J Clin Epidemiol Vol. 51, No. 11, pp. 903–912, 1998

0895-4356/98 $–see front matter
PII S0895-4356(98)00081-X

Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project

John E. Ware, Jr., and Barbara Gandek, for the IQOLA Project*
Health Assessment Lab at the Health Institute, New England Medical Center, Boston, Massachusetts

ABSTRACT. This article presents information about the development and evaluation of the SF-36 Health Survey, a 36-item generic measure of health status. It summarizes studies of reliability and validity and provides administrative and interpretation guidelines for the SF-36. A brief history of the International Quality of Life Assessment (IQOLA) Project is also included. J CLIN EPIDEMIOL 51;11:903–912, 1998. © 1998 Elsevier Science Inc.

KEY WORDS. Health status indicators, reliability, validity, SF-36 Health Survey, translations, cross-cultural research

INTRODUCTION

The SF-36 Health Survey is a multi-purpose, short-form health survey which contains 36 questions. It yields an eight-scale profile of scores as well as summary physical and mental measures. The SF-36 is a generic measure of health status as opposed to one that targets a specific age, disease, or treatment group. Accordingly, the SF-36 has proven useful in comparing general and specific populations, estimating the relative burden of different diseases, differentiating the health benefits produced by a wide range of different treatments, and screening individual patients [1]. The International Quality of Life Assessment (IQOLA) Project was established in 1991 to translate the SF-36 Health Survey and to validate, norm, and document the translations as required for their use internationally. This overview summarizes the development and evaluation of the SF-36, including studies of reliability and validity, and provides administrative and interpretation guidelines. It also presents a brief history of the IQOLA Project.

It should be noted that most of the references provided in this overview are for the U.S.-English version of the SF-36. IQOLA Project researchers replicated methods used in the United States and utilized new methods to test scaling assumptions and the reliability and validity of the SF-36 translations. Thus, much of the information provided in this article can be seen as a benchmark against which the psychometric properties of the translations can be compared.

*Address for correspondence: Barbara Gandek, MS, Health Assessment Lab, 750 Washington Street, NEMC #345, Boston, MA 02111.
Accepted for publication on 7 July 1998.

CONSTRUCTION OF THE SF-36

Much remains to be discovered about population health in terms of functional health and well-being, the relative burden of disease, and the relative benefits of alternative treatments. One reason for this has been the lack of practical measurement tools appropriate for widespread use across diverse populations. The SF-36 was constructed to provide a basis for such comparisons.

The SF-36 was constructed to satisfy minimum psychometric standards necessary for group comparisons. The eight health concepts measured in the SF-36 were selected from dozens included in the Medical Outcomes Study (MOS) [2] and represent the most frequently measured concepts in widely-used health surveys that have been shown to be affected by disease and treatment [3,4]. SF-36 items also represent multiple operational definitions of health, including function and dysfunction, distress and well-being, objective reports and subjective ratings, and both favorable and unfavorable self-evaluations of general health status [3]. The verbatim content of the SF-36 items and response choices is presented in Table 1.

Most SF-36 items have their roots in instruments that have been in use since the 1970s and 1980s [2], including the General Psychological Well-Being Inventory [5], various physical and role functioning measures [6–9], the Health Perceptions Questionnaire [10], and other measures that proved to be useful during the Health Insurance Experiment (HIE) [11]. MOS researchers selected and adapted questionnaire items from these and other sources and developed new measures for a 149-item Functioning and Well-Being Profile (FWBP) [2]. The FWBP was the source for SF-36 items and instructions, which were first made available in a "developmental" form in 1988 and in "standard" form in 1990 [12,13]. As documented elsewhere [3],

TABLE 1. Content of the SF-36 Health Survey

Label	SF-36 QUESTIONS
GH1	1. In general, would you say your health is:
HT	2. **Compared to one year ago**, how would you rate your health in general **now**?
	3. The following items are about activities you might do during a typical day. Does **your health now limit you** in these activities? Is so, how much?
PF01	a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports
PF02	b. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf
PF03	c. Lifting or carrying groceries
PF04	d. Climbing **several** flights of stairs
PF05	e. Climbing **one** flight of stairs
PF06	f. Bending, kneeling, or stooping
PF07	g. Walking **more than a mile**
PF08	h. Walking **several blocks**
PF09	i. Walking **one block**
PF10	j. Bathing or dressing yourself
	4. During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of your physical health**?
RP1	a. Cut down on the **amount of time** you spent on work or other activities
RP2	b. **Accomplished less** than you would like
RP3	c. Were limited in the **kind** of work or other activities.
RP4	d. Had **difficulty** performing the work or other activities (for example, it took extra effort)
	5. During the **past 4 weeks**, have you had any of the following problems with your work or other regular daily activities **as a result of any emotional problems** (such as feeling depressed or anxious)
RE1	a. Cut down on the **amount of time** you spent on work or other activities
RE2	b. **Accomplished less** than you would like
RE3	c. Didn't do work or other activities as **carefully** as usual
SF1	6. During the **past 4 weeks**, to what extent has your physical health or emotional problems interfered with your normal social activities with family, frends, neighbors, or groups?
BP1	7. How much **bodily** pain have you had during the **past 4 weeks**?
BP2	8. During the **past 4 weeks**, how much did **pain** interfere with your normal work (including both work outside the home and housework)?
	9. These questions are about how you feel and how things have been with you **during the past 4 weeks**. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the **past 4 weeks**—
VT1	a. Did you feel full of pep?
MH1	b. Have you been a very nervous person?
MH2	c. Have you felt so down in the dumps that nothing could cheer you up?
MH3	d. Have you felt calm and peaceful?
VT2	e. Did you have a lot of energy?
MH4	f. Have you felt downhearted and blue?
VT3	g. Did you feel worn out?
MH5	h. Have you been a happy person?
VT4	i. Did you feel tired?
SF2	10. During the **past 4 weeks**, how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting with friends, relatives, etc.)?
	11. How TRUE or FALSE is **each** of the following statements for you?
GH2	a. I seem to get sick a little easier than other people
GH3	b. I am as healthy as anybody I know
GH4	c. I expect my health to get worse
GH5	d. My health is excellent

SF-36 RESPONSE CHOICES	
1.	Excellent, Very good, Good, Fair, Poor
2.	Much better now than one year ago, Somehwat better now than one year ago, About the same as one year ago, Somewhat worse now than one year ago, Much worse now than one year ago
3.	Yes, limited a lot; Yes, limited a little; No, not limited at all
4. & 5.	Yes, No
6.	Not at all, Slightly, Moderately, Quite a bit, Extremely
7.	None, Very mild, Mild, Moderate, Severe, Very severe
8.	Not at all, A little bit, Moderately, Quite a bit, Extremely
9.	All of the time, Most of the time, A good bit of the time, Some of the time, A little of the time, None of the itme
10.	All of the time, Most of the time, Some of the time, A little of the time, None of the time
11.	Definitely true, Mostly true, Don't know, Mostly false, Definitely false

the standard form eliminated more than one-fourth of the words contained in MOS versions of the 36 items and instructions, and adopted improvements in format and scoring.

SF-36 MEASUREMENT MODEL

Figure 1 illustrates the taxonomy of items and concepts underlying the construction of the SF-36 scales and summary measures. The taxonomy has three levels: (1) items; (2) eight scales that aggregate 2–10 items each; and (3) two summary measures that aggregate scales. All but one of the 36 items (self-reported health transition) are used to score the eight SF-36 scales. Each item is used in scoring only one scale.

The eight scales are hypothesized to form two distinct higher-ordered clusters resulting from the physical and mental health variance that they have in common. Factor analytic studies have confirmed physical and mental health factors that account for 80–85% of the reliable variance in the eight scales in the U.S. general population [14], among MOS patients [14,15], and in general populations in Sweden [16] and the UK [17]. These studies have also been replicated in seven other countries [18].

The discovery that 80–85% of the reliable variance in the eight SF-36 scales was accounted for by two factors led to the construction of psychometrically-based physical and mental health summary measures. Three scales (Physical Functioning, Role-Physical, Bodily Pain) correlate most highly with the physical component and contribute most to the scoring of the Physical Component Summary (PCS) measure [14]. The mental component correlates most highly with the Mental Health, Role-Emotional, and Social Functioning scales, which also contribute most to the scoring of the Mental Component Summary (MCS) measure. Three of the scales (Vitality, General Health, and Social Functioning) have noteworthy correlations with both components. It was hypothesized that the summary measures would make it possible to reduce the number of statistical comparisons involved in analyzing the SF-36 (from eight to two) without loss of the potential for distinguishing between physical and mental health outcomes. In both cross-sectional and longitudinal studies reported to date, this objective appears to have been achieved [14,19]. The advantages and disadvantages of analyzing the eight-scale SF-36 profile versus the two summary measures are illustrated and discussed elsewhere [14,19].

The importance of these findings is illustrated in the discussion of empirical validity. Specifically, scales that load highest on the physical component are most responsive to

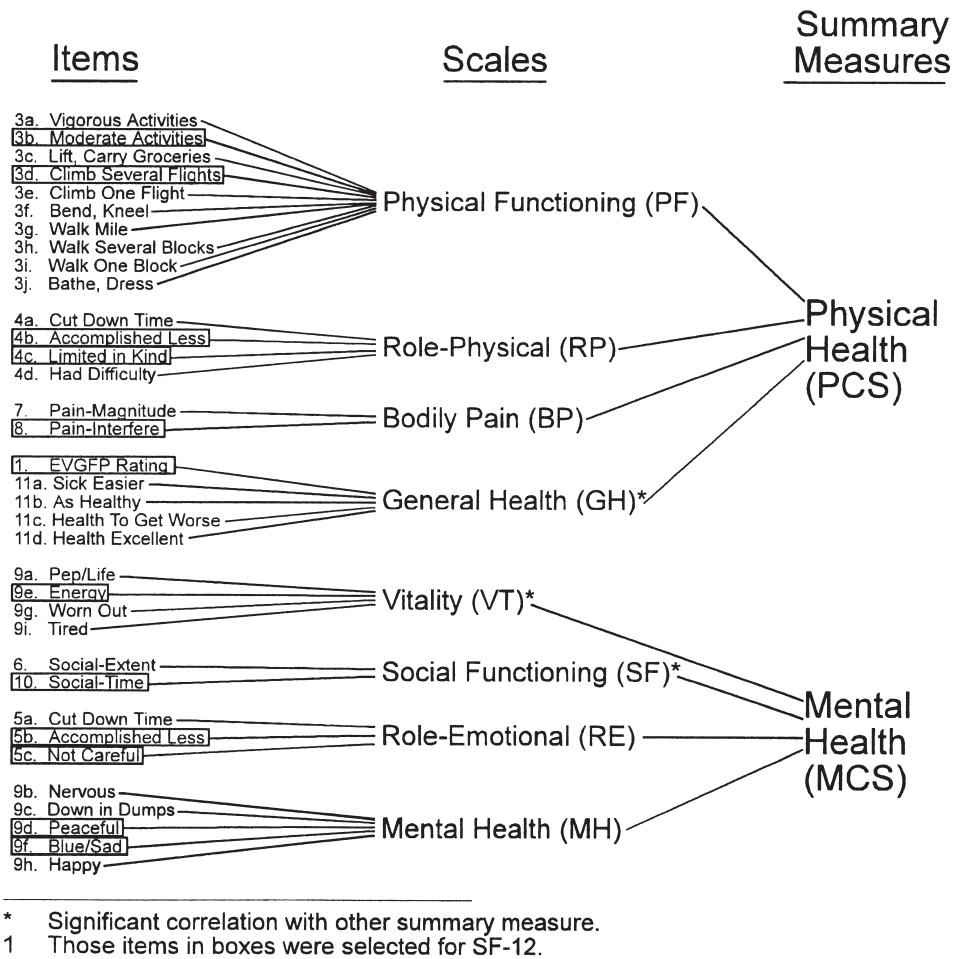

FIGURE 1. SF-36 and SF-12 measurement model (Source: [14,42]).

treatments that change physical morbidity, whereas scales loading highest on the mental component respond most to therapies that target mental health.

Finally, the SF-36 self-evaluated health transition item (five levels from "much better than one year ago" to "much worse than one year ago"), which is not used in scoring the scales or summary measures, has been shown to be useful in estimating average change in health status during the year prior to its administration. In the MOS, measured changes in health status during a 1-year follow-up period corresponded substantially, on average, to self-evaluated health transitions at the end of the year. Using the 0–100 General Health Rating Index [20] as a "criterion," those who evaluated their health as "much better" improved an average of 13.2 points. The average change was 5.8 points for those who reported that they were "somewhat better." An average decline of −10.8 was observed for those who reported that their health was "somewhat worse" and −34.4 for those reporting "much worse." (It should be noted that the latter category had only 29 patients.) Change scores over a 1-year period for those choosing the "about the same" category averaged 1.6 points. These results are encouraging with regard to the use and interpretation of the SF-36 self-evaluated transition items at the group level. Pending results from ongoing studies of the reliability of responses to this item, it should be interpreted with caution at the individual level. Additional results and their implications are discussed elsewhere [3].

TESTS OF SCALING AND SCORING ASSUMPTIONS

A major objective in constructing the SF-36 was achievement of high psychometric standards. Guidelines for testing were derived from those recommended for use in validating psychological and educational measures by the American Psychological Association, the American Education Research Association, and the National Council on Measurement in Education [21]. Extensive psychometric testing initially was conducted on the SF-36 in the United States [15,22,23] and later in numerous other countries. On the strength of favorable results from tests to date, nearly all studies have used the method of summated ratings and standardized SF-36 scoring algorithms documented elsewhere to score the eight SF-36 scales [3,24]. This method assumes that items shown in the same scale in Table 1 can be aggregated without item score standardization or weighting. Standardization of items within a scale was avoided by selecting or constructing items with roughly equivalent means and standard deviations. Weighting was avoided by using roughly equally related items (that is, items with roughly equivalent relationships to the underlying scale dimension). All items have been shown to correlate substantially (greater than 0.40, corrected for overlap) with their hypothesized scales with rare exceptions [3,23]. These results support their analysis as quasi-interval measurement scales.

RELIABILITY AND CONFIDENCE INTERVALS

The reliability of the eight scales and two summary measures has been estimated using both internal consistency and test-retest methods. With rare exceptions, published reliability statistics have exceeded the minimum standard of 0.70 recommended for measures used in group comparisons [1]; in a summary of 15 studies, most exceeded 0.80 [3]. Reliability estimates for physical and mental summary scores usually exceed 0.90 [14]. In addition, a reliability of 0.93 has been reported for the Mental Health scale using the alternate forms method, suggesting that the internal-consistency method underestimated the reliability of that scale by about 3% [25]. These trends in reliability coefficients for the SF-36 scales and summary measures have also been replicated across 24 patient groups differing in sociodemographic characteristics and diagnoses [3,14,23]. While studies of subgroups indicate slight declines in reliability for more disadvantaged respondents, reliability coefficients consistently exceeded recommended standards for group level analysis.

Standard errors of measurement, 95% confidence intervals for individual scores, and distributions of change scores from test-retest and 1-year stability studies have been published [3,14,26,27]. Confidence intervals around individual scores are much smaller for the two summary measures than for the eight scales (±6–7 points versus ±13–22 points, respectively) [14]. Estimates of sample sizes required to detect differences in average scores of various magnitudes have been documented for five different study designs for each of the eight scales and for the two summary measures [3,14].

VALIDITY AND INTERPRETATION

Studies of validity are about the meaning of scores and whether or not they have their intended interpretations. Because of the widespread use of the SF-36 across a variety of applications, evidence of all types of validity is relevant. Studies to date have addressed content, concurrent, construct, criterion, and predictive validity.

The brevity of the SF-36 was achieved by focusing on only eight of 40 health concepts studied in the MOS and by measuring each concept with a short-form scale. The content validity of the SF-36 has been compared to that of other widely used generic health surveys [3]. Systematic comparisons indicate that the SF-36 includes eight of the most frequently represented health concepts. Among the content areas included in widely-used surveys, but not included in the SF-36, are sleep adequacy, cognitive functioning, sexual functioning, health distress, family functioning, self-esteem, eating, recreation/hobbies, communication, spirituality, and symptoms/problems that are specific to one

condition. Symptoms and problems that are specific to a particular condition are not included in the SF-36 because the SF-36 is a generic measure.

To facilitate the consideration of concepts not included, the SF-36 user's manuals include tables of correlations between the eight scales and the two summary measures and 32 measures of other general concepts [3,14] and 19 specific symptoms. SF-36 scales correlate substantially ($r = 0.40$ or greater) with most of the omitted general health concepts and with the frequency and severity of many specific symptoms and problems. A noteworthy exception is sexual functioning, which correlates relatively weakly with SF-36 scales and is an example of a good candidate for inclusion in questionnaires that have the space to supplement the SF-36.

The scales chosen for the SF-36 (excluding General Health) have been shown to explain about two-thirds of the reliable variance in individual evaluations of current health status in the United Kingdom, United States, and Sweden [28]. Addition of 14 multi-item measures (e.g., sleep problems, family and sexual functioning) added only about 5% to the variance explained in general health evaluations.

Relative to other published measures, SF-36 scales have performed well in most tests published to date. The SF-36 annotated bibliography cites studies comparing the SF-36 with 225 other measures [1].

Because most SF-36 scales were constructed to reproduce longer scales, attention was initially given to how well the short-form versions perform in empirical tests relative to the full-length versions. Relative to the longer MOS measures they were constructed to reproduce, SF-36 scales have been shown to perform with about 80–90% empirical validity in studies involving physical and mental health "criteria" [22]. This disadvantage of the SF-36 should be weighed against the fact that some of these long-form measures require 5–10 times greater respondent burden. Empirical studies of this tradeoff suggest that the SF-36 provides a practical alternative to longer measures and that the eight scales and two summary scales rarely miss a noteworthy difference in physical or mental health status in group level comparisons [3,14,29]. Regardless, the fact that the SF-36 represents a documented compromise in measurement precision (relative to longer MOS measures) leading to a reduction in the statistical power of hypothesis testing should be taken into account in planning clinical trials and other studies. In relation to longer non-MOS measures, such as the Sickness Impact Profile, the SF-36 has performed equally well or better in detecting average group differences or changes over time [29,30].

The validity of each scale has been shown to differ markedly from the other scales, as would be expected from factor analytic studies of their construct validity (see Figure 2) [14,15,19]. The same has been shown for the two summary measures. Specifically, the Mental Health, Role-Emotional, and Social Functioning scales and the MCS summary measure have been shown to be the most valid mental health measures in both cross-sectional and longitudinal tests using the method of known-groups validity. The Physical Functioning, Role-Physical, and Bodily Pain scales and the PCS have been shown to be the most valid physical health measures. Criteria used in the initial known-groups validation of the SF-36, which include accepted clinical indicators of diagnosis and severity of depression, heart disease, and other conditions, are well documented in peer-reviewed publications and in the two user's manuals [3,14,15,19,31].

The Mental Health scale has been shown to be useful in screening for psychiatric disorders [14,32], as has the MCS summary measure [14]. For example, using a cutoff score of 42, the MCS had a sensitivity of 74% and a specificity of 81% in detecting patients diagnosed with depressive disorder [14].

Results from clinical studies comparing scores for patients before and after treatment have largely supported hypotheses developed from results of factor analytic studies about the validity of SF-36 scales. For example, clinical studies have shown that three of the scales (Physical Functioning, Role-Physical, and Bodily Pain) with the most physical factor content (Figure 2) tend to be most responsive to the benefits of knee replacement [33], hip replacement [29,34], and heart valve surgery [35]. In contrast, the three scales with the most mental factor content (Mental Health, Role-Emotional, and Social Functioning) in factor analytic studies have been shown to be most responsive in comparisons of patients before and after recovery from depression [19]; change in the severity of depression [36]; as well as drug treatment and interpersonal therapy for depression [37]. Experience to data from nearly 100 studies suggests that the SF-36 will be useful in evaluating the benefits of other treatments as well [1].

Predictive validity studies have linked SF-36 scales and summary measures to utilization of health care services [14], the clinical course of depression [36,38], loss of job within 1 year [14], and 5-year survival [14].

The interpretation value of general and specific population norms, which was demonstrated well for the Sickness Impact Profile [39] and later for the MOS SF-20 [40,41] and other measures, has also been demonstrated for the SF-36. General population normative data for the SF-36 has been collected in the United States [3,14,42] and as documented elsewhere in this issue, in 12 other countries [43].

Whereas some of the initial descriptive studies using the SF-36 were performed primarily to validate scale scores [22], on the strength of validation studies to date, SF-36 scales appear to be increasingly accepted as valid health measures for purposes of documenting disease burden. The advantage of a generic survey (such as the SF-36) in estimating disease burden is well illustrated in articles describing more than 130 diseases and conditions. Among the most frequently studied conditions, with more than 20 SF-36 publications each, are arthritis, back pain, depres-

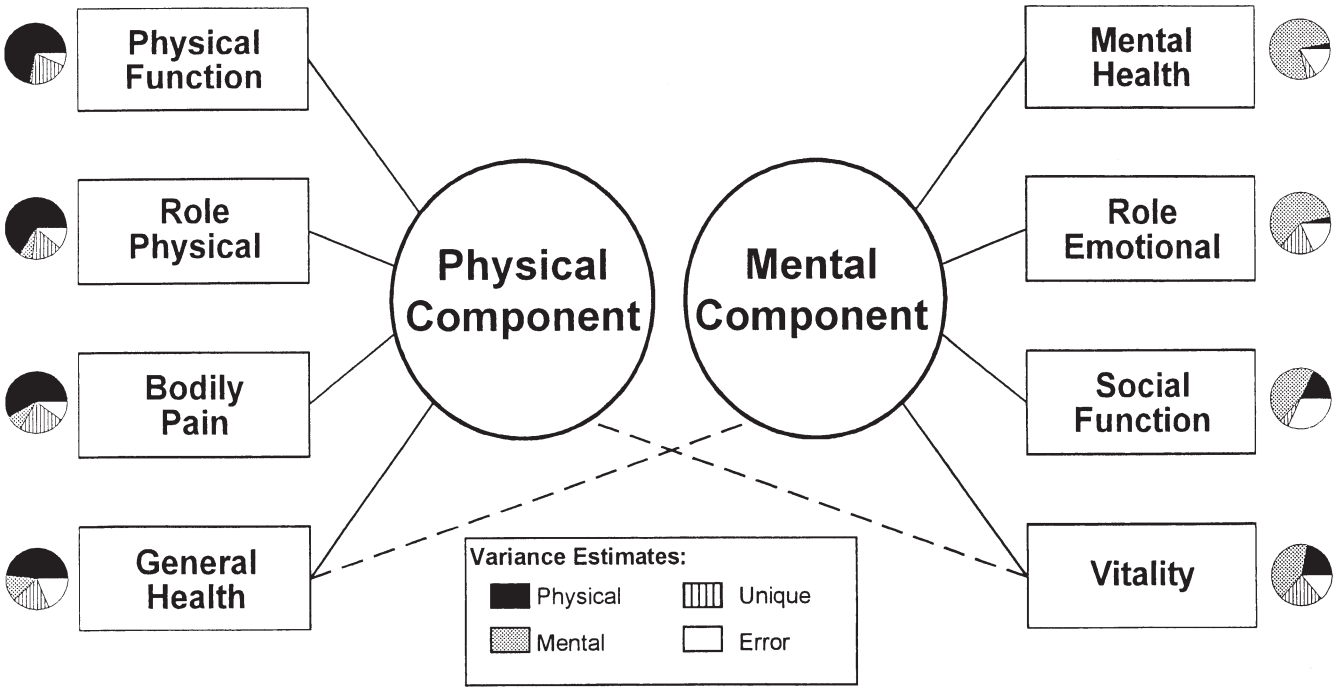

FIGURE 2. Construct validation of the SF-36 two-component model (Source: [14]).

sion, diabetes, and hypertension [1]. The standardization of measurement across studies is producing considerable information about norms and benchmarks useful in comparing "well" and "sick" populations and for estimating the burden of specific conditions.

ADMINISTRATIVE GUIDELINES

The SF-36 is suitable for self-administration, computerized administration, or administration by a trained interviewer in person or by telephone, to persons age 14 and older. The SF-36 has been administered successfully in general population surveys in the United States and other countries [43] as well as to young and old adult patients with specific diseases [3,23]. It can be administered in 5–10 minutes with a high degree of acceptability and data quality [3]. Indicators of data quality that have yielded satisfactory results in studies to date include very high item-completion rates and favorable results for a response consistency index based on 15 pairs of SF-36 items, which is scored at the individual level [3]. Computer administered and telephone voice recognition interactive systems of administration are currently being evaluated.

Although many studies appear to be relying on the SF-36 as the principal measure of health outcome, among the most useful studies are those that use it as a "generic core." A generic core battery of measures makes it possible to compare results across studies and populations and accelerates the accumulation of interpretation guidelines that are essential to determining the clinical, economic, and social relevance of differences in health status and outcomes. However, because it is short, the SF-36 can be reproduced in a questionnaire with ample room for other more precise general and specific measures to follow. Numerous studies [1,29,44,45] have adopted this strategy and have illustrated the advantages of supplementing the SF-36 with other measures. When included with other batteries, the SF-36 is usually administered first to preserve the standardization necessary for norm-based interpretation.

INTERPRETATION GUIDELINES

Table 2 summarizes information about the eight SF-36 scales and two summary measures that is important in their use and interpretation. The eight scales are ordered in terms of their factor content (i.e., construct validity) as they are in the SF-36 profile to facilitate interpretation. The first scale is Physical Functioning (PF), which has been shown to be the best all around measure of physical health; the last scale, Mental Health (MH) is the most valid measure of mental health in studies to date [3,14,15]. Interestingly, MH is the poorest measure of the physical component and PF is the poorest measure of the mental component. Scales in between are ordered according to their validity in U.S. studies in measuring physical and mental health. The Vitality and General Health scales have substantial or moderate validity for both components of health status and should be interpreted accordingly.

The number of items and levels and the range of states defined by each scale are also shown in Table 2. These attributes have been linked to their empirical validity

TABLE 2. Summary of information about SF-36 scales and physical and mental component summary measures

Scales	Correlations[a]		Number of[b]		Mean[a]	SD[a]	Reliability[a]	CI[c]	Definition (% observed)	
	PCS	MCS	Items	Levels					Lowest possible score (floor)[d]	Highest possible score (ceiling)[d]
Physical Functioning (PF)	.85	.12	10	21	84.2	23.3	.93	12.3	Very limited in performing all physical activities including bathing or dressing (0.8%)	Performs all types of physical activities including the most vigorous without limitations due to health (38.8%)
Role-Physical (RP)	.81	.27	4	5	80.9	34.0	.89	22.6	Problems with work or other daily activities as a result of physical health (10.3%)	No problems with work or other daily activities (70.9%)
Bodily Pain (BP)	.76	.28	2	11	75.2	23.7	.90	15.0	Very severe and extremely limiting pain (0.6%)	No pain or limitations due to pain (31.9%)
General Health (GH)	.69	.37	5	21	71.9	20.3	.81	17.6	Evaluates personal health as poor and believes it likely to get worse (0.0%)	Evaluates personal health as excellent (7.4%)
Vitality (VT)	.47	.65	4	21	60.9	20.9	.86	15.6	Feels tired and worn out all of the time (0.5%)	Feels full of pep and energy all of the time (1.5%)
Social Functioning (SF)	.42	.67	2	9	83.3	22.7	.68	25.7	Extreme and frequent interference with normal social activities due to physical and emotional problems (0.6%)	Performs normal social activities without interference due to physical or emotional problems (52.3%)
Role-Emotional (RE)	.16	.78	3	4	81.3	33.0	.82	28.0	Problems with work or other daily activities as a result of emotional problems (9.6%)	No problems with work or other daily activities (71.0%)
Mental Health (MH)	.17	.87	5	26	74.7	18.1	.84	14.0	Feelings of nervousness and depression all of the time (0.0%)	Feels peaceful, happy, and calm all of the time (0.2%)
Physical Component Summary (PCS)			35	567[b]	50.0	10.0	.92	5.7	Limitations in self-care, physical, social, and role activities, severe bodily pain, frequent tiredness, health rated "poor" (0.0%)	No physical limitations, disabilities, or decrements in well-being, high energy level, health rated "excellent" (0.0%)
Mental Component Summary (MCS)			35	493[b]	50.0	10.0	.88	6.3	Frequent psychological distress, social and role disability due to emotional problems, health rated "poor" (0.0%)	Frequent positive affect, absence of psychological distress and limitations in usual social/role activities due to emotional problems, health rated "excellent" (0.0%)

[a]U.S. data.
[b]Number of levels observed at baseline; scores rounded to the first decimal place (n = 2,474).
[c]CI = 95% confidence interval.
[d]Percentage observed comes from general U.S. population sample
Source: [14].

[3,14,15]. The most precise (least coarse) scales are those with 20 or more levels (PF, GH, VT, and MH) and have the smallest standard deviations. They also define the widest range of health states and, therefore, usually produce the least skewed score distributions. The relatively coarse role disability scales (RP and RE) each measure only four or five levels across a restricted range and, therefore, usually have the most problems with ceiling and floor effects and the largest standard deviations.

Means and standard deviations for each of the eight scales in the general U.S. adult population are also presented. These can be used to determine whether a group or individual in question scores above or below the U.S. average. Detailed normative data including frequency distributions of scores and percentile ranks are documented in the two user's manuals [3,14].

Table 2 illustrates the practical implications of a number of theoretical advantages of the PCS and MCS summary measures including reliability, as well as the number and range of levels covered.

SF-36 LITERATURE

The experience through 1996 with the SF-36 has been documented in more than 400 publications, which have been summarized in annotated bibliographies [1,46]; an additional 300 publications were published in 1997 and will be documented in the 1997 update to the bibliography. The most complete information about the history and development of the SF-36, its psychometric evaluation, studies of reliability and validity, and normative data is available in the first of three user's manuals [3]. A second manual documents the development and validation of the SF-36 summary measures and presents norms for those measures [14]. A third presents similar information for the SF-12 Health Survey, an even shorter version constructed from a subset of 12 items [42]. One of the most complete independent accounts of SF-36 development along with a critical commentary is offered by McDowell and Newell [47]. Additional publications are listed on the Internet at the SF-36 web page (http://www.sf-36.com) and a third edition of the annotated bibliography is in preparation.

HISTORY OF THE IQOLA PROJECT

The International Quality of Life Assessment (IQOLA) Project began in 1991, with the goal of developing validated translations of a health status questionnaire for use in multinational clinical trials and other international studies of health [48,49]. Although the SF-36 Health Survey has become an increasingly popular measure since 1991, at that time the SF-36 was only beginning to be widely used. Thus, much consideration was given to the questionnaire that was to be translated in the IQOLA Project. The SF-36 was first made available in "developmental" form in 1988 and in "standard" form in 1990. By 1991, a number of studies had documented the acceptability, reliability, and validity of SF-36 scales [32,50–53], and research had indicated that it was applicable across heterogeneous populations in the United States [23]. Research in several countries using preliminary SF-36 translations and a translated "parent" full-length questionnaire [16,54–56] suggested that the SF-36 could be translated successfully. In addition, the SF-36 was a comprehensive measure of generic health status, and because it was brief it could be supplemented with other generic and disease-specific measures in clinical studies. Thus, the SF-36 was chosen as the health status measure to be translated in the IQOLA Project.

The first meeting of the IQOLA Project took place in September 1991 in Paris. Participants included the IQOLA Project Principal Investigator and other U.S.-based Health Assessment Lab research staff, National Principal Investigators from the first five sponsored countries (France, Germany, Italy, the Netherlands, and Sweden), and members of the France-based Mapi Research Institute, who assisted in the coordination of the project in its early stages. During this meeting, basic project policies and procedures were discussed. The project adopted a policy of granting royalty-free permission for other academic researchers to use the IQOLA Project translations while they were in development and testing, upon signature of a user's agreement. Since 1991, more than 1000 researchers have been granted permission to use IQOLA Project translations, while the translations were being tested prior to their publication. Researchers also agreed in the meeting that the IQOLA Project translations would be made available royalty-free to all users upon publication, through a non-profit organization.

Additional sponsored researchers joined the project in 1992 and 1993. Research began in five additional sponsored countries in 1992 (Australia, Belgium, Canada, Japan, and Spain), and a representative from the United Kingdom also joined the project in that year. Sponsored investigators from Denmark and Norway joined the project in 1993. IQOLA research procedures were refined and augmented in subsequent meetings, which included representatives from all 14 countries.

Since 1993, interest in the SF-36 has increased worldwide. As of June 1998, researchers were translating and studying the SF-36 in more than 40 other countries, including: Argentina, Austria, Bangladesh, Brazil, Bulgaria, Cambodia, Chile, China, Colombia, Croatia, Czech Republic, Estonia, Finland, Greece, Honduras, Hong Kong, Hungary, Iceland, India, Indonesia, Iran, Ireland, Israel, Korea, Malaysia, Mexico, New Zealand, Poland, Portugal, Romania, Russia, Singapore, Slovak Republic, South Africa, Switzerland, Taiwan, Tanzania, Thailand, Turkey, Ukraine, the United Kingdom (Welsh), the United States (Chinese, Japanese, Vietnamese), and Yugoslavia.

To promote the standardization of the administration, scoring, and interpretation of the SF-36 translations,

TABLE 3. IQOLA Project methodology: Steps and products

Steps	Products
I. Translation	Survey form
II. Scale construction	Scoring algorithms
III. Validations and norming	Interpretation

IQOLA researchers have written scoring documentation and user's manuals, which as of June 1998 were available for Australia, Canada, Denmark, Germany, Italy, Spain, Sweden, the United Kingdom, and the United States. Other manuals are forthcoming. In addition, more than 200 publications using the SF-36 translations and English-language adaptations have been published; for more information see the articles in this issue and [1]. For up-to-date information about SF-36 user's manuals and other IQOLA Project publications, please go to http://www.sf-36.com or http://www.iqola.org.

The IQOLA Project followed a three-stage research protocol for translating and testing the SF-36, including translation following a standard process; formal psychometric tests of the assumptions underlying item scoring and construction of multi-item scales; and studies to evaluate validity and the equivalence of interpretations across countries [28,57]. As noted in Table 3, the products of these three research stages are: (1) questionnaires which can be used in data collection; (2) scoring algorithms which can be used to make standardized comparisons; and (3) validation and norming studies that provide a basis for interpretation. These research stages are discussed further in this supplement, in articles by Bullinger *et al.* [58], Wagner *et al.* [59], Keller *et al.* [60], and Ware and Gandek [43,61]. Articles which demonstrate the application of these methods in 15 countries follow.

References

1. Manocchia M, Bayliss MS, Connor J, Keller SD, Shiely J-C, Tasai C, *et al.* **SF-36 Health Survey Annotated Bibliography: Second Edition (1988–1996).** Boston, MA: The Health Assessment Lab, New England Medical Center; 1998.
2. Stewart AL, Ware JE. **Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.** Durham, NC: Duke University Press; 1992.
3. Ware JE, Snow KK, Kosinski M, Gandek B. **SF-36 Health Survey Manual and Interpretation Guide.** Boston, MA: New England Medical Center, The Health Institute; 1993.
4. Ware JE. The status of health assessment 1994. **Annu Rev Public Health** 1995; 16: 327–354.
5. Dupuy HJ. The Psychological General Well-Being (PGWB) Index. In: Wenger NK, Mattson ME, Furberg CD, Elinson, J, Eds. **Assessment of Quality of Life in Clinical Trials of Cardiovascular Disease.** New York: Le Jacq Publishing, Inc.; 1984: 170–183.
6. Patrick DL, Bush JW, Chen MM. Toward an operational definition of health. **J Health Soc Behav** 1973; 14: 6–21.
7. Hulka BS, Cassel JC. The AAFP-UNC study of the organization, utilization and assessment of primary medical care. **Am J Public Health** 1973; 63(6): 494–501.
8. Reynolds WJ, Rushing WA, Miles DL. The validation of a functional status index. **J Health Soc Behav** 1974; 15: 271–289.
9. Stewart AL, Ware JE, Brook RH. Advances in the measurement of functional status: Construction of aggregate indexes. **Med Care** 1981; 19(5): 473–488.
10. Ware JE. Scales for measuring general health perceptions. **Health Serv Res** 1976; 11(4): 396–415.
11. Brook RH, Ware JE, Davies-Avery A, Stewart AL, Donald CA, Rogers WH, *et al.* Overview of adult health status measures fielded in RAND's Health Insurance Study. **Med Care** 1979; 17(Suppl. 7): 1–131.
12. Ware JE. **How to Score the Revised MOS Short-Form Health Scale (SF-36).** Boston, MA: The Health Institute, New England Medical Center Hospitals; 1988.
13. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. **Med Care** 1992; 30(6): 473–483.
14. Ware JE, Kosinski M, Keller SD. **SF-36 Physical and Mental Health Summary Scales: A User's Manual.** Boston, MA: The Health Institute; 1994.
15. McHorney CA, Ware JE, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. **Med Care** 1993; 31(3): 247–263.
16. Sullivan M, Karlsson J, Ware JE. The Swedish SF-36 Health Survey: I. Evaluation of data quality, scaling assumptions, reliability and construct validity across general populations in Sweden. **Soc Sci Med** 1995; 41(10): 1349–1358.
17. Jenkinson C, Layte R, Lawrence K. Development and testing of the Medical Outcomes Study 36-Item Short Form Health Survey summary scale scores in the United Kingdom. **Med Care** 1997; 35: 410–416.
18. Ware JE, Kosinski M, Gandek B, Aaronson NK, Apolone G, Bech P, *et al.* The factor structure of the SF-36 Health Survey in ten countries: Results from the IQOLA Project. **J Clin Epidemiol** 1988; 51: 1159–1165.
19. Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, Raczek A. Comparison of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: Summary of results from the Medical Outcomes Study. **Med Care** 1995; 33(Suppl. 4): AS264–AS279.
20. Davies AR, Ware JE. **Measuring Health Perceptions in the Health Insurance Experiment.** Santa Monica, CA: Rand Corporation; 1981; R-2711-HHS.
21. American Psychological Association. **Standards for Educational and Psychological Tests.** Washington, DC: American Psychological Association; 1985.
22. McHorney CA, Ware JE, Rogers W, Raczek A, Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: Results from the Medical Outcomes Study. **Med Care** 1992; 30(Suppl. 5): MS253–MS265.
23. McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. **Med Care** 1994; 32(4): 40–66.
24. Medical Outcomes Trust. **How to Score the SF-36 Health Survey.** Boston, MA: Medical Outcomes Trust; 1991.
25. McHorney CA, Ware JE. Construction and validation of an alternate form general mental health scale for the Medical Outcomes Study Short-Form 36-Item Health Survey. **Med Care** 1995; 33(1): 15–28.
26. Brazier JE, Harper R, Jones NMB, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. Validating the SF-36 Health Sur-

vey Questionnaire: New outcome measure for primary care. **Br Med J** 1992; 305: 160–164.
27. Ware JE. Tech Notes: Confidence intervals for individual scores. **Medical Outcomes Trust Bulletin** 1994; 2(1): 3.
28. Ware JE, Keller SD, Gandek B, Brazier JE, Sullivan M, IQOLA Project Group. Evaluating translations of health status questionnaires: Methods from the IQOLA Project. **Int J Technol Assess Health Care** 1995; 11(3): 525–551.
29. Katz JN, Larson MG, Philips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. **Med Care** 1992; 30(10): 917–925.
30. Beaton DE, Bombardier C, Hogg-Johnson S. Choose your tool: A comparison of the psychometric properties of five generic health status instruments in workers with soft tissue injuries. **Qual Life Res** 1994; 3: 50–56.
31. Kravitz RL, Greenfield S, Rogers WH, Manning WG, Zubkoff M, Nelson E, *et al.* Differences in the mix of patients among medical specialties and systems of care: Results from the Medical Outcomes Study. **JAMA** 1992; 267(12): 1617–1623.
32. Berwick DM, Murphy JM, Goldman PA, Ware JE, Barsky AJ, Weinstein MC. Performance of a five-item mental health screening test. **Med Care** 1991; 29: 169–176.
33. Kantz ME, Harris WJ, Levitsky K, Ware JE, Davies AR. Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. **Med Care** 1992; 30(Suppl. 5): MS240–MS252.
34. Lansky D, Butler JBV, Waller FT. Using health status measures in the hospital setting: From acute care to "outcomes management." **Med Care** 1992; 30(Suppl 5): MS57–MS73.
35. Philips RC, Lansky DJ. Outcomes management in heart valve replacement surgery: Early experience. **J Heart Valve Dis** 1992; 1(1): 42–50.
36. Beusterien KM, Steinwald B, Ware JE. Usefulness of the SF-36 Health Survey in measuring health outcomes in the depressed elderly. **J Geriatr Psychiatry Neurol** 1996; 9: 13–21
37. Coulehan JL, Schulberg HC, Block MR, Madonia MJ, Rodrigues E. Treating depressed primary care patients improves their physical, mental, and social functioning. **Arch Intern Med** 1997; 157: 1113–1120.
38. Wells KB, Burnam MA, Rogers W, Hays R, Camp P. The course of depression in adult outpatients: Results from the Medical Outcomes Study. **Arch Gen Psychiatry** 1992; 49: 788–794.
39. Bergner M, Bobbitt RA, Carter WB, Gilson BS. The Sickness Impact Profile: Development and final revision of a health status measure. **Med Care** 1981; 19(8): 787–805.
40. Stewart AL, Hays RD, Ware JE. The MOS Short-Form General Health Survey: Reliability and validity in a patient population. **Med Care** 1988; 26(7): 724–735.
41. Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, *et al.* Functional status and well-being of patients with chronic conditions; Results from the Medical Outcomes Study. **JAMA** 1989; 262(7): 907–913.
42. Ware JE, Kosinski M, Keller SD. **SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. Second Edition.** Boston, MA: The Health Institute, New England Medical Center; 1995.
43. Gandek B, Ware JE. Methods for validating and norming translations of health status questionnaires: The IQOLA Project approach. **J Clin Epidemiol** 1998; 51: 953–959.
44. Wagner AK, Keller SD, Kosinski M, Baker GA, Jacoby A, Hsu M-A, *et al.* Advances in methods for assessing the impact of epilepsy and antiepileptic drug therapy on patient's health-related quality of life. **Qual Life Res** 1995; 4: 115–134.
45. Nerenz DR, Repasky DP, Whitehouse FW, Kahkonen DM. Ongoing assessment of health status in patients with diabetes mellitus. **Med Care** 1992; 30(Suppl. 5): MS112–MS124.
46. Shiely J-C, Bayliss MS, Keller SD, Tsai C, Ware JE. **SF-36 Health Survey Annotated Bibliography: The First Edition (1988–1995).** Boston, MA: The Health Institute, New England Medical Center; 1996.
47. McDowell I, Newell C. **Measuring Health: A Guide to Rating Scales and Questionnaires, 2nd Edition.** New York: Oxford University Press; 1996.
48. Aaronson NK, Acquadro C, Alonso J, Apolone G, Bucquet D, Bullinger M, *et al.* International Quality of Life Assessment (IQOLA) Project. **Qual Life Res** 1992; 1: 349–351.
49. Ware JE, Gandek B, IQOLA Project Group. The SF-36 Health Survey: Development and use in mental health research and the IQOLA Project. **Int J Ment Health** 1994; 23: 49–73.
50. Cleary PD, Greenfield S, McNeil BJ. Assessing quality of life after surgery. **Controlled Clin Trials** 1991; 12: 189S–203S.
51. Gelberg L, Linn LS. Psychological distress among homeless adults. **J Nerv Ment Dis** 1989; 177: 291–295.
52. Lancaster TR, Singer DE, Sheehan MA, Seehan MA, Oertel LB, Maraventano SW, Hughes RA, *et al.* The impact of long-term warfarin therapy on quality of life: Evidence from a randomized trial. **Arch Int Med** 1991; 151: 1944–1949.
53. Weinberger M, Samsa GP, Hanlon JT, *et al.* An evaluation of a brief health status measure in elderly women. **J Am Geriatr Soc** 1991; 39: 691–694.
54. Liang J, Wu SC, Krause NM, Chiang TL, Wu HY. The structure of the mental health inventory among Chinese in Taiwan. **Med Care** 1992; 30: 659–676.
55. Leplege A, Mesbah M, Marquis P. Preliminary psychometric analysis of the French version of an international quality of life questionnaire: The MOS SF-36 (Version 1.1). **Revue d'Epidemiologie et de Sante Publique** 1995; 43: 371–379.
56. Apolone G, Mosconi P. The Italian SF-36 Health Survey: Translation, validation and norming. **J Clin Epidemiol** 1988; 51: 1025–1036.
57. Ware JE, Keller SD, Gandek B, Brazier JE, Sullivan M. Evaluating translations of health status questionnaires: Methods from the IQOLA Project. **Int J Technol Assess Health Care** 1995; 11: 525–551.
58. Bullinger M, Alonso J, Apolone G, Leplege A, Sullivan M, Wood-Dauphinee S, *et al.* Translating health status questionnaires and evaluating their quality: The IQOLA Project approach. **J Clin Epidemiol** 1988; 51: 913–923.
59. Wagner AK, Gandek B, Aaronson NK, Acquadro C, Alonso J, Apolone G, *et al.* Cross-cultural comparisons of the content of SF-36 translations across ten countries: Results from the IQOLA Project. **J Clin Epidemiol** 1988; 51: 925–932.
60. Keller SD, Ware JE, Gandek B, Aaronson NK, Alonso J, Apolone G, *et al.* Testing the equivalence of translations of widely-used response choice labels: Results from the IQOLA Project. **J Clin Epidemiol** 1988; 51: 933–944.
61. Ware JE, Gandek B. Methods for testing data quality, scaling assumptions, and reliability: The IQOLA Project approach. **J Clin Epidemiol** 1988; 51: 945–952.