

SF-36 Health Survey Update

***John E. Ware Jr., PhD
QualityMetric Incorporated
and Tufts University Medical School***

Reprinted with permission from:

***The Use of
Psychological Testing
For Treatment Planning
and Outcomes Assessment, Third Edition
Volume 3, pp. 693-718
Mark E. Maruish, Editor
© 2004 Lawrence Erlbaum Associates***

SF-36 Health Survey Update

John E. Ware Jr.
QualityMetric Inc. and Tufts University Medical School

The SF-36 is a multipurpose, short form health survey with 36 questions. It yields an eight scale profile of functional health and well-being scores, as well as psychometrically based physical and mental health summary measures and a preference based health utility index. It is a generic measure, as opposed to one that targets a specific age, disease, or treatment group. Accordingly, the SF-36 has proven useful in surveys of general and specific populations, comparing the relative burden of diseases and in differentiating the health benefits produced by a wide range of different treatments. This chapter summarizes the steps in the construction of the SF-36; how it led to the development of an even shorter (one-page, 2 minute) survey form—the SF-12, which is based on a subset of 12 SF-36 items; the improvements reflected in Version 2.0 of the SF-36; psychometric studies of assumptions underlying scale construction and scoring, how they have been translated for use in more than 60 countries and languages as part of the International Quality of Life Assessment (IQOLA) Project, and studies of reliability and validity. The strengths and weaknesses of the SF-36 and SF-12 forms as well as the new SF-8 form, and considerations in choosing among them for various applications are discussed on the Internet (www.sf-36.org) and in various user's manuals (Ware, Kosinski, & Dewey, 2003; Ware, Kosinski, Turner-Bowker, & Gandek, 2002).

SF-36 LITERATURE

The experience to date with the SF-36 has been documented in nearly 5,000 publications, 2,060 citations for those published in 1988 through 2000 are documented in a bibliography covering the SF-36 and other instruments in the "SF" family of tools (Turner-Bowker, Bartley, & Ware, 2002). 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 SF-36 users' manuals (Ware, Snow, Kosinski, & Gandek, 1993). This information was also summarized in the first two peer-reviewed articles about the SF-36 Health Survey (McHorney, Ware, & Raczek, 1993; Ware & Sherbourne, 1992). A second manual documents the development and validation of the SF-36 physical and mental component summary measures and presents norms for those measures (Ware, Kosinski, & Dewey, 2000).

Ware, Kosinski, & Keller, 1994). These user's manuals have been revised to include more up-to-date norms and other findings and to document the much improved Version 2.0 (SF-36v2), which are discussed below (Ware et al., 2000; Ware & Kosinski, 2001; Ware et al., 2003). A fourth manual, first published in 1995 (Ware, Kosinski, & Keller, 1995) and recently updated (Ware, Kosinski, Turner-Bowker, & Gandek, 2002), presents similar information for the SF-12 Health Survey, an even shorter version constructed from a subset of 12 SF-36 items.

One of the most complete independent accounts of the development of the SF-36, along with a critical commentary, was offered by McDowell and Newell (1996). More recently, the SF-36 was judged to be the most widely evaluated generic patient assessed health outcome measure in a bibliographic study of the growth of "quality of life" measures published in the *British Medical Journal* (Garratt, Schmidt, Mackintosh, & Fitzpatrick, 2002). Additional information about the SF-36 literature and a community forum for discussing old and new publications and the interpretation of results are available on the SF-36 Web page (<http://www.sf-36.org>).

The usefulness of the SF-36 in estimating disease burden and comparing disease-specific benchmarks with general population norms is illustrated in articles describing more than 100 diseases and conditions. Among the most frequently studied diseases and conditions, with 100 or more SF-36 publications each, are: arthritis, asthma, back pain, cancer, cardiovascular disease, chronic obstructive pulmonary disease, depression, diabetes, gastro intestinal disease, migraine headache, HIV/AIDS, kidney disease, low back pain, multiple sclerosis, musculoskeletal conditions, osteoarthritis, psychiatric diagnoses, renal disease, rheumatoid arthritis, stroke, surgical procedures, trauma, vascular disease and women's health issues (Turner-Bowker et al., 2002).

Translations of the SF-36 have been the subject of nearly 800 publications involving investigators in 60 countries. Ten or more studies have been published from 13 countries.

CONSTRUCTION OF THE SF-36

The SF-36 was constructed to satisfy minimum psychometric standards necessary for group comparisons. The eight health concepts were selected from 40 included in the Medical Outcomes Study (MOS; Stewart & Ware, 1992). Those chosen represent the most frequently measured concepts in widely used health surveys and those most affected by disease and treatment (Ware, 1995; Ware et al., 1993). The questionnaire items selected 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).

Most SF-36 items have their roots in instruments that have been in use since the 1970s and 1980s (Stewart & Ware, 1992), including items from: the General Psychological Well-Being Inventory (Dupuy, 1984); various physical and role functioning measures (Fulka & Cassel, 1973; Patrick, Bush, & Chen, 1973; Reynolds, Rushing, & Miles, 1974; Stewart, Ware, & Brook, 1981); the Health Perceptions Questionnaire (Ware, 1976); and other measures that proved to be useful during the Health Insurance Experiment (HIE; Brook et al., 1979). MOS researchers selected and adapted questionnaire items from these and other sources, and developed new measures for the 149-item Functioning and Well-Being Profile (FWBP; Stewart & Ware, 1992). The FWBP was the source for questionnaire items and instructions adapted for use in the

SF-36. The SF-36 was first made available by the Health Institute at New England Medical Center in a "developmental" form in 1988 and in "standard" form in 1990 (Ware, 1988; Ware & Sherbourne, 1992). As documented elsewhere (Ware et al., 1993), the standard form eliminated more than one fourth of the words contained in MOS versions of the 36 items and also incorporated improvements in item wording, format, and scoring.

VERSION 2.0

In 1996, Version 2.0 of the SF-36 (SF-36v2) was introduced to correct deficiencies identified in the original version. Those improvements, which are documented in the SF-36v2 user's manual (Ware, Kosinski, & Dewey, 2000; Ware et al., 2003), were implemented after careful study using both qualitative and quantitative methods. Briefly, the SF-36v2 improvements include

- Improvements in instructions and questionnaire items to shorten and simplify the wording and make it more familiar and less ambiguous
- An improved layout for questions and answers in the self-administered forms that makes it easier to read and complete, and that reduces missing responses
- Greater comparability with translations and cultural adaptations widely used in the United States and in other countries
- Five-level response choices in place of dichotomous response choices for seven items in the two role-functioning scales
- Five-level (in place of six-level) response categories to simplify items in the Mental Health (MH) and Vitality (VT) scales.

These and other improvements are briefly explained later.

Layout

All responses to questions in Version 2.0 are printed in a left-to-right (also referred to as horizontal) format, rather than with the mixture of horizontal and vertical listings of response choices that were printed below questions in the MOS and in the original SF-36. Mixed formats of response choices confuse respondents and cause missing and inconsistent responses, particularly among older people. Other improvements include more consistent use of indenting, numbering of instructions, deletion of useless item labels, and a simpler formatting of boxes that are checked by respondents.

Type Size and Bolding

A larger type size has been adopted throughout. Only instructions, as opposed to response choices, are bolded to simplify the "look and feel" of Version 2.0. These and other refinements were adopted on the basis of lessons learned in health care and from surveys in other fields.

Wording Changes

Evidence from numerous focus group studies, formal cognitive tests, and from empirical studies in more than a dozen countries support the improvements in item

wording and the changes in some terms used to identify health concepts adopted in Version 2.0. These improvements make the English-language SF-36 easier to understand and administer, as well as making it more objective. Version 2.0 is also more comparable with translations of the SF-36. Because most of the improvements in item wording were developed during the process of translating and adapting the SF-36 for use in other countries during the IQOLA Project, Version 2.0 is sometimes referred to as the international version.

Five-Choice Response Scales

There is considerable empirical evidence that the Version 2.0 five-level response scales substantially improve the two SF-36 role functioning scales. Version 2.0 response scales extend the range measured and greatly increase score precision without increasing respondent burden. Specifically, Version 2.0 achieves a fivefold increase in the number of levels defined by both role scales, a substantially smaller standard deviation, and substantially reduces the percentage of respondents who score at both the ceiling and floor for both role scales. The elimination of one of the six response choices ("a good bit of the time") from the MH and VT items was based on the finding that this response choice is not consistently ordered between adjacent categories in studies of item responses in Version 1.0 or in translations of the SF-36. Eliminating this choice simplified the format of the form with little or no loss of information.

Scoring and Norms

With the release of SF-36v2, norms were updated using data from the 1996 National Survey of Functional Health Status and norm-based scoring (NBS) algorithms were introduced for all eight scales (Ware et al., 2000). NBS, which employs a linear *T*-score transformation with mean = 50 and standard deviation = 10, makes it possible to meaningfully compare scores for the eight-scale profile and the physical and mental summary measures in the same graph. SF-36v2 scoring software also yields less biased estimates of missing responses and makes it possible to estimate scores for more respondents with incomplete data (Kosinski, Bayliss, Bjorner, & Ware, 2000).

Comparability of Results

To make Version 1.0 easier to interpret and directly comparable to published results based on Version 2.0, cross-sectional and longitudinal norms for general and specific populations were reestimated for Version 1.0 using NBS for all eight scales and for the two summary measures. Further, national calibration studies were fielded in the United States in 1998 and 1999 to evaluate the effect of all improvements and to assure the comparability of average scores across Versions 1.0 and 2.0 (Ware et al., 2000).

Acute (1-Week Recall) Form

The SF-36 is now available in both standard (4-week) and acute (1-week) recall versions. The more recently developed acute form was designed for applications in which health status would be measured weekly or biweekly. It was created by changing the recall period for six of the eight scales [Role-Physical (RP), Bodily Pain (BP), VT, Social Functioning (SF), Role-Emotional (RE) and MH] from "the past four weeks" to "the past week." Two scales, Physical Functioning (PF) and General Health (GH), do not

24 SF-36 HEALTH SURVEY UPDATE

have a recall period; the items and instructions for these scales are identical across acute and standard forms

The rationale behind a form with a 1-week recall period was that it would be more sensitive to recent changes in health status. This hypothesis was tested by comparing results for both the 1-week and original 4-week recall forms administered three times during a clinical trial of treatments for asthma (Keller et al., 1997). As hypothesized, answers to SF-36 questions with a 1-week recall period tended to be more responsive to recent changes in disease state as estimated using several clinical criteria defining the severity of asthma. For example, changes in acute (1-week recall) SF-36 scale scores were generally more highly related to 1-week changes in asthma severity. Of some concern, from a normative perspective, the study also revealed higher mean scores for the acute version scales in comparison with the standard form scales. One explanation offered was a lower prevalence of negative events during the shorter recall period defined by the acute form. If so, this potential difference in mean scores would have implications for the norm-based interpretation of acute form scores. However, the findings from this one clinical trial were not replicated during the 1998 norming of the acute and standard forms in the general U.S. population (Ware et al., 2000). Accordingly, the developers recommend use of the same norms for purposes of the norm-based interpretation of both the standard and acute versions.

PSYCHOMETRIC CONSIDERATIONS

SF-36 Measurement Model

Figure 24.1 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 to 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, because of 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% to 85% of the reliable variance in the eight scales in the U.S. general population (Ware, Kosinski, & Keller, 1994), among MOS patients (McHorney et al., 1993; Ware, Kosinski, & Keller, 1994), and in general populations in Sweden (Sullivan, Karlsson, & Ware, 1995) and the United Kingdom (Ware, Kosinski, & Keller, 1994). As of 1998, these studies had been replicated in more than a dozen countries (Fukahara, Ware, Kosinski, Wada, & Gandek, 1998; Ware et al., 1998).

Three scales (PF, RP, BP) correlate most highly with the physical component and contribute most to the scoring of the Physical Component Summary (PCS) measure (Ware, Kosinski, & Keller, 1994). The mental component correlates most highly with the MH, RE, and SF scales, which also contribute most to the scoring of the Mental Component Summary (MCS) measure. Three of the scales (VT, GH, and SF) have noteworthy correlations with both components.

The importance of these findings is illustrated below in the discussion of empirical validity. Specifically, scales that load highest on the physical component are most responsive to treatments that change physical morbidity, whereas scales loading highest on the mental component respond most to drugs and therapies that target mental health.

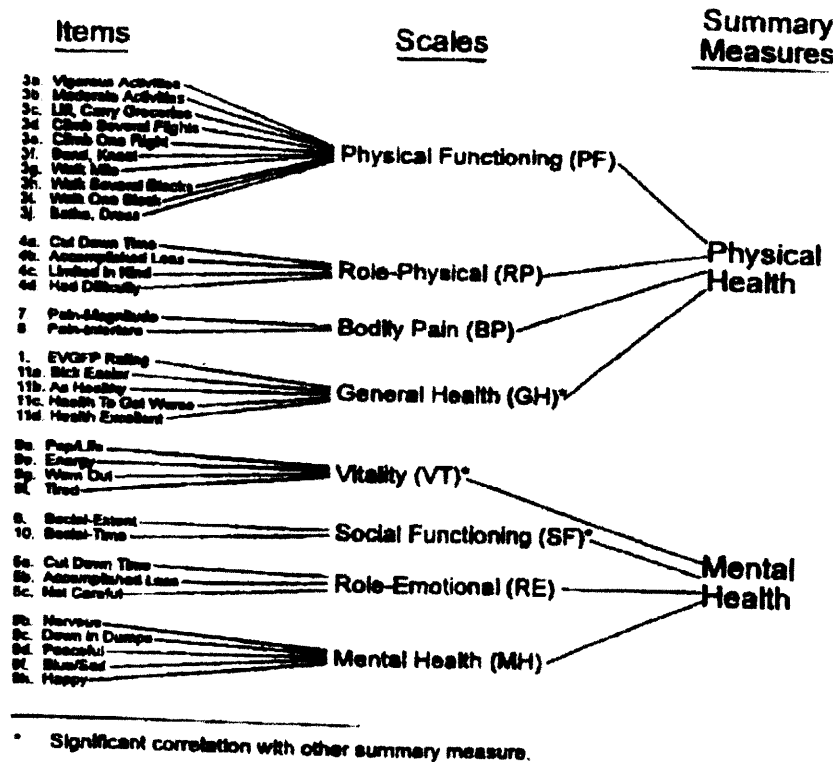


FIG. 24.1. SF-36 and SF-12 Measurement Model. From Ware, Kosinski, and Keller (1994, 1996). Those items in boxes were selected for SF-12.

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 (American Psychological Association, 1974). Extensive psychometric testing has been conducted on the SF-36 in the United States (Garratt, Ruta, Abdalla, Buckingham, & Russell, 1993; Jenkinson, Coulter, & Wright, 1993; McHorney, Ware, Lu, & Sherbourne, 1994; Wagner et al., 1995) and other countries (Bullinger, 1995; McCallum, 1995; Rampal, Martin, Marquis, Ware, & Bonfils, 1994; Sullivan, Karlsson, & Ware, 1994; Sullivan et al., 1995). Using the same tests of scaling and scoring assumptions that were used in developing the SF-36, results have been compared across general population studies in 10 countries (Gandek & Ware, 1998).

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 (Medical Outcomes Trust, 1991; Ware et al., 1993). This method