PART I:
INTRODUCTION

1

Introduction

The SF-36v2® Health Survey is a multipurpose, short-form health survey with 36 questions that yields an eight-scale profile of functional health and well-being, as well as two psychometrically based physical and mental health summary measures and a preference-based health utility index. Like its predecessor, the SF-36® Health Survey, the SF-36v2® Health Survey is a generic measure of health status, as opposed to one that targets a specific age, disease, or treatment group. It has proven useful in surveys of general and specific populations, in comparing the relative burden of diseases, and in differentiating the health benefits produced by a wide range of treatments.

The main purpose of this chapter is to provide a summary of circumstances and events that led to the development of the SF-36v2® Health Survey. The evolution of this instrument is presented through a brief review of major health status studies that have employed its predecessors and have resulted in its subsequent improvements. This chapter also describes the developments in assessment technology, such as item response theory (IRT), computerized adaptive testing (CAT), and QualityMetric Incorporated's item banks, that have enabled better empirical demonstrations of improvements in the original SF-36® Health Survey. Finally, a new conceptual framework for health status assessment is presented. This model involves the development of disease-specific measures of the impact of illness that are standardized across measures in both content and scoring, allowing for comparison with the specific impact of other diseases.

Context for Health Status Assessment

During the 1980s, one of the more important developments in the healthcare field was the recognition of the centrality of the patient's point of view in monitoring the quality of medical care outcomes (Geigle & Jones, 1990). A *medical outcome* has come to mean the extent to which a change in a patient's behavioral functioning or well-being meets the patient's needs or expectations. This sentiment was well expressed in medical literature during the 20th century (Codman, 1914; Lembcke, 1952, as cited in Silver, 1990). More than 50 years ago, Lembcke (1952) wrote:

> The best measure of quality is not how well or how frequently a medical service is given, but how closely the result approaches the fundamental objectives of prolonging life, relieving distress, restoring function and preventing disability.

In the 1980s, these objectives were echoed by those arguing that the goal of medical care for most patients is the achievement of a more effective life (McDermott, 1981) and the preservation of function and well-being (American College of Physicians, 1988; Cluff, 1981; Ellwood, 1988; Schroeder, 1987; Tarlov, 1983). Although the patient is the best source of information regarding the achievement of these goals, information from patients about their experiences of disease and treatment had not previously been routinely collected in clinical research or medical practice. Because this information was not a part of the medical record, it was unavailable for routine analysis in the current healthcare database.

In the 1990s, clinical investigators evaluating new treatments and technologies and practicing physicians and other providers trying to achieve the best possible patient outcomes began to use information about functional status, well-being, and other important health outcomes. Policy analysts also began to utilize this information to compare the costs and benefits of competing ways of organizing and financing healthcare services, as did managers of healthcare organizations seeking to produce the best value for each healthcare dollar. Today, the primary source of new information

on general health outcomes is rapidly becoming the standardized patient surveys that have been serving researchers effectively for the past several decades.

Several advances in the methods for assessing patient perspectives about functional status, well-being, and other important healthcare outcomes occurred during the 1980s and 1990s. These advances have been the subjects of numerous conferences (Department of Health and Human Services Agency for Health Care Policy and Research, 1999; Katz, 1987; Lohr, 1989, 1992; Lohr & Ware, 1987; Patrick & Chiang, 2000; Reeve, 2004; Wenger, Mattson, Furberg, & Elinson, 1984). Some significant advances were (a) an improved understanding of the major dimensions of health and the validity of specific measurement scales in relation to those dimensions (Hays & Stewart, 1990; Liang, 1986; Ware, Brook, Davies, & Lohr, 1981), (b) demonstration of the usefulness of standardized health surveys in clinical trials (Bombardier et al., 1986; Croog et al., 1986; Fowler et al., 1988), (c) evaluations of health policy (Brook et al., 1983; Ware et al., 1986; Ware, Bayliss, et al., 1996), and (d) development of general population health surveys (Bergner, Bobbitt, Carter, & Gilson, 1981; McHorney, Kosinski, & Ware, 1994; Stewart, Hays, & Ware, 1988; Stewart et al., 1989; Ware et al., 1986).

These advances were followed by (a) the use of self-assessed well-being in medical practice (Nelson & Berwick, 1987), (b) the formation of professional societies such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the International Society for Quality of Life Research (ISOQOL), (c) the introduction of item response theory (IRT) to the field of health status measurement (Avlund, Kreiner, & Schultz-Larsen, 1993; Bech et al., 1981; Granger, Hamilton, Linacre, Heinemann, & Wright, 1993; McHorney, Haley, & Ware, 1997), and (d) the introduction of computerized adaptive testing (CAT; Bjorner & Ware, 1998; Revicki & Cella, 1997; Ware, Bjorner, et al, 2000; Ware et al., 2003).

Improvement of Health Status Surveys

The use of standardized surveys to assess functional status and well-being can be traced back over 300 years. Methodological interest, however, has been greatest during the last half of the 20th century (Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963). Most health measures used prior to the 1970s were not based upon methods of scale construction, even though psychometric techniques of scale construction had been available for most of the past century (Guttman, 1944; Likert, 1932; Thurstone & Chave, 1929). In the last 50 years, however, psychometric techniques have been used successfully in constructing numerous health status scales (Berki & Ashcraft, 1979; DiCocco & Apple, 1958; Dupuy, 1984; Ware, 1976a; Williams & Lindem, 1976).

Both the techniques for constructing health measures and the content of the measures have changed over time. The primary focus of measures was previously limited to the presence or absence of negative health status, functional limitations, symptoms of disease, and acute and chronic problems. Some health measures still focus exclusively on such negative content (Kaplan, 1989). During the last half of the 20th century, however, the content of published measures of functioning and well-being has undergone well-documented changes (Maruish, 2004a, 2004b, 2004c; McDowell & Newell, 1987; McHorney, 1997; Ware, Davies-Avery, & Brook, 1978; Ware, Johnston, Davies-Avery, & Brook, 1979; Stewart & Ware, 1992; Ware, 1987, 1995).

In recent years, more sophisticated psychometric methods, specifically IRT methodology (Fischer & Molenaar, 1995; van der Linden & Hambleton, 1997) and structural equation models for categorical data (Muthen, 1984), have been applied in the analyses of health status surveys (e.g., Bjorner, Kosinski, & Ware, 2003a; Bjorner, Kosinski, & Ware, 2003b; Bjorner, Kosinski, & Ware, 2003c; Haley, McHorney, & Ware, 1994; McHorney & Cohen, 2000; Orlando, Sherbourne, & Thissen, 2000). These techniques can be used to obtain a more realistic assessment of measurement precision, to achieve better analyses of dimensionality (Bjorner, Kosinski, & Ware, 2003a; Bjorner & Ware, 1998), and to evaluate differential item functioning (i.e., whether the survey performs in the same way in different subgroups; see Bjorner, Kreiner, Ware, Damsgaard, & Bech, 1998; Groenvold, Bjorner, Klee, & Kreiner, 1995; Raczek et al., 1998). Moreover, IRT provides a rationale for selecting the most informative items for a particular person or group (Ware et al., 2003; Ware, Bjorner, & Kosinski, 2000). Such item selection is utilized in computerized adaptive testing (CAT; van der Linden & Glas, 2000; Wainer et al., 2000). Both IRT and CAT are discussed further later in this chapter.

The Evolution of Short Form Health Status Surveys

The Health Insurance Experiment (HIE)

One of the first extensive applications of psychometric theory and methods to the development and refinement of health status surveys took place during the

Health Insurance Experiment (HIE; Brook et al., 1983; Newhouse et al., 1993; Valdez et al., 1989; Ware et al., 1986). The HIE constructed scales for measuring a broad array of functional status and well-being concepts for group-level longitudinal analyses of data from children and non-aged adults. Data collection for the HIE took place between 1974 and 1981. The work was summarized in an eight-volume set of RAND Corporation technical reports and in *Medical Care* (Brook, Ware, Davies-Avery et al., 1979; Eisen, Donald, Ware, & Brook, 1980). The HIE clearly demonstrated the potential of scales constructed from self-administered surveys to be reliable and valid tools, yielding high quality data for assessing changes in health status in the general population. It also demonstrated that with vigorous follow-up, use of measures such as these could yield high completion rates. The HIE, however, left two basic questions unanswered: (a) Can methods of data collection and scale construction such as those used in the HIE work with individuals who are older and those who have more health problems, and (b) can more efficient scales be constructed? Answering these questions was the challenge for the Medical Outcomes Study.

The Medical Outcomes Study (MOS)

The Medical Outcomes Study (MOS; see Stewart & Ware, 1992; Tarlov et al., 1989; Ware et al., 1996) was a 4-year longitudinal, observational study of the variations in practice styles and of the health outcomes for chronically ill patients. The MOS began at the University of Chicago in 1981 and was continued at the RAND Corporation and Tufts-New England Medical Center, with institutional collaborators from the University of Washington and Dartmouth Medical School. Over 23,000 patients from the practices of 362 medical clinicians and 161 mental health providers in Boston, Chicago, and Los Angeles participated in the study. The MOS provided the opportunity for a large-scale test of the feasibility of self-administered patient questionnaires and generic health scales for those with chronic conditions, including the elderly. Pilot studies began in the early 1980s, with data collection taking place between 1986 and 1990 and data analyses taking place through the early 1990s.

The MOS surveys, like the HIE surveys, were based on a multidimensional model of health. The MOS surveys, however, were more comprehensive, assessing 40 health concepts. These standardized questionnaires include the items that were selected and adapted by the principal investigator of the MOS to develop the SF-36® Health Survey. The SF-36® Health Survey represents eight of the most important health concepts

included in the MOS and other widely used health surveys. The MOS surveys also included other questions measuring health concepts not addressed by the SF-36® Health Survey, including cognitive functioning, sleep, health distress, social support, family and marital functioning, sexual functioning, and physical and psycho-physiologic symptoms.

The International Quality of Life Assessment (IQOLA) Project

In 1991, The Health Institute at Tufts-New England Medical Center began an organized effort to expand worldwide the use of health status instruments. The goal of this undertaking, referred to as the International Quality of Life Assessment (IQOLA) Project, was to develop validated translations of a single health status questionnaire that could be used in multinational clinical studies and other international studies of health. The SF-36® Health Survey was selected as the measure to be translated and used in the IQOLA Project for several reasons. For example, it is a brief, comprehensive measure of generic health status that can easily be supplemented with other generic or disease-specific measures. In addition, research on preliminary translations suggested that it could be successfully translated into several languages.

During its first year, five countries (France, Germany, Italy, Sweden, and the Netherlands) participated in the IQOLA Project. Additional researchers from other countries joined the project in 1992 and 1993; by that time, 14 countries were represented. Interest in developing translations of the SF-36® Health Survey continued such that it was translated for use in more than 70 countries by 2006. The development and validation of these translated versions contributed to the improvements in item wording and response categories, leading to the development of the SF-36v2® Health Survey. The methods and results from the translations and adaptations studies of the SF-36® Health Survey that were conducted for the IQOLA Project are described in a series of articles published in a special issue of the *Journal of Clinical Epidemiology* (Gandek & Ware [Eds.], 1998b). Visit http://www.iqola.org/ for further information about the IQOLA Project and its translation methodology.

The Medicare Health Outcomes Study (HOS)

In 1997, the U.S. Congress passed the Balanced Budget Act (BBA), which, among other provisions, directed Medicare to begin focusing on the health status of its enrollees and to begin gathering data on the effectiveness of disease management strategies in this

population (Haffer et al., 2003; Stevic, Haffer, Cooper, Adams, & Michael, 2000). Toward this end, the Centers for Medicare and Medicaid Services (CMS) worked with the National Committee for Quality Assurance (NCQA) to incorporate the Medicare population into the Health Plan Employer Data and Information Set (HEDIS®), which is widely used to measure the performance of managed healthcare plans. CMS was also interested in expanding the HEDIS outcome measures to include more generic outcomes, or outcomes that relate to patients regardless of their underlying diagnoses.

Partly in response to the findings reported by Ware, Bayliss, Rogers, Kosinski, and Tarlov (1996), an NCQA technical expert panel determined that the SF-36® Health Survey should be used as the core measure for the Medicare Health Outcomes Survey (HOS), the annual assessment of the physical and mental health of Medicare beneficiaries enrolled in managed care plans (NCQA, 2004). From 1998 to 2004, the primary outcomes in the HOS were the Physical Component Summary (PCS) and Mental Component Summary (MCS) measures, scored from the SF-36® Health Survey (using 1998 SF-36® Health Survey U.S. general population norms), and mortality. The HOS survey instrument also includes questions to obtain information about limitations in activities of daily living (ADLs) and data for use in case-mix and risk adjustment.

1998 National Survey of Functional Health Status (NSFHS)

Key to the development of the SF-36v2® Health Survey was the 1998 National Survey of Functional Health Status (NSFHS). U.S. general population norms were derived from responses to both the SF-36v2® Health Survey and the original SF-36® Health Survey forms. Households were drawn from the sampling frames maintained by National Family Opinion (NFO) Research. Panel households were balanced demographically according to the four census regions and the nine census divisions, and in the correct proportion by state within each of the nine divisions. The NFO used a two-stage area probability sample design. In the first stage, quota sampling was used based on age, sex, and income. The primary sampling units (PSUs) used were Standard Metropolitan Statistical Areas, or non-metropolitan counties stratified by region, market size, age, income, and household size before selection. The units of selection at the second stage were households stratified according to age, sex, and race.

The National Research Corporation (NRC) collected data for 12 weeks between October and December 1998 using a single wave of questionnaires mailed to randomly selected members of the NFO panel. At the end of the data collection period, the overall response rate for the survey was 67.8%. A total of 7,069 respondents completed the standard (4-week) form and 7,837 completed the acute (1-week) form. Sampling weights were applied to adjust the sample to match the age, gender, and age-by-gender distribution of the 1998 census. Norms were developed separately for the standard (N = 7,069) and acute (N = 7,837) forms. The Missing Data Estimator (MDE) from QualityMetric Health Outcomes™ Scoring Software (Saris-Baglama et al., 2004; see Chapter 5) was employed to maximize the amount of useable data. All health domain scales and component summary measures from both sets of published norm-based scoring (NBS) norms have a mean of 50 and standard deviation of 10. Norms for the SF-6D, a health state utility index derived from the SF-36® Health Survey (Brazier, Usherwood, Harper, & Thomas, 1998; see Chapter 2) were also developed based on a scale ranging from 0.0 (worst health state) to 1.0 (best health state). Because health status scores for some domains differ significantly across age groups and for men and women, norms were developed for the total population (by both combined and separate age groups) and separately for males and females (also by both combined and separate age groups).

As part of the data gathering effort, participants were asked to indicate whether they were suffering from one or more of 18 diseases or physically impairing conditions. This information enabled the development of specific sets of norms for each of the conditions and disease states, which can provide important comparison information when interpreting SF-36v2® Health Survey results from individual patients or groups of patients (see Chapter 7).

Improvements in Standards for Measurement Evaluation

Over the past few decades, several technological and psychometric advances have led to improvements in the way in which health status and quality of life can be measured. These advances have not only increased the efficiency for gathering health-related data, but have also led to improvements in measurement precision itself. The following are innovations that are particularly notable.

New Standards for Health Status Measurement: The SF-36® Health Surveys

The development of psychometrically sound measures of physical and mental health status has been

guided by standards that have served the needs of the healthcare researchers and clinical communities for a number of decades. A brief overview of some well-accepted sets of these standards is presented in Chapter 13. The realities of late 20th century healthcare delivery and research, however, created a context which necessitated some redefinition or flexibility of traditional standards of measurement in order to meet the demands of the context in which healthcare measurement currently takes place.

Adoption of new standards became necessary for two reasons. First, the old standards addressed the wrong questions for the MOS approach. Traditionally, longer measures are generally found to be more reliable and more valid (Manning, Newhouse, & Ware, 1982). The best tests, however, are those most clearly approximating the intended use of the measure (Kerlinger, 1973; McHorney, Ware, & Raczek, 1993; Ware, 1990a). The new direction in the assessment of health outcomes called for new standards formulated to address two questions: (a) What concepts should be measured, and (b) how much measurement precision is enough for each concept and for a particular purpose?

The second reason for adopting new standards was that considerations of respondent burden and the cost of data collection were prompting rethinking of measurement goals and, accordingly, the criteria used to construct and evaluate a standardized health survey. It is no longer adequate for a battery of health measures to excel in relation to traditional psychometric standards of reliability, validity, and precision. Today, psychometric measures must be sensitive to the demands (i.e., burden) they place on both the respondent and the examiner in terms of time and cost. They must demonstrate an adequate range of measurement to avoid floor and ceiling effects while maintaining acceptable validity and reliability across the range of possible scores. Measures of health status must also be understandable to respondents and other stakeholders in patients' care. Moreover, they must be translatable and acceptable across a wide range of languages and cultural groups. Consequently, opportunities to measure health status routinely demand the best compromise between traditionally defined psychometric rigor and the new standard of feasibility and practicality. The SF-36® Health Survey was developed with both of these considerations in mind.

SF-36® Health Survey

The SF-36® Health Survey was first made available in "developmental" form in 1988 (Ware, 1988) and in the standard form (i.e., SF-36® Health Survey) in 1990 (Ware et al., 1993). It was constructed to satisfy minimum psychometric standards necessary for group comparisons. The eight health domains represented in the SF-36® Health Survey profile were selected from the 40 domains that were included in the MOS (Stewart & Ware, 1992). Those chosen represent the health domains most frequently measured in widely used health surveys and those believed to be most affected by disease and health conditions (Ware, 1995; Ware et al., 1993). The items also represent multiple operational indicators of health, including behavioral function and dysfunction, distress and well-being, objective reports and subjective ratings, and both favorable and unfavorable self-evaluations of general health status (Ware et al., 1993).

Lengthier research tools served as a point of departure in the development of the SF-36® Health Survey. It was more practical than other available health status measures because it is shorter; consequently, it requires less in terms of respondent time and the cost of collecting and processing data. Also, for the great majority of respondents, the SF-36® Health Survey can be self-administered. The reliance on self-administration as the primary mode of data collection, even for surveys with more than 250 questions, was based in part on the successful use of relatively lengthy self-administered questionnaires in the MOS (Stewart & Ware, 1992). Self-administered surveys were adopted for use in the MOS on the strength of pilot studies in which self-administration worked well with chronically ill and elderly patients.

For the SF-36® Health Survey, a new standard of evaluation was established. The MOS team evaluated its scales in terms of their relative performance as judged by formal tests using external criteria, such as their validity in discriminating among diagnostic groups known to differ in morbidity and in predicting subsequent utilization of healthcare resources. Others have published the results of such tests and have expanded their efforts to include tests of sensitivity to change over time (Katz, Larson, Phillips, Fossel, & Liang, 1992).

SF-36v2® Health Survey

Although the SF-36® Health Survey proved to be useful for many purposes, 10 years of experience revealed the need and potential for improvements. A need to improve item wording and response choices resulting from the IQOLA Project and the translation of the SF-36® Health Survey form, as well as an opportunity to update normative data, led to a revision of the survey. In the early 1990's, studies were initiated to address problems with the meaning of words in some items

and to address well-documented shortcomings of the two role functioning scales (Ware & Kosinski, personal communication, September, 1996). The result of these efforts was the development of the SF-36v2® Health Survey.

Like its predecessor, the SF-36v2® Health Survey is a multi-purpose, 36-item health survey yielding a profile of two health component summary measures and eight health domain scales. It can be used across all adult patient and nonpatient populations for a variety of purposes, such as screening individual patients, monitoring the results of care, comparing the relative burden of diseases, and comparing the benefits of different treatments (Afdhal, 2003; Afdhal et al., 2004; Baird, Sanders, Woolfenden, & Bearhart, 2004; Bertagnoli & Kumar, 2002; Camilleri-Brennan, Munro, & Steele, 2003; Camilleri-Brennan & Steele, 2001, 2002; Carter, 2002; Chang et al., 2006; Chong et al., 2003; Cicero et al., 2004; Drescher, Monga, Williams, Promecene-Cook, & Schneider, 2003; Ellis & Reddy, 2002; Fernandez-Fairen, Sala, Ramirez, & Gil, 2007; Fitzgibbons et al., 2006; Han, Lee, Iwaya, Kataoka, & Kohzuki, 2004; Hawn et al., 2006; Jenkinson & Stewart-Brown, 1999; Jenkinson, Stewart-Brown, Petersen, & Paice 1999; Kelly, Brillante, Kushner, Robey, & Collins, 2005; Lanman & Hopkins, 2004; Linder & Singer, 2003; Lloyd, 1999; Martin et al., 2005; McCune et al., 2006; McManus, Mitchison, Chung, Stubbings, & Martin, 2003; Morfeld, Bullinger, Nantke, & Brahler, 2005; Morrison, Thomson, & Petticrew, 2004; Motallebzadeh, Bland, Markus, Kaski, & Jahangiri, 2006; Nicholson, Ross, Sasaki, & Weil, 2006; Perry et al., 2003; Poole & Mason, 2005; Razvi, Ingoe, McMillan, & Weaver, 2005; Reissman et al., 2004; Ricci et al., 2004; Wang, Taylor, Pearl, & Chang, 2004; Ware, Kosinski, & Bjorner, 2004; Wrennick, Schneider, & Monga, 2005; Wyrwich, Fihn, Tierney, et al., 2003; Wyrwich et al., 2006; Wyrwich, Nelson, Tierney, et al., 2003; Wyrwich, Spertus, Kroenke, et al., 2004). Relative to the SF-36® Health Survey, however, the SF-36v2® Health Survey has also incorporated (a) improved instructions and minimized ambiguity and bias in item wording, (b) improved layout of questions and answers, (c) increased comparability in relation to translations and cultural adaptations, (d) five-level response choices in place of dichotomous choices for the seven items in the Role-Physical and Role-Emotional scales, and (e) elimination of a response option from the items of the Mental Health and Vitality scales. These improvements were implemented after thorough evaluation of their advantages. The SF-36v2® Health Survey—sometimes

referred to as the "international version"— was made available for use by the research and clinical communities in 1996 (Ware & Kosinski, 1996). It represents an improved measurement tool that maintains comparability with the original version in terms of purpose, content, scores, and the psychometric rigor with which it was developed.

Without increasing the number of questions, the SF-36v2® Health Survey improvements substantially increase the reliability and validity of scores and make the survey easier to understand and complete. Further, the norm-based scoring (NBS) algorithms make it possible to compare results across both versions of the SF-36® Health Surveys, eliminating concerns about loss of comparability. Additionally, the NBS linear transformations do not change the interpretation of significance of difference in group-level comparisons. Using NBS, the health domain scales and component summary measures all have a mean of 50 and a standard deviation of 10, based on the results from a large sample of the U.S. general population in 1998 (see above; see also Chapter 14).

Studies of diverse populations in both the United States and abroad provide clear evidence that the advantages of SF-36v2® Health Survey are substantial (Jenkinson, Stewart-Brown, Petersen, & Paice, 1999). Its domains were shown to have improved reliability over the previous version of the United Kingdom SF-36® Health Survey. Furthermore, enhancements to wording and response categories reduced the extent of floor and ceiling effects in the role performance health domain scales (see Chapter 13). These advances are likely to lead to better precision as well as greater responsiveness in longitudinal studies.

Although standardized comprehensive measures of generic functional status and well-being existed prior to the SF-36® Health Survey (e.g., the Sickness Impact Profile [Bergner et al., 1981]), no instrument had received widespread adoption, nor had any one measure been shown to be suitable for use across diverse populations and healthcare settings. As a result, the opportunity to describe differences in functioning and well-being for both the sick and the well was lost. Little was known about how patients suffering from various chronic medical or psychiatric conditions differed from each other in terms of functional status and well-being. The SF-36v2® Health Survey maintains comparability with the SF-36® Health Survey and, like its predecessor, provides a common metric to compare those patients with chronic health problems to those sampled from the general population.

The SF-36v2® Health Survey is now the key member of a "family" of static (fixed-length) short-form

measures—the SF-8™ Health Survey, SF-12® Health Survey, SF-12v2® Health Survey, SF-36® Health Survey, and SF-36v2® Health Survey—that can each be scored by norm-based methods, thus enabling rough comparability between the scores obtained on different forms. Thus, using 1998 norms, results of the physical and mental health summary measures can be compared across the five forms. Further, scores on the eight health domain scales can be compared across most of the forms (see Chapter 3). Estimates of scores differ across forms primarily in terms of their precision at the individual level and their proneness to floor and ceiling effects. The SF-36v2® Health Survey can be completed and scored (using both NBS and the original 0-100 scoring systems) on the QualityMetric Incorporated Web site (http://www.QualityMetric.com/) or by purchasing the new QualityMetric Health Outcomes™ Scoring Software 2.0 (see Chapter 5). Moreover, as discussed below, it is the core instrument for the development of item banks, with its items serving as the basis for linking items from different assessment instruments with each other.

QualityMetric's Item Banks and Computerized Adaptive Testing (CAT) Tool

While the SF-12v2® Health Survey and SF-8™ Health Survey represent options for assessing the eight Short Form domains using fewer items, QualityMetric's item banks and computerized adaptive testing (CAT) system provide the option of assessing those domains with even higher precision and greater range coverage than the SF-36v2® Health Survey. In 2000, seven national norming studies were conducted to develop item banks for seven of the eight health domains. These studies included nearly 6,500 assessments that were completed via the Internet and another 4,500 assessments completed by telephone interview. Internet respondents were recruited from AOL's Opinion Place (see Ware, Kosinski, Dewey, & Gandek, 2001 for detailed description). Quotas were used to ensure that the final sample was approximately representative of the distribution of age and gender in the U.S. general population.

In total, seven item banks, one each for seven of the eight SF health domain scales (General Health scale not included), were developed from the seven national norming studies. Each national norming study consisted of a survey with items from one of the eight scales along with items selected from 52 published health status instruments measuring the same health concept as the scale. In total, there were 305 items surveyed in these seven national norming studies to build the item banks, ranging from 18 to 61 items per health domain. For each health domain, IRT methods were used to calibrate and score the items from the various instruments on a single, unidimensional scale.

As previously noted, an item bank for the General Health (GH) scale was not part of the seven national norming studies conducted in 2000. The data for the GH item bank came from the Medical Outcomes Study (Stewart & Ware, 1992), which fielded the entire battery of items ($N = 31$) from the General Health Rating Index (GHRI; Davies & Ware, 1981; Ware, Davies-Avery, & Donald, 1978). The baseline dataset ($N = 3,445$) was used to identify and calibrate a homogeneous set of 12 items using IRT methods.

The QualityMetric item banks serve to cross-calibrate items from the SF-36® Health Survey and SF-36v2® Health Survey with items from other established measures, thus providing a means of better understanding the breadth of their coverage across each domain and helping to identify their areas of strength and weakness in the measurement of health status. The item banks also allow CAT assessment of the eight health domains with even higher precision and less floor and ceiling problems than the SF-36v2® Health Survey. The basic notion of a CAT system is to mimic what an experienced clinician would do: direct questions at the individual's approximate level of health and functioning (Bjorner, Kosinski, & Ware, 2005; Ware, Bjorner, & Kosinski, 1999). For example, an adult who is able to "walk 50 feet" need not be asked a question about "walking 10 feet." CAT systems employ a simple form of artificial intelligence that selects questions tailored to the test-taker, scores everyone on a standard metric so that results can be compared, shortens or lengthens the test to achieve the desired precision, and displays results instantly (van der Linden & Glas, 2000; Wainer et al., 2000; Weiss, 1983; see Figure 1.1). By altering the stopping rule, it becomes possible to match the level of score precision to the specific purpose of measurement for each individual (Bjorner et al., 2005; Ware et al., 2003). For example, more precision in scoring will be needed to monitor individual progress than to assess the health status of a patient group.

QualityMetric Incorporated offers CAT assessment of the generic and disease-specific health domains using the DYNHA® Computer Adaptive Health Assessments

Figure 1.1   Logic of Computerized Adaptive Testing

Adapted from Wainer et al. (2000)
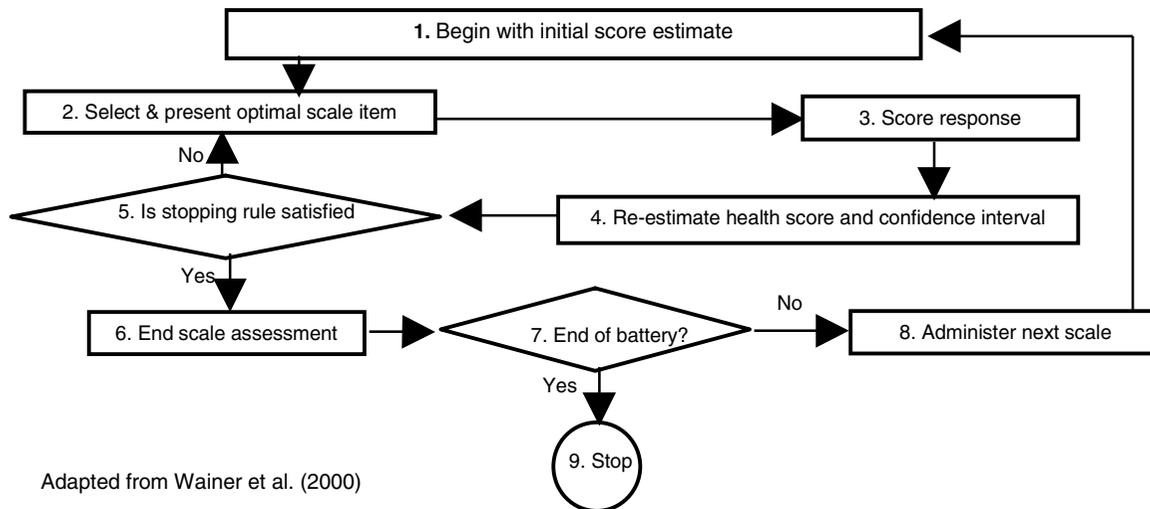
engine (Ware et al., 2003a). The DYNHA® Computer Adaptive Health Assessments engine builds on principles from item response theory and CAT logic (Fischer et al., 1995; van der Linden et al., 1997), a set of psychometric models that describes item response probabilities as a function of item characteristics and the individual's level of health-related quality of life (HRQOL).

A New Conceptual Framework for Health Status Assessment

The substantial growth in the number of tools assessing health status over the past decades has broadened the range of domains available for assessment and enabled researchers and clinicians to better understand the impact of disease from the patient's perspective (McHorney, 1997; Ware, 2003). However, it is difficult to compare results across different measurement tools. This is particularly true for *disease-, condition-,* or *procedure-specific measures,* which focus on the particulars of a specific disease or diagnostic group (e.g., diabetes, cancer), condition (e.g., congestive heart failure, low back pain), or treatment (e.g., hip or knee replacement).

In contrast to disease-specific measures, all of the Short Form family of instruments are *generic,* or *general, measures*; that is, they assesses health concepts that represent basic human values that are relevant to everyone's functional status and well-being, regardless of age, disease, or treatment group (Ware, 1987, 1990a). The term *generic* not only implies that they are univer-

sally valued but also that they are not age-, disease-, condition-, or treatment-specific.

Despite their contribution to the assessment of health status, generic health measures are not designed or intended to serve as substitutes for traditional measures of clinical endpoints. To the contrary, the greatest advances in this field during this decade are likely to come from studies that test generic health measures in parallel with clinical measures. The findings from these measures will not always be parallel; however, understanding the differences will lead to progress in this field of endeavor. The potential of such comparisons is illustrated in the profiles of functional status and well-being for patients with different medical and psychiatric conditions and in contrast to the U.S. general population (see Chapter 14). These comparisons serve at least two important purposes. First, they test the validity of SF-36v2® Health Survey health domain scales and component summary measures in describing groups of patients known to differ in functional status and well-being. Second, they facilitate understanding among clinicians of the meaning of differences in SF-36v2® Health Survey scores because of their familiarity with these diagnostic groups.

Typically, evaluation of the impact of disease on health status has been performed with both generic and disease-specific measures. In general, disease-specific measures demonstrate greater sensitivity (Bombardier et al., 1995; Kantz, Harris, Levitsky, Ware, & Davies, 1992) and specificity than generic measures (Kantz et al., 1992), while generic measures better capture the total burden of disease (Bombardier et al., 1995; Ware, 1995). In the presence of

comorbid conditions, generic measures reflect the combined effects of primary and comorbid conditions, whereas disease-specific measures reflect mainly the primary disease (Kantz et al., 1992).

A conceptual framework for constructing and describing the relationships between disease-specific and generic HRQOL measures for clinical outcomes research is presented in Figure 1.2. This framework makes important distinctions between domains of health and their operational definitions. Figure 1.2 portrays a specific-generic continuum (Ware, 1995; Ware, 2003; Wilson & Cleary, 1995) rather than a simple categorization of specific and generic concepts and measures. For example, as one moves from the left to the right on the figure, the measures change from being the most highly specific and objective clinical measures (1), to disease-specific symptoms (2), to specific measures of disease impact (3), to generic measures that are applicable across chronic disease and treatment groups (4). Measures in 3 and 4 attempt to capture specific and generic HRQOL impact, for example, with questions about limitations in role participation due to a specific disease versus questions about the same limitations without attribution to a specific disease, respectively.

Measures on the left (1 and 2) are the most specific and, therefore, useful in making a diagnosis and in determining the severity of a specific condition (Deyo & Patrick, 1989; Patrick & Deyo, 1989; Patrick & Erickson, 1988). In contrast, measures on the right (3 and 4) are more useful in understanding the impact (on functioning and well-being) of disease and treatment in the more distal HRQOL terms that matter most to patients. In compari-son with 2, measures in 3 are HRQOL measures because they capture the social and economic impact of disease and treatment. In comparison with 3, the most generic measures in 4 (e.g., Sickness Impact Profile, SF-36v2® Health Survey) are not specific to a disease or treatment and, therefore, permit meaningful comparisons across disease and treatment groups (e.g., Bergner et al., 1976; Stewart et al., 1989).

As conceptualized and measured to date, the gains in specificity achieved using disease-specific HRQOL measures (3) have been achieved at the expense of being able to make meaningful comparisons of burden across diseases and benefit across treatments using those measures. QualityMetric Incorporated has launched the *Disease Impact Project* to standardize domain content and scoring algorithms across a number of tools with disease-specific attributions (e.g., limited in social activity because of diabetes, limited in social activity because of heart failure). With such standardization, the aim is to achieve comparability even among scores from specific instruments for different diseases (see Figure 1.3).

The conceptual framework in Figure 1.2 also makes useful distinctions between the content of measures and helps to illustrate the importance of un-confounding measures across the four types. For example, when symptom frequency and/or severity is assessed and scored separately (2) and the associated specific impact is assessed and scored separately (3), the implications of different symptoms can be meaningfully studied and interpreted in terms of their impact on HRQOL in specific (3) or generic (4) terms.

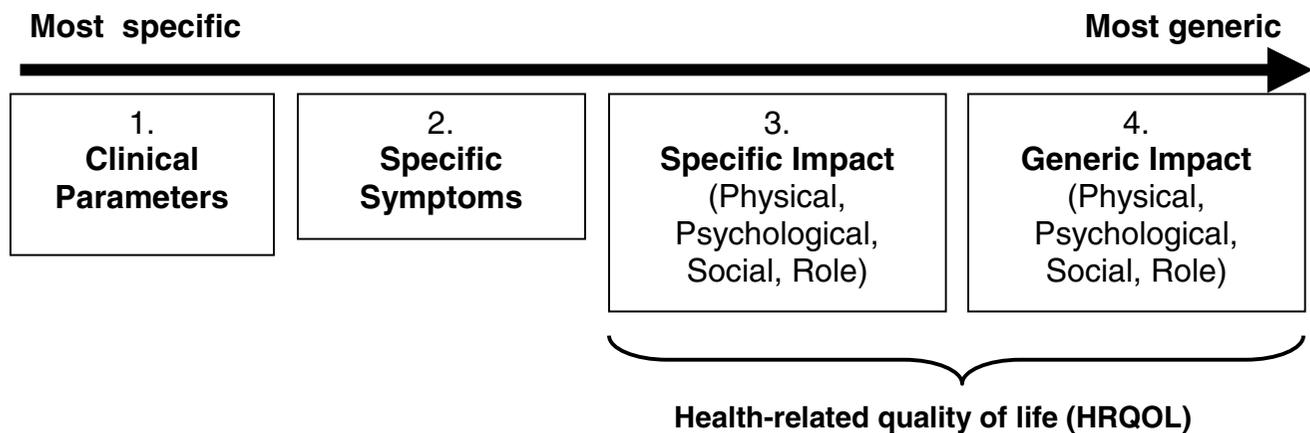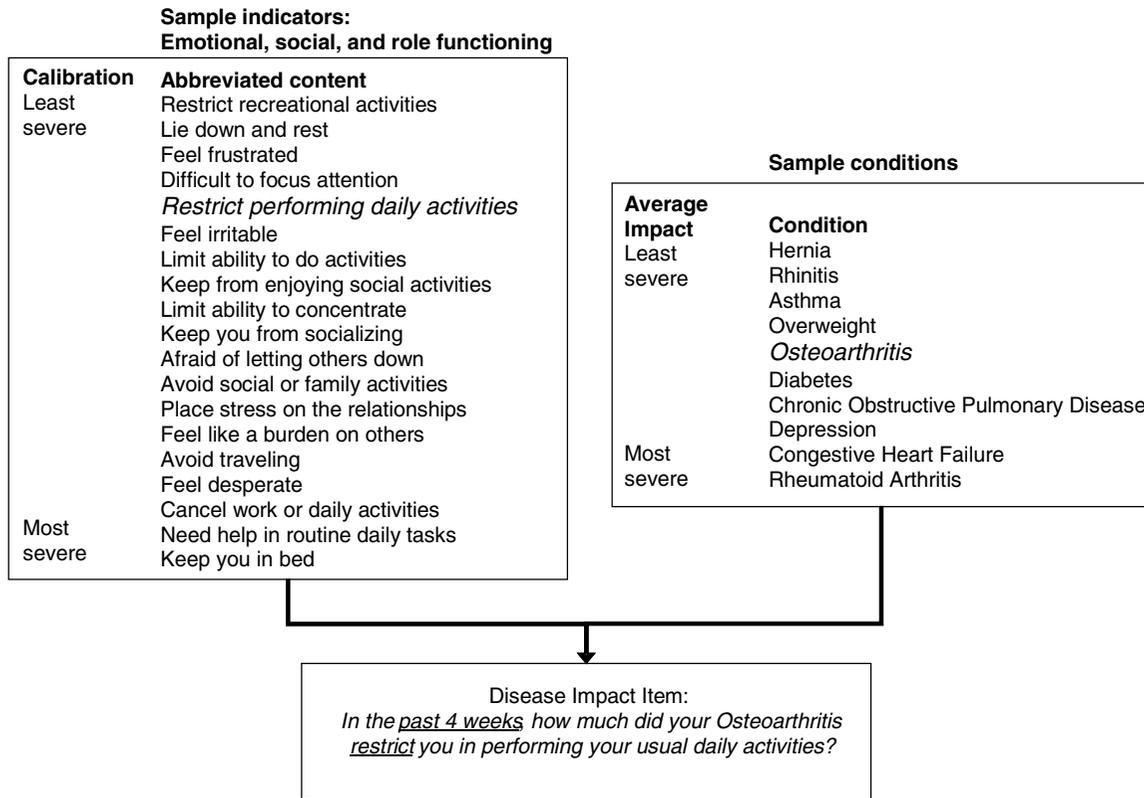**Figure 1.2** Patient-Reported Outcomes (PRO) Conceptual Framework

Figure 1.3   Components of Disease Impact Items

Sample indicators:
Emotional, social, and role functioning

Calibration	Abbreviated content
Least severe	Restrict recreational activities Lie down and rest Feel frustrated Difficult to focus attention *Restrict performing daily activities* Feel irritable Limit ability to do activities Keep from enjoying social activities Limit ability to concentrate Keep you from socializing Afraid of letting others down Avoid social or family activities Place stress on the relationships Feel like a burden on others Avoid traveling Feel desperate Cancel work or daily activities
Most severe	Need help in routine daily tasks Keep you in bed

Sample conditions

Average Impact	Condition
Least severe	Hernia Rhinitis Asthma Overweight *Osteoarthritis* Diabetes Chronic Obstructive Pulmonary Disease Depression
Most severe	Congestive Heart Failure Rheumatoid Arthritis

Disease Impact Item:
In the <u>past 4 weeks</u> how much did your Osteoarthritis <u>restrict</u> you in performing your usual daily activities?

Use of This Manual

The *User's Manual for the SF-36v2® Health Survey, Second Edition* was developed to provide those using the SF-36v2® Health Survey—clinicians, researchers, quality improvement organizations, healthcare organizations, and others—all the information necessary to evaluate and use the instrument. This edition of the manual is organized differently than any previously published manuals for the SF family of instruments, including the first edition of this manual. Information that is most useful for those who want to quickly begin using the survey is now presented in Parts II and III. This includes information that will help the user properly administer, score, and interpret the SF-36v2® Health Survey. The edges of the pages contained in Parts II and III are screened in gray for easy location. Information regarding the development, norms, and psychometric properties is presented in Part IV of this manual.

Regardless of the intended use, it is recommended that all users of the survey familiarize themselves with the content of the entire manual. A guide to finding specific information in this manual is offered in the How to Use This Manual section that immediately precedes Part I.

The *User's Manual for the SF-36v2® Health Survey, Second Edition* presents the most current information regarding the SF-36v2® Health Survey at the time of its publication. With time, this store of information will be enhanced by information from other published articles, books, and reports that will stem from efforts to further investigate the utility and psychometric integrity of the instrument. Although QualityMetric Incorporated will strive to keep users apprised of published information representing significant strides in understanding the survey and how it can be used, those employing the SF-36v2® Health Survey for any purpose are encouraged to keep abreast of the literature on the instrument as it becomes available.