmatory factor analysis, structural equation modeling, and methods based on item response theory (IRT).

Review criteria

Developers should:

- State what broad concept (or concepts) the instrument is trying to measure – for example, functional status, well-being, health-related quality of life, QoL, satisfaction with health care, or others. In addition, if the instrument is designed to assess multiple domains within a broad concept (e.g., multiple scales assessing several dimensions of health-related quality of life), then provide a listing of all domains or dimensions.
- Describe the conceptual and empirical basis for generating the instrument content (e.g., items) and for combining multiple items into a single scale score and/or multiple scale scores.
- State the methods and involvement of the target populations for obtaining the final content of the instrument and for ascertaining the appropriateness of the instrument's content for that population, for example by use of focus groups or pretesting in target population(s).
- Provide information on dimensionality and distinctiveness of multiple scales, because both classical and modern test approaches assume appropriate dimensionality (usually unidimensionality) of scales.
- Provide evidence that the scale has adequate variability in a range that is appropriate to its intended use – for example, information on central tendency and dispersion, skewness, ceiling and floor effects, and pattern of missing data.
- State the intended level of measurement (e.g., ordinal, interval, or ratio scales) with available supportive evidence.
- Describe the rationale and procedures for deriving scale scores from raw scores and for transformations (such as weighting and standardization); for preference-weighted measures or utility measures, provide a rationale and empirical basis for the weights.

Reliability

Definition

The principal definition of test reliability is the degree to which an instrument is free from random error. Classical approaches for examining test re-

liability include (a) internal consistency reliability, typically using Cronbach's coefficient α, and (b) reproducibility (e.g., test–retest or inter-observer (interviewer) reliability). The first approach requires one administration of the instrument; the latter requires at least two administrations.

In modern test theory applications, the degree of precision of measurement is commonly expressed in terms of error variance, standard error of measurement (SEM) (the square root of the error variance), or test information (reciprocal of the error variance). Error variance (or any other measure of precision) takes on different values at different points along the scale.

Internal consistency reliability. In the classical approach, Cronbach's coefficient α provides an estimate of reliability based on all possible split-half correlations for a multi-item scale. For instruments employing dichotomous response choices, an alternative formula, the Kuder–Richardson formula 20 (KR-20), is available. Commonly accepted minimal standards for reliability coefficients are 0.70 for group comparisons and 0.90–0.95 for individual comparisons. Reliability requirements are higher when applying instrument scores for individualized use because confidence intervals of those scores are typically computed based on the SEM. The SEM is computed as the standard deviation (SD) x $\sqrt{1\text{-reliability}}$. Reliability coefficients lower than 0.9–0.95 provide too wide (e.g., more than one to two thirds of the score distribution) intervals to be useful for monitoring individual's score.

In the IRT approach, measurement precision is generally evaluated at one or more points on the scale. The scale's precision should be characterized over the measurement range likely to be encountered in actual research. A single value, marginal reliability, can be estimated as an analog to the classical reliability coefficient. This value is most useful for tests in which measurement precision is relatively stable across the scale.

Reproducibility. A second approach to reliability can be obtained by judging the reproducibility or stability of an instrument over time (test–retest) and inter-rater agreement at one point in time. In classical applications, the stability of an instrument is often expressed as a single value, but IRT applications describe specific levels of stability for specific levels of the scale. As with internal consistency reliability, minimal standards for reproducibility coefficients are also typically considered to be 0.70 for group comparisons and 0.90–0.95 for individual measurements over time.

Test–retest reproducibility is the degree to which an instrument yields stable scores over time among respondents who are assumed not to have changed on the domains being assessed. The influence of test administration on the second administration may overestimate reliability. Conversely, variations in health, learning, reaction, or regression to the mean may yield test–retest data that underestimate reproducibility. Bias and limits-of-agreement statistics can indicate the range within which 95% of retest scores can be expected to lie. Despite these cautions, information on test–retest reproducibility data is important for the evaluation of the instrument. For instruments administered by an interviewer, test–retest reproducibility typically refers to agreement among two or more observers.

Review criteria

Internal consistency reliability and test information. Developers should:

- Describe clearly the methods employed to collect reliability data. This should include (a) methods of sample accrual and sample size; (b) characteristics of the sample (e.g., sociodemographics, clinical characteristics if drawn from a patient population, etc.); (c) the testing conditions (e.g., where and how the instrument of interest was administered); and (d) descriptive statistics for the instrument under study (e.g., means, SDs, floor and ceiling effects).
- For classical applications, report reliability estimates and SEs for all elements of an instrument, including both the total score and subscale scores, where appropriate.
- For IRT applications, provide a plot showing the SEM over the range of the scale. In addition, the marginal reliability of each scale may be reported where this information is considered useful.
- Where developers have reason to believe that reliability estimates or SEM may differ substantially for the various populations in which an instrument is to be used, present these data

for each major population of interest (e.g., different chronic disease populations, different language or cultural groups).

Reproducibility. Developers should:
- Describe clearly the methods employed to collect reproducibility data. This description should include (a) methods of sample accrual and sample size; (b) characteristics of the sample (e.g.,sociodemographics, clinical characteristics if drawn from a patient population, etc.); (c) the testing conditions (e.g., where and how the instrument of interest was administered); and (d) descriptive statistics for the instrument under study (e.g., intraclass correlation coefficient, receiver operator characteristic, the test–retest mean, limits of agreement, etc.).
- Provide test–retest reproducibility information as a complement to, not as a substitute for, internal consistency.
- Give a well-argued rationale to support the design of the study and the interval between first and subsequent administrations to support the assumption that the population is stable. This can include self-report about perceived change in health over the time interval or other measures of general and specific health or functional status. Information about test and retest scores should include the appropriate central tendency and dispersion measures of both test and retest administrations.
- In classical applications for instruments yielding interval-level data, include information on test–retest reliability (reproducibility) and inter-rater reliability based on intraclass correlation coefficients (ICC, the bias statistic or test–retest mean, or limits of agreement); for nominal or ordinal scale values, κ and weighted κ, respectively, are recommended.
- In IRT applications, also provide information on the comparability of the item parameter estimates and on measurement precision over repeated administrations.

Validity

Definition

The validity of an instrument is defined as the degree to which the instrument measures what it purports to measure. Evidence for the validity of an instrument has commonly been classified in three ways discussed just below. (We note that validation of a preference-based measure will need to employ constructs relating to preferences per se, not simply descriptive constructs, and these can differ from the criteria set out below for nonutility measures.)

1. *Content-related:* Evidence that the content domain of an instrument is appropriate relative to its intended use. Methods commonly used to obtain evidence about content-related validity include the use of lay and expert panel (clinician) judgments of the clarity, comprehensiveness, and redundancy of items and scales of an instrument. Often, the content of newly developed self-report instruments is best elicited from the population being assessed or experiencing the health condition.
2. *Construct-related:* Evidence that supports a proposed interpretation of scores based on theoretical implications associated with the constructs being measured. Common methods to obtain construct-related validity data include examining the logical relations that should exist with other measures and/or patterns of scores for groups known to differ on relevant variables. Ideally, developers should generate and test hypotheses about specific logical relationships among relevant concepts or constructs.
3. *Criterion-related:* Evidence that shows the extent to which scores of the instrument are related to a criterion measure. Criterion measures are measures of the target construct that are widely accepted as scaled, valid measures of that construct. In the area of self-reported health status assessment, criterion-related validity is rarely tested because of the absence of widely accepted criterion measures, although exceptions occur such as testing shorter versions of measures against longer versions. For testing screening instruments, criterion validity is essential to compare the screening measure against a criterion measure of the diagnosis or condition in question, using sensitivity, specificity, and receiver operating characteristic.

Review Criteria

Developers should:
- Explain the rationale that supports the particular mix of evidence presented for the intended uses.

- Provide a clear description of the methods employed to collect validity data. This should include (a) methods of sample accrual and sample size; (b) characteristics of the sample (e.g., sociodemographics, clinical characteristics if drawn from a patient population); (c) the testing conditions (i.e., where and how the instrument of interest was administered); and (d) descriptive statistics for the instrument under study (e.g., means, SDs, floor and ceiling effects).
- Describe the composition of the sample used to examine the validity of a measure in sufficient detail to make clear the populations to which the instrument applies and selective factors that might reasonably be expected to influence validity, such as gender, age, ethnicity, and language.
- When reasons exist to believe that validity will differ substantially for the various populations in which an instrument is to be used, present the above data for each major population of interest (e.g., different chronic disease populations, different language or cultural groups, different age groups). Because validity testing and use of major instruments are ongoing, we encourage developers to continue to present such data as they accumulate them.
- When presenting construct validity, provide the hypotheses tested and data relating to the tests.
- When data related to criterion validity are presented, provide a clear rationale and support for the choice of criteria measures.

Responsiveness

Definition

Sometimes referred to as sensitivity to change, responsiveness is viewed as an important part of the longitudinal construct validation process. Responsiveness refers to an instrument's ability to detect change. The criterion of responsiveness requires asking whether the measure can detect differences in outcomes, even if those differences are small. Responsiveness can be conceptualized also as the ratio of a signal (the real change over time that has occurred) to the noise (the variability in scores seen over time that is not associated with true change in status).

Assessment of responsiveness involves statistical estimation of an effect size statistic – that is, an estimate of a measure of the magnitude of change in health status (sometimes denoted the 'distance' or difference between before and after scores). No agreement or consensus exists on the preferred statistical measure. Effect size statistics translate the before-and-after changes into a standard unit of measurement; essentially they involve dividing the change score by one or another variance denominator. (Said another way, these statistics are the amount of observed change over the amount of observed variance.) In these statistics, the numerator is always a change score, but the denominator differs depending on the statistic being used (e.g., standardized response mean, responsiveness statistic, SE of the mean).

Moreover, different methods may be used to evaluate effect size. Common approaches include comparing scale scores before and after an intervention that is expected to affect the construct and comparing changes in scale scores with changes in other, related measures that are assumed to move in the same direction as the target measure.

Responsiveness, as some authors have suggested, can be construed as a 'meaningful' level of change and, accordingly, defined as the minimal change considered to be important by persons with the health condition, their significant others, or their providers. We suggest, however, that this connotation of responsiveness might better be considered an element of how the data from an instrument are interpreted. Interpretation of effects is discussed in Interpretability. This includes minimally important differences or changes. We make this distinction in part because, although responsiveness and interpretability are related concepts, one focuses on performance characteristics of the instrument at hand and the other focuses on the respondents' views about the domains being studied.

Review criteria

Developers should:

- For any claim that an instrument is responsive, provide evidence on the changes in scores found in field tests of the instrument. Apart from this information, change scores can also be expressed as effect sizes, standardized response means, SEM, or other relative or adjusted measures of distance between before and after scores. The methods and formulae used to calculate the responsiveness statistics should be explained.

– Preferably, cite longitudinal data that compare a group that is expected to change with a group that is expected to remain stable.
– Clearly identify the population(s) on which responsiveness has been tested, including the time intervals of assessment, the interventions or measures involved in evaluating change, and the populations assumed to be stable.

Interpretability

Definition
Interpretability is defined as the degree to which one can assign easily understood meaning to an instrument's quantitative scores. Interpretability of a measure is facilitated by information that translates a quantitative score or change in scores to a qualitative category or other external measure that has a more familiar meaning. Interpretability calls for explanation of the rationale for the external measure, the change scores, and the ways that those scores are to be interpreted in relation to the external measure.

Several types of information can aid in the interpretation of scores:
– comparative data on the distributions of scores derived from a variety of defined population groups, including, when possible, a representative sample of the general population;
– results from a large pool of studies that have used the instrument in question and reported findings on it, thus bringing familiarity with the instrument that will aid interpretation;
– the relationship of scores to clinically recognized conditions, need for specific treatments, or interventions of known effectiveness;
– the relationship of scores or changes in scores to socially recognized life events (such as the impact of losing a job);
– the relationship of scores or changes in scores to subjective ratings of the minimally important changes by persons with the condition, their significant others, or their providers; and
– how well scores predict known relevant events (such as death or need for institutional care).

Review criteria
Developers should:
– Clearly describe the rationale for selection of external criteria or populations for purposes of comparison and interpretability of data. As with validity (attribute 3), this should include (a) rationale for selection of external criteria or comparison population; (b) methods of sample accrual and sample size; (b) characteristics of the sample; (c) the testing conditions; and (d) descriptive statistics for the instrument under study
– Provide information regarding the ways in which data from the instrument should be (or have been) reported and displayed in order to facilitate interpretation.
– Cite meaningful 'benchmarks' (comparative or normative data) to facilitate interpretation of the scores.

Burden

Definition
Respondent burden is defined as the time, effort, and other demands placed on those to whom the instrument is administered. Administrative burden is defined as the demands placed on those who administer the instrument.

Review criteria: respondent burden
Developers should:
– Give information on the following properties:
(1) average and range of time needed to complete the instrument on a self-administered basis or, as an interviewer-administered instrument, for all population groups for which the instrument is intended;
(2) the reading and comprehension level needed for all population groups for which the instrument is intended;
(3) any special requirements or requests that might be placed on respondents, such as the need to consult health care records or copy information about medications used; and
(4) the acceptability of the instrument, for example by indicating the level of missing data and refusal rates and the reasons for both.
– For instruments that are not, on the face of it, harmless and for those that appear to have excessive rates of missing data, provide evidence that the instrument places no undue physical or emotional strain on the respondent (for instance, that it does not include questions that a significant minority of patients finds too upsetting or confrontational).

– Indicate when or under what circumstances their instrument is not suitable for respondents.

Review criteria: administrative burden
Developers should provide information about any resources required for administration of the instrument, such as the need for special or specific computer hardware or software to administer, score, or analyze the instrument.

For interviewer-administered instruments, developers should:
– Document the average time and range of time required of a trained interviewer to administer the instrument in face-to-face interviews, by telephone, or with computer-assisted formats/applications, as appropriate;
– Indicate the amount of training and level of education or professional expertise and experience needed by administrative staff to administer, score, or otherwise use the instrument;
– Indicate the availability of scoring instructions.

Alternative modes of administration

Definition
Alternative modes of administration used for the development and application of instruments can include self-report, interviewer-administered, trained observer rating, computer-assisted self-report, computer-assisted interviewer-administered, and performance-based measures. In addition, alternative modes may include self-administered or interviewer-administered versions of the original source instrument that are to be completed by proxy respondents such as parents, spouses, providers, or other substitute respondents.

Review criteria
Developers should:
– Make available information evidence on reliability, validity, responsiveness, interpretability, and burden for each mode of administration;
– Provide information on the comparability of alternative modes; whenever possible, equating studies should be conducted so that scores from alternative modes can be made comparable to each other or to scores from an original instrument.

Cultural and language adaptations or translations

Definition
Many instruments are adapted or translated for applications across regional and national borders and populations. In the MOT and SAC context, cultural and language adaptations have referred to situations in which instruments have been fully adapted from original or source instruments for cultures or languages different from the original. Language adaptation might well be differentiated from translation. As a case in point: an instrument developed in Spanish or English may be adapted for different 'versions' (e.g., country- or region-specific dialects) of those basic languages, whereas an instrument developed in Swedish and translated into French or German would be quite a different matter. In any case, the SAC has held the view that measurement properties of each cultural and/or language adaptation ought to be judged separately for evidence of reliability, validity, responsiveness, interpretability, and burden.

The cross-cultural adaptation of an instrument involves two primary steps: (1) assessment of conceptual and linguistic equivalence, and (2) evaluation of measurement properties. Conceptual equivalence refers to equivalence in relevance and meaning of the same concepts being measured in different cultures and/or languages. Linguistic equivalence refers to equivalence of question wording and meaning in the formulation of items, response choices, and all aspects of the instrument and its applications. In all such cases, it is useful if developers provide empirical information on how items work in different cultures and languages.

Review criteria
Developers should:
– Describe methods to achieve linguistic equivalence. The commonly recommended steps are (a) at least two forward translations from the source language that yields a pooled forward translation; (b) at least one, preferably more, backward translations to the source language that results in another pooled translation; (c) a review of translated versions by lay and expert panels with revisions; and (d) field tests to provide evidence of comparability.

– Provide information about methods to achieve conceptual equivalence between or among different versions of the same instrument. For this step, assessment of content validity of the instrument in each cultural or language group to which the instrument is to be applied is commonly recommended. In a cross-cultural perspective, some items in a given instrument may well function differently in one language than in another. Thus, IRT and confirmatory factor analysis (using SEM approaches, for example) can be used to evaluate cross-cultural equivalence through examination of differential item functioning (DIF).
– Identify and explain any significant differences between the original and translated versions.
– Explain how inconsistencies were reconciled.

Conclusions

International interest and support for health status and QoL assessment in biomedical and health services research, clinical care, and even health policymaking are expanding rapidly. These developments are occurring in an environment featuring both traditional and emerging methods for measuring outcomes of health care, and they offer exciting opportunities to expand such applications and to enhance the confidence with which clinicians, investigators, and policymakers can use such instruments. Nonetheless, the field ought not to proceed in too free-wheeling a manner. Thus, we offer these definitions of attributes and criteria for judging instruments in the strong hope that the field will use them as a jumping-off place for debate and discussion about challenges that lie ahead. These challenges include continued refinement of existing measures and development of measures to cover clear gaps relating to patient populations and disease groups; improving instruments to make them more culturally appropriate and comparable across diverse populations; dealing with the differences and understanding the complementarities of instru-ment developed with different conceptual frameworks; and enhancing the ways that results of such instruments can be interpreted in ordinary terms.

Acknowledgments

We extend our deepest appreciation to Alvin Tarlov, MD, founding president of the Medical Outcomes Trust, for unstinting encouragement and support to the Scientific Advisory Committee (SAC) and its efforts to develop and promulgate rigorous criteria for reviewing health status instruments. We also thank Les Lipkind, the MOT Executive Director during the time this article was being prepared, for substantial background and coordinating support to the SAC. Maria Orlando, PhD, at The RAND Corporation, Santa Monica, California, assisted with the development of material encompassing modern test theory.

We are especially grateful to the following individuals who provided comments and suggestions on an earlier version of these criteria: Jacob Bjorner, AMI, Copenhagen, Denmark; John Brazier, PhD, University of Sheffield, Sheffield, United Kingdom; Yen-Pin Chiang, PhD, Agency for Healthcare Research and Quality, Rockville, Maryland; Pennifer Erickson, PhD, College Station, Pennsylvania; Rowan Harwood, MA, MSc, MD, MRCP(UK), London, England; Ronald Hays, PhD, University of California at Los Angeles and The RAND Corporation, Santa Monica, California; and Cathy D. Sherbourne, PhD, The RAND Corporation, Santa Monica, California. This acknowledgment of interest and assistance on the part of reviewers does not necessarily imply endorsement of the SAC final criteria.

This work (Lohr) was supported in part by the Center for Education and Research in Therapeutics at the University of North Carolina (Cooperative Agreement No. U18 HS10397) and the University of North Carolina Program on Health Outcomes.

Further reading

1. Aday LA. Designing and Conducting Health Surveys: A Comprehensive Guide. 2nd edn., San Francisco, Calif.: Jossey-Bass, 1996.
2. American Psychological Association. Standards for Educational and Psychological Testing. Washington, DC: APA, 1985.
3. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307–310.

4. Bjorner JB, Thunedborg K, Kristensen TS, Modvig J, Bech P. The Danish SF-36 health survey: Translation and preliminary validity studies. J Clin Epidemiol 1998; 51; 991–999.
5. Bjorner JB, Kreiner S, Ware JE Jr, Damsgaard MT, Bech P. Differential item functioning in the Danish translation of the SF-36. J Clin Epidemiol 1998; 51: 1189–1202.
6. Bowling A. Measuring Health: A Review of Quality of Life Measurement Scales. 2nd edn., London: Open University Press, 1997.
7. Carmines E. Reliability and Validity Assessment. Newbury Park, Calif.: Sage Publications. 1997.
8. Cronbach LJ. Essentials of Psychological Testing. 4th edn., New York: Harper and Row. 1984.
9. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16: 297–334.
10. Devellis R. Scale Development: Theory and Applications. Vol. 26. Applied Social Research Methods Series. Newbury Park, Calif.: Sage Publications, 1991.
11. Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of Item Response Theory. Newbury Park, Calif.: Sage Publications, 1991.
12. Lohr KN. Health outcomes methodology symposium. Summary and recommendations. Med. Care 2000; 38 (9) (Suppl. II):II194–II208.
13. Lord FM, Norvick MR. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.
14. McDonald RP. Test Theory: A Unified Treatment. Mahwah, NJ: Lawrence Erlbaum Associates, 1999.
15. McDowell I, Newell C. Measuring Health. A Guide to Rating Scales and Questionnaires. 2nd edn., New York: Oxford University Press, 1996.
16. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: Are available health surveys adequate? Qual Life Res 1995; 4: 293–307.
17. Nunnally JC, Bernstern IH. Psychometric Theory. 3rd edn., New York: McGraw-Hill, 1994.
18. Payne SL. The Art of Asking Questions. Princeton, NJ: Princeton University Press, 1951.
19. Patrick DL, Chiang Y-P. (eds) Health outcomes methodology symposium. Med Care 2000; 38 (9) (Suppl. II): II3–II208.
20. Reise SP, Widaman KF, Pugh RH. Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychol Bull 1993; 114: 552–566.
21. Staquet M, Hays R, Fayers P. Quality of Life Assessment in Clinical Trials: Methods and Practice. New York: Oxford University Press, 1998.
22. Streiner DL, Norman GR. Health Measurement Scales. 2nd edn., Oxford: Oxford University Press, 1995.
23. Wainer H, Dorans NJ, Flaugher R, et al. Computerized Adaptive Testing: A Primer. Mahwah, NJ: Lawrence Erlbaum Associates, 1990.
24. Wilkin D, Hallam L, Doggett MA. Measures of Need and Outcome for Primary Health Care. New York: Oxford University Press, 1992.

Author for correspondence: Kathleen N. Lohr, Ph.D.,
Chief Scientist, Health, Social, and Economic Research,
RTI International, PO Box 12194, 3040 Cornwallis Road,
Research Triangle Park, NC, 27709-2194 USA
Phone: +1-919-541-6512; +1-919-541-7384
E-mail: klohr@rti.org