THE SHORT-FORM-36 HEALTH SURVEY
(Rand Corporation and John E. Ware Jr., 1990, revised 1996)

Purpose

The 36-item short form of the Medical Outcomes Study questionnaire (SF-36) was designed as a generic indicator of health status for use in population surveys and evaluative studies of health policy. It can also be used in conjunction with disease-specific measures as an outcome measure in clinical practice and research (1).

Conceptual Basis

The SF-36 derived from the work of the Rand Corporation of Santa Monica during the 1970s. Rand's Health Insurance Experiment compared the impact of alternative health insurance systems on health status and utilization (2; 3, p2:3). The outcome measures developed for the study have been widely used and several are described in this book. They were subsequently refined and used in Rand's Medical Outcomes Study (MOS), which focused more narrowly on care for chronic medical and psychiatric conditions (4; 5). The MOS surveys were comprehensive, covering 40 physical and mental health concepts, and several abbreviated forms were produced. An 18-item scale was produced in 1984, followed by the 20-item Short-Form (SF-20) in 1986. The SF-36 was constructed to answer criticisms of limitations in the SF-20; it covered the eight most important of the original 40 concepts (3, p2:3). The SF-20 was reviewed in the 1996 edition of Measuring Health, but has largely been replaced by a more recent abbreviation, the SF-12, which is reviewed separately in this edition. Note that the physical functioning items in the SF-36 are also contained in the MOS Physical Functioning Measure, reviewed in Chapter 3; the mental health questions are derived from the Mental Health Inventory that is described in Chapter 5. Fuller details of the origins of the items in the SF-36 are given in the SF-36 manual, which shows the derivation of each item (3, Table 3.4). The development of the SF-36 is now coordinated by QualityMetric, Incorporated, based in Rhode Island (see Address section), although Rand continues to provide information on its version of the SF-36, designated the “Rand 36-item Health Survey 1.0” (www.rand.org/health/surveys/sf36item/).

As a generic instrument, the SF-36 was designed to be applicable to a wide range of types and severities of conditions. Generic instruments are useful for monitoring patients with multiple conditions, for comparing the health status of patients with different conditions, and for comparing patients to the
general population (4, p912). Ware argued that a generic measure should cover both physical and mental concepts and should measure each concept in several contrasting ways. These include behavioral functioning, perceived well-being, social and role disability, and personal evaluations of health in general (6, p3:2). Measures of behavioral functioning and role limitations include questions on work, self-care, mobility, etc. Perceived well-being is subjective and cannot be completely inferred from behavior; hence the SF-36 includes questions on feeling states. The questions on overall evaluation of health provide a summary indicator and capture the impact of health problems not directly included in the other questions (6, p3:3).

Description

The items in the SF-36 were drawn from the original 245-item MOS questionnaire; nine items from the SF-20 were retained in the SF-36 while a further five were reworded, as were several of the answer categories. Several of the mental health items originated from the General Well-Being Schedule of Dupuy. Ware and Sherbourne give an extended description of the origins of the SF-36 and of its links with the other MOS instruments (1). The SF-36 includes multi-item scales to measure the following 8 dimensions (the question numbers refer to those in Exhibit 10.29):

- **PF** - Physical functioning (ten items in question 3)
- **RP** - Role limitations due to physical health problems (four items in question 4)
- **BP** - Bodily pain (questions 7 and 8)
- **SF** - Social functioning (questions 6 and 10)
- **MH** - General mental health, covering psychological distress & well-being (five items: questions 9 b, c, d, f, and h)
- **RE** - Role limitations due to emotional problems (questions 5 a, b, and c)
- **VT** - Vitality, energy or fatigue (four items: questions 9 a, e, g, and i)
- **GH** - General health perceptions (five items: questions 1 and 11 a to d)

In addition, question 2 covers change in health status over the past year; this is not counted in scoring the eight dimensions, but is used to estimate change in health from a cross-sectional administration of the SF-36 (6, p9:15). Version 1 of the SF-36 was described in the 1996 edition of this book.

To address criticism of the layout and wording of some items, Version 2 of the SF-36 changed the dichotomous answers for the role questions to 5-point scales, and slightly altered the wording of several items (7, Figure 1). Modifications to the SF-36 wording over time are indicated in the manual (6, Table 3.4). The Exhibit shows Version 2.

The standard version uses a four-week recall period, but an acute version uses a one-week recall and is suitable for use when the measure is administered repeatedly on a weekly basis. The SF-36 may be self-administered or used in personal or telephone interviews; machine-readable forms and instruction sheets for each version are available from The Health Institute. McHorney et al. compared mail and telephone survey administration; mail was significantly cheaper, and provided a higher response rate;
mail also identified a higher level of disability (8). The questions generally take five to ten minutes to complete; elderly respondents may require about 15 minutes (9; 10). Self-administration appears acceptable and feasible for most patients, although the optical scan forms may be difficult for people with vision problems (11). A version for computerized administration appears acceptable to respondents (12). Nonresponse rates averaged 3.9% across the 36 items in a large study of people with chronic conditions (13, p46). Detailed administration instructions are contained in Chapter 4 of the manual.

Scoring. Two sets of scores are derived from the SF-36: a profile of eight section scores, and two summary scores, one for the physical component (PCS) and one for the mental component (MCS) summary scores. For each set of scores, two alternative approaches may be used in calculating scores: a normal, additive approach that produces 0-to-100 scores for the eight scales (3, Chapter 6), and a norm-based approach that adjusts these raw scores to have a mean of 50 and a standard deviation of 10 (14, Chapter 5).

The first step is to check for out-of-range values, and then to orient all item scores so that high scores correspond to better health. The codes shown in the exhibit are replaced for several items In the first approach, each of the item scores is oriented so that a higher score represents better health. Values for items 1, 7, and 8 are recoded, using weights derived from Likert analyses. For item 1, excellent is scored 5.0, very good = 4.4, good = 3.4, fair = 2.0, and poor = 1.0. For item 7, none = 6.0, very mild = 5.4, mild = 4.2, moderate = 3.1, severe = 2.2, and very severe = 1.0. Scores for item 8 take account of the answers to item 7: if no pain is recorded on either item, then item 8 is scored 6. If item 8 is answered not at all, but item 7 > none, then item 8 is scored 5. For the remaining categories of item 8, a little bit = 4, moderately = 3, quite a bit = 2, and extremely = 1. However, it item 7 was not answered, the values for item 8 are: not at all = 6.0, a little bit = 4.75, moderately = 3.3, quite a bit = 2.25, and extremely = 1 (3, Table 6.3). Next, scores for items on each of the eight scales (see above) are added to give scale scores. Finally, these are linearly transformed to a 0-to-100 scale (6, Table 6.11; 13; 15). The formula is

\[
\text{Transformed scale} = \frac{\text{actual score} - \text{lowest possible score}}{\text{possible raw score range}} \times 100
\]

A missing value is given for a scale if over half of the items are missing; where fewer items are missing, these are replaced by that respondent's mean scores on the remaining items in the scale (13, p44). However, the scoring software from www.sf-36.com provides a more sophisticated adjustment for missing values (14, p41).

The 0-to-100 scales can also be transformed into norm-based scores which translate raw scores into a position on the population distribution of scores and so requires population norms. Ideally, norms
should be taken from the country in which the study is being undertaken. This simplifies interpretation, as the user no longer has to remember the (different) norms for each scale. It also has the advantage of ensuring comparability across different versions of the SF-36 (14, p26). First, the 0 - 100 scores for the eight subscales are standardized using a z-score transformation, which involves subtracting the population mean score for that scale from each respondent’s score, and dividing the difference by the population standard deviation. For example, the raw score for Physical Function is transformed by subtracting 83.29094 and dividing by 23.75883; the formulae are given in the scoring manual (14, Table 6.12). Next, to give a mean of 50 and standard deviation of 10, the z-score is multiplied by 10 and 50 is added to the product.

To produce the PCS physical summary score, the z-scores for each of the eight scales is multiplied by a factor score coefficient and the resulting scores summed over the eight subscales. The same is done for the MCS score and formulae are given in the manual (14, p51). Finally, the PCS and MCS scores are translated into t-scores (with a mean of 50 and a SD of 10) by multiplying the PCS and MCS scores by 10 and adding 50 to the product.

The influence of health economics and the demand for estimates of quality-adjusted survival have led to the development of ways to transform SF-36 scores into a utility score. One approach is to use regression analysis on a data set containing both SF-36 scores and a utility measure; this has been undertaken using the Health Utilities Index (HUI) (16), the Quality of Well-Being scale (17) and the EuroQol EQ-5D (18). Alternatively, selected permutations of response categories from the SF-36 can be valued using a utility estimation procedure such as the standard gamble. Utility scores for other permutations not directly evaluated can be interpolated using regression methods. For example, a regression method for transforming SF-36 scores into equivalent scores on the Quality of Well-Being scale has been described (17).

Reliability

McHorney et al. based comprehensive analyses of item response, reliability and validity on a sample of 3,445 patients with chronic medical or psychiatric conditions drawn from the MOS study (13). Item analyses confirmed the assignment of the items to the eight scales; this was replicated in different patient groups (13, p51). In the International Quality of Life Assessment project (IQOLA), studies in 11 countries showed that items generally correlated more highly with their own scales than with others (19, Table 5).

Alpha internal consistency coefficients for the eight scales have been reported from many studies; Ware et al. listed results from 14 (6, Table 7.2), and then from 11 countries under the IQOLA project (19, Table 6). Combining results from these studies, the median alpha reliability for all scales exceeds 0.80, except for the two-item social functioning scale (0.76); in Essink-Bot’s study the mean alpha for the SF-36 scales was 0.84, compared to 0.72 for the Nottingham Health Profile (20, pp 528-9). Typical results are illustrated in Table 10.6, in which the first line shows values from the study of over 163,000 elderly people. Slightly lower values were reported by Kurtin et al. (11, Table 5). All scales
appear sufficiently reliable for comparing groups, and the physical functioning scale appears reliable enough for comparing individuals. The intraclass correlation was 0.85 for patients with musculoskeletal problems (21). Item-total correlations typically lie in the mid-0.70s (6, Table 5.2).

Two-week test-retest correlations exceeded 0.8 for physical function, vitality, and general health perceptions; the lowest coefficient was 0.6 for social function (22, Table II). Assessing agreement, the mean of the differences in scores did not exceed one point on the 100 point scale (22, p162). Test-retest correlations for the scales after a delay of 6 months ranged between 0.60 and 0.90, except for the pain dimension, with a correlation of 0.43 (23, Table 5). Slightly lower results were reported in a British study, in which 6-month retest Spearman correlations ranged from 0.28 (SF) to 0.70 (VT) (24, Table 2).

Ware et al. provided a series of tables indicating the estimated sample size requirements, based on the reliability of the scales, for showing a given difference in scores as being statistically significant (3, Tables 7.4 to 7.8).
### Table 10.6 Cronbach Alpha Coefficients for SF-36 Scales from Various Studies

<table>
<thead>
<tr>
<th>Physical function (PF)</th>
<th>Role, physical (RP)</th>
<th>Bodily pain (BP)</th>
<th>Social functioning (SF)</th>
<th>Mental health (MH)</th>
<th>Role, emotional (RE)</th>
<th>Vitality (VT)</th>
<th>General health perceptions (GH)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.93</td>
<td>0.91</td>
<td>0.88</td>
<td>0.85</td>
<td>0.83</td>
<td>0.88</td>
<td>0.87</td>
<td>0.83</td>
<td>(25, Table 5)</td>
</tr>
<tr>
<td>0.88</td>
<td>0.90</td>
<td>0.80</td>
<td>0.77</td>
<td>0.82</td>
<td>0.80</td>
<td>0.88</td>
<td>0.83</td>
<td>(26)</td>
</tr>
<tr>
<td>0.93</td>
<td>0.84</td>
<td>0.82</td>
<td>0.85</td>
<td>0.90</td>
<td>0.83</td>
<td>0.87</td>
<td>0.78</td>
<td>(13)</td>
</tr>
<tr>
<td>0.93</td>
<td>0.96</td>
<td>0.85</td>
<td>0.73</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>(22)</td>
</tr>
<tr>
<td>0.90</td>
<td>0.88</td>
<td>0.82</td>
<td>0.76</td>
<td>0.83</td>
<td>0.80</td>
<td>0.85</td>
<td>--</td>
<td>(27)</td>
</tr>
<tr>
<td>0.93</td>
<td>0.89</td>
<td>0.90</td>
<td>0.68</td>
<td>0.84</td>
<td>0.82</td>
<td>0.86</td>
<td>0.81</td>
<td>(6)</td>
</tr>
<tr>
<td>0.92</td>
<td>0.89</td>
<td>0.86</td>
<td>0.80</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.83</td>
<td>(28, Table 1)</td>
</tr>
<tr>
<td>0.92</td>
<td>0.95</td>
<td>0.85</td>
<td>0.85</td>
<td>0.84</td>
<td>0.92</td>
<td>0.84</td>
<td>0.80</td>
<td>(7, Table 2)</td>
</tr>
<tr>
<td>0.91</td>
<td>0.86</td>
<td>0.81</td>
<td>0.56</td>
<td>0.80</td>
<td>0.83</td>
<td>0.84</td>
<td>0.66</td>
<td>(24, Table 1)</td>
</tr>
<tr>
<td>0.93</td>
<td>0.88</td>
<td>0.85</td>
<td>0.82</td>
<td>0.86</td>
<td>0.83</td>
<td>0.83</td>
<td>0.77</td>
<td>(29, Table 1)</td>
</tr>
</tbody>
</table>
Validity

The SF-36 manual presents criterion validity information on the scales, comparing scale scores to ability to work, symptoms, utilization of care, and to a range of criteria for the mental health scale (6, Chapter 9). Each comparison suggested significant and consistent associations with the validation criteria. Item 2, on self-reported change, was evaluated in a study that applied a general health rating twice, at an interval of one year (6, Table 9.11). There was substantial agreement, although there were some errors at the ends of the scale: 6.9% of those who said they were much better had worsened, and 3.4% of those who reported being much worse had improved. McHorney et al. compared SF-36 scale scores for patients with varying levels of medical and psychiatric conditions and with combinations of both. The scales discriminated between types and levels of disease and were also able to distinguish people with a chronic medical condition alone from those who had a medical disorder combined with a psychological one (6, pp9:21-9:23; 30, pp255-259). From these analyses, McHorney et al. provided guidelines for interpreting the eight scales. The physical functioning and mental health scales are relatively pure, being specific to medical or to psychiatric disorders. The two role scales chiefly reflect physical or mental conditions, but not exclusively. By design, the social functioning and vitality scales reflect both physical and mental conditions. The general health perceptions scale appears to be most sensitive to physical health problems. Nerenz et al. compared patient SF-36 ratings with physicians' ratings of the eight dimensions; these fell between 0.39 and 0.64 (23, pMS117). In general, physicians rated patients healthier than did the patients themselves.

Predictive validity for physician visits was greatest for the pain and health perceptions scales; the mental health and role - emotional scales best predicted hospital admissions, while vitality, general health perceptions and physical function scales best predicted mortality over four years (10, pp576-8).

Principal components analyses of the MOS scale scores indicated that a general health dimension was common to all eight SF-36 scales, explaining 55% of the variance (30, p254). When a two-component solution is extracted, physical and mental dimensions are clearly distinguished; the GH and VT scales tend to load on both, with GH loading mainly on the physical factor and VT on the mental factor (6, Table 9.12; 30, Table 1; 31, Tables 2 to 4). A slightly different two-factor solution was derived from analysis of the eight scales by Kazis et al. The PF, RP, BP, GH and VT loaded on one factor, and the SF, RE and MH loaded on the second, supporting the interpretation of physical and emotional dimensions (32, Table 3). This was further supported in the analysis of over 163,000 respondents in the Health Outcomes Study (25, p19). An analysis of the individual items provided an eight-factor solution that corresponded to the original structure of the SF-36. However, a nine-factor solution fit the data significantly better (33). Correlations among the eight scales were reported by Nerenz et al. (23, Table 2). These broadly correspond to the groupings found by McHorney, except that the social functioning scale was more closely associated with physical than mental functioning in the Nerenz study. A structural equation modeling analysis using a large data set assembled from 10
countries in general supported the conceptual structure of the SF-36 as comprising eight first-order domains, three second-order domains (physical and mental health, and well-being, which comprised the GH and VT scales), and a single underlying construct of health (34, p1185). Jette et al. provided an interesting analysis in which they combined items from the PF scale, the Functional Independence Measure (FIM), the Minimum Data Set (MDS) and the Outcome and Assessment Information Set for Home Health Care (OASIS). Using a Rasch analysis, they presented the scope of coverage of the four scales on a 0 (greatest disability) to 100 (no disability) scale. The PF scale items covered higher levels of function, running from 50 to 100 (35, Figure 2). The FIM ranged from 25 to 86, while the MDS and OASIS scales covered almost the entire disability range.

Correlations with the Sickness Impact Profile (SIP) were 0.78 for 106 hip surgery patients (36, Table 3); in a separate study the correlations were 0.73 for overall functioning, 0.78 for physical function, and 0.67 for social function (N = 25 elderly males) (9, Table 2). A comparison of the SF-36 and the Quality of Well-Being Scale (QWB) for 916 respondents indicated that the SF-36 accounted for 55.7% of the variance in the QWB scores; conversely, the QWB accounted for 63.7% of the variance in SF-36 physical functioning scores, for 33.8% of variance in general health perceptions, and for 49.2% of physical role functioning (37). Correlations of the eight scales of the British version of the SF-36 with the EuroQol Quality of Life Index ranged from 0.48 to 0.60 (p < 0.01) (38, p173). Ware et al. present a table listing correlations between the SF-36 and 15 other health measures. Correlations for the mental health scale range from 0.51 to 0.82 with the corresponding scales in other leading measures; equivalent correlations for the physical function scale range from 0.52 to 0.85 (6, Table 9.16). Brazier et al. present a full multitrait-multimethod correlation matrix comparing the SF-36 dimension scores with those of the Nottingham Health Profile. While the correlations between comparable dimensions in the two instruments exceeded those between non-comparable dimensions, the coefficients were not high (0.52 for the physical scale, 0.55 for pain, 0.67 for mental health, and 0.68 for vitality) (22, Table IV).

The SF-36 appears sensitive to change: an effect size of 0.67 in a study of musculoskeletal patients was higher than that for the Nottingham Health Profile, SIP, or the Duke DUHP (21). Effect sizes in a study of hip replacement (6-12 months after the operation) were 0.82 (overall), 1.45 for the Pain scale, 1.33 for PF, 1.22 for RP, and 0.8 for Social. These were markedly higher than corresponding statistics for the London Handicap Scale (39, Table 4). In another study of effect sizes for hip replacements, the effect size for the PCS was 1.26; that for the Pain scale was 1.73 and PF was 1.37. These figures were higher than figures for the HUI-2 or HUI-3 (40, Table 2). Effect sizes for detecting long standing illness in elderly patients ranged from 0.31 (MH) to 0.96 (PF); these were comparable to the effect size for the EuroQol EQ-5D (24, Table 5). In a comparison of sensitivity to change in a study of hip replacement patients, the ranking of five instruments depended on whether the overall scores were used or the physical or psychological scores (36, Table 4). Using the overall scores, SF-36 was more sensitive to change than the SIP but less so than the Arthritis Impact Measurement Scales or the Functional Status Questionnaire. The physical score of the SF-36 was more sensitive than all but one of the other measures, but the psychological score was the least sensitive to change. In a study of migraine, the SF-36 produced larger effect sizes than did the COOP charts (41, Tables 2, 6).
Likewise, the SF-36 was slightly superior to the NHP in a study of myocardial infarction (42). Jenkinson et al. comment that the SF-36 may not be adequately sensitive to change when used in survey research (43), although it was superior to the EQ-5D (44). Essink-Bot compared the ability of scores on four measures to discriminate between people absent from work due to illness and others. The SF-36 was the most discriminating (mean area under the ROC curve 0.72), followed by the COOP Charts (mean AUC 0.64), the EuroQol EQ-5D (mean AUC 0.61) and the Nottingham Health Profile (mean AUC 0.60) (20, Table 5). In a study of hernia surgery, the SF-36 and COOP charts provided comparable effect size statistics; values for the physical, pain and social function scales for both scales fell in the range 1.0 to 1.85 (45, Table 4). The SF-36 manual presents confidence intervals and estimates of the sample sizes required to detect differences of various sizes under a variety of statistical designs for each SF-36 scale (6, Tables 7.4-7.9). Ferguson et al. have calculated reliable change index thresholds for the eight subscales by age and sex (46, Table 2). Values (in terms of the T-scores) ranged from 7.35 for the PF scale to 15.8 for the SF scale.

McHorney and others have reviewed ceiling and floor effects (10, Table 1; 13, p53; 19, Table 7; 25, Table 6). Both effects are most common in the RP and RE scales, but the ceiling effect is worse, and PF, BP and SF produce important ceiling effects (see, for example, reference 19, Table 6). However, in a sample of relatively well people, the SF-36 identified minor levels of discomfort that were missed by the Nottingham Health Profile (22). In a general population sample, the SF-36 demonstrated less of a ceiling effect than the EuroQol, on which between 64% and 95% of respondents achieved maximum scores on the dimensions of the instrument, compared to 37% to 72% for the various dimensions of the SF-36 (38, p173).

O’Brien et al. compared scores on the SF-6D (described below) with the Health Utilities Index for a sample of 310 cardiac patients. The results showed poor agreement between the two scales: the HUI showed a much wider range of utility scores (-0.21 to 1.00) than the SF-6D (0.3 to 0.95). O’Brien obtained an intraclass correlation between the SF-6D and the HUI of 0.42 ($R^2 = 0.34$) (47, p978). The correlations among the six dimensions of the SF-6D were higher than for the HUI, but the low correlations are an intentional design feature of the HUI. Similar findings in terms of a narrower distribution of scores for the SF-6D and a Pearson correlation of 0.69 with the HUI, and 0.66 with the EQ-5D were reported in a study of rheumatology patients (48).

**Alternative Forms**

There are several abbreviations of the SF-36, of which the SF-12 has become sufficiently well-tested to merit a separate entry in this chapter; the second edition of *Measuring Health* also contained a separate entry for the SF-20, but this has now largely been supplanted by the SF-12. Among the other abbreviations, the SF-8 uses one question to represent each of the SF-36 domains and requires one to
two minutes to complete. Physical and mental component scores are produced, along with eight single-item domain scores (49, p16). Unlike the SF-12, however, items in the SF-8 are not a subset of those in the SF-36. Ware et al. reported a test-retest reliability of 0.73 for the PCS-8 and 0.74 for the MCS-8 (50). The scale may be scored using an online scoring service at www.qualitymetric.com. In a study of migraine patients, correlations between the SF-8 and corresponding SF-36 scores ranged from 0.67 (role-emotional scale) to 0.84 (pain) (51, Table 2).

Minor variants of the SF-36 include the acute version in which the time referent is the past week. The questions are otherwise identical. The Veterans SF-36 Health Survey has made changes to the RP and RE scales and altered the response options for various questions (32; 52). The Rand Corporation has presented an alternative version of the SF-36 that differs in detail of the wording of two questions and in the scoring method (53). A condition-specific variant of the SF-36 has been described for use with knee replacement patients, in which patients were asked to report only limitations in function due to their knee condition (26). A comparison of the condition-specific and generic versions indicated that some of the condition-specific scales (e.g., pain, role limitations) were more sensitive to the effects of treatment than were the equivalent generic scales, which reflected the combined effect of the knee condition and comorbidities (26, pMS248). A 30-item version for use with HIV patients, the MOS-HIV, takes five to 10 minutes to administer and has been widely used in clinical trials (54-56, p339).

Brazier et al. have developed an abbreviated version of the SF-36 that can be scored using a utility-based approach. This is called the SF-6D, and includes 8 items from the SF-12 and 3 more from the SF-36 covering six dimensions: physical functioning, role limitations, social functioning, pain, mental health and vitality. Between two and six levels are specified in each dimension (57, Figure 1; 58, Table 1). Utility weights have been developed using a standard gamble (57; 58).

Through the IQOLA project, the SF-36 has been translated and adapted for use in many languages, including Swedish (59-61), German, Spanish, French, Italian, Danish (62), Dutch, and Japanese (6, p12:7; 63; 64, p381). A Hebrew version has been described (65). A Chinese version has been designed for use in the USA (66-68). A British-English version alters the phrasing of six items; for example, "block" was standardized to "100 yards" and "feeling blue" was changed to "feeling low" (7; 22; 27).

An alternative form of the five-item mental health scale has been validated; this is useful in studies requiring repeated administrations of the scale over short time periods (69).

Reference Standards

Norms for the general U.S. population were presented for seven age groups and by sex in the original SF-36 manual, as were norms for 13 different medical conditions (6, Chapter 10; 8). The manuals for version 2 provide extensive sets of norms for the U.S. population (3, pp10:14 to 10:38; 14, pp63-102). U.S. norms for each subscale are also shown in (70, Table 1). Reference standards for elderly people aged 65 and above are presented for the eight scales by disability status, based on a sample of
177,714 participants in the Health Outcomes Study (25, Figure 1). The same source also gives means and standard deviations for each item (Table 1). Norms for both the SF-36 and SF-12 from a much smaller U.S. sample of 2329 are available (71, Tables 4 to 12).

Norms from large British samples show mean scores by age, gender, socioeconomic class, and by the presence of chronic conditions (7, Tables 3 and 5; 22, Table III). Several other sets of British norms are available (24, Table 3; 28, Table 3; 72; 73). Canadian norms are available (74), as are figures from a survey of 42,000 Australian women (75), and norms from Romania (76). Table 10.7 illustrates some of these results. The Rand figures in the table were based on 2,471 responses from the MOS and use the Rand scoring approach described earlier.

An interpretation of effect sizes can be derived from the comparison of patients with different types of chronic disease. McHorney et al. noted that a difference of 23 points on the physical functioning scale reflects the impact of a complicated chronic medical condition, while a difference of 27 points on the mental health scale is equivalent to the impact of serious depressive symptoms (30, p261).
Table 10.7  Population Norms for SF-36 Scales.  *Note:* Figures show mean scores on the 0 - 100 scale (standard deviations in parentheses)

<table>
<thead>
<tr>
<th>Sample</th>
<th>Physical function (PF)</th>
<th>Role, physical (RP)</th>
<th>Bodily pain (BP)</th>
<th>Social functioning (SF)</th>
<th>Mental health (MH)</th>
<th>Role, emotional (RE)</th>
<th>Vitality (VT)</th>
<th>General health perceptions (GE)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>70.6 (27.4)</td>
<td>53.0 (40.8)</td>
<td>70.8 (25.5)</td>
<td>78.8 (25.4)</td>
<td>70.4 (22.0)</td>
<td>65.8 (40.7)</td>
<td>52.2 (22.4)</td>
<td>57.0 (21.1)</td>
<td>(53)</td>
</tr>
<tr>
<td>Elderly males</td>
<td>52.0 (30.7)</td>
<td>-</td>
<td>-</td>
<td>66.2 (35.6)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>(9)</td>
</tr>
<tr>
<td>U.S. general population</td>
<td>84.2 (23.3)</td>
<td>81.0 (34.0)</td>
<td>75.2 (23.7)</td>
<td>83.3 (22.7)</td>
<td>74.7 (18.1)</td>
<td>81.3 (33.0)</td>
<td>60.9 (21.0)</td>
<td>72.0 (20.3)</td>
<td>(6)</td>
</tr>
<tr>
<td>Canada</td>
<td>85.8 (20.0)</td>
<td>82.1 (33.2)</td>
<td>75.6 (23.0)</td>
<td>86.2 (19.8)</td>
<td>77.5 (15.3)</td>
<td>84.0 (31.7)</td>
<td>65.8 (18.0)</td>
<td>77.0 (17.7)</td>
<td>(74)</td>
</tr>
<tr>
<td>Australian women 45-49 years</td>
<td>85.1 (18.7)</td>
<td>79.6 (35.2)</td>
<td>70.7 (23.8)</td>
<td>81.4 (23.7)</td>
<td>72.1 (18.0)</td>
<td>77.0 (36.3)</td>
<td>58.1 (20.9)</td>
<td>71.9 (20.6)</td>
<td>(75)</td>
</tr>
<tr>
<td>British males</td>
<td>89.8 (18.8)</td>
<td>89.0 (21.1)</td>
<td>81.3 (22.2)</td>
<td>84.7 (22.6)</td>
<td>74.3 (17.2)</td>
<td>88.1 (19.9)</td>
<td>60.8 (18.9)</td>
<td>70.9 (20.3)</td>
<td>(7, Table 3)</td>
</tr>
<tr>
<td>British females</td>
<td>86.7 (20.2)</td>
<td>85.8 (22.5)</td>
<td>77.0 (23.4)</td>
<td>81.3 (23.6)</td>
<td>70.1 (18.7)</td>
<td>84.1 (21.8)</td>
<td>55.9 (19.9)</td>
<td>71.3 (20.5)</td>
<td>(7, Table 3)</td>
</tr>
</tbody>
</table>
Commentary

The SF-36 achieved a meteoric rise to prominence; the New England Health Institute estimated that by 1992 a million forms were being administered each year, even before evidence on reliability and validity had accumulated. The publications followed; from a handful in 1991 to 250 in 1997, 300 in 1998 and 400 in 1999. The instrument clearly met a need and was also carefully promoted; there are several advantages in this. Attention is being paid to ensuring its standard administration. This was facilitated through the Medical Outcomes Trust, a nonprofit organization created to support the development and distribution of standardized outcome measures. Permission to use the SF-36 should be obtained from QualityMetric, which provides updates on its administration, scoring, and interpretation. The SF-36 is in the public domain, and no royalties are required for using it. The IQOLA project was established to translate and adapt the SF-36 for use outside the USA (61); the project evaluates the psychometric properties of the instrument, gathers general population norms and documents findings in accessible documents. The SF-36 Manual is comprehensive and exemplary, and is a necessity for anyone wishing to administer the instrument (3). Attention to detail is outstanding. For example, 15 internal consistency checks are programmed into the SF-36 scoring to identify respondents who may be making internally inconsistent responses (6, p7:16). The manuals provide clear guidelines for administration, details on scoring and an extensive summary of validity results and normative data (3; 14).

The history of the SF-36 reflects the challenges inherent in designing any general health measurement. The tradeoff between covering many topics superficially and providing detailed coverage of a few translates into comprehensiveness versus precision. In this vein, the first abbreviation of the original MOS instruments, the SF-20, was soon criticized as being too narrow. There ensued an explosion of rival abbreviations beginning with a 30-item short form (54); subsequently, 36-, 38-, and 56-item versions appeared (77). The SF-36 now replaces these; the SF-20 has largely been supplanted by the more recent SF-12 and SF-8 Surveys. A comparison of the shortened versions is included in the QualityMetric web site.

The scope of the SF-36 is broad and the “quality of life” label appears justified as factors outside the health domain correlate with scores. The SF-36 is one of the few generic scales to include items on positive health (20). As with any brief, generic instrument, the scope of the SF-36 can be criticized, and some have commented on the absence of cognitive function or distress (77). The physical activity items focus on gross activities such as walking, bending, and kneeling, while coordinated actions that might be captured by items such as shopping or cooking are not covered (64, p380). Floor effects were reported for the role scales in version 1 (11; 64, p379), suggesting that these scales might not detect further deterioration in the condition of relatively sick patients. The revisions in Version 2 appear to have reduced the problem (7, p49). The use of the SF-36 with cognitively impaired elderly people has received some attention, and it appears that it can be used reliably by such patients (10, p579).

Various comments have been made on item wording and Lessler gave a summary (78). The first question concerning health in general may be difficult: is the respondent to undertake a form of
averaging where their health has fluctuated? Is "in general" intended to exclude specific health problems? Or, does "in general" refer to the time preceding the current, acute symptoms they are experiencing? Some items seem inappropriate for elderly people ("... vigorous activities, such as running, lifting heavy objects"; "How much did pain interfere with your normal work?") (79). Question 10 refers to the impact of physical health or emotional problems on social activities. "Physical health" suggests a continuum, whereas "emotional problems" is clearly negative, and may even imply that the person did experience emotional problems. It is also curious that two items (numbers 6 and 10 in the Exhibit) should so closely replicate each other.

A debate has arisen over the appropriateness of the scoring system for the PCS and MCS which was based on factor-loadings. Taft et al. argued that the negative weights given to physical items on the mental scale (and vice-versa) could contaminate each scale. That is, a perfect score on the MCS can be obtained only when the respondent scores 100 on each scale in the MCS, and there are no subtractions due to scores on the physical scale (i.e., the physical score must be zero, and vice-versa for the PCS calculation) (80). Ware and Kosinski responded, arguing mainly on empirical grounds that this does not occur frequently enough to be an issue (81), but other authors have noted the same problem with the subscale scoring (82; 83).

With the active debate over the choice between health indexes and health profiles, the comparisons between the SF-6D and the HUI are instructive. While both used variants of a standard gamble approach to derive utility weights, the agreement between the two methods is low (47; 48). This may in part reflect differing item content: the HUI focuses on a ‘within a the skin’ definition of health, whereas the SF questions cover handicap and function in the person’s environment. The SF instruments also cover positive aspects of health not included in the HUI. There were also significant methodological differences in the way weights were derived (47, p980). The bottom line is that QALYs estimated from the two instruments cannot be directly compared and it is not clear which provides the more accurate picture.

It is clear that the SF-36 will continue to be the leading general health measure for many years to come.

**Addresses**

The SF-36 web site is at www.sf-36.org, but extensive information on ordering copies of the scale, manuals, etc., is available from www.qualitmetric.com/products/SFSurveys.shtml . Information on licencing is available through www.qualitmetric.com/products/descriptions/sflicenses.shtml

An educational series concerned with understanding health measures is available from www.healthstatprod.com

Information on translations is available from the International Quality of Life Assessment (IQOLA)
References


(11) Kurtin PS, Davies AR, Meyer KB, DeGiacomo JM, Kantz ME. Patient-based health status


(20) Essink-Bot M-L, Krabbe PFM, Bonsel GJ, Aaronson NK. An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOOP/WONCA charts, and the EuroQol instrument. Med Care 1997; 35:522-537.


(30) McHorney CA, Ware JE, Jr., 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; 31:247-263.


(33) Wolinsky FD, Stump TE. A measurement model of the Medical Outcomes Study 36-item Short-Form Health Survey in a clinical sample of disadvantaged, older, black and white men and women. Med Care 1996; 34:537-548.
(34) Keller SD, Ware JE, Jr., Bentler PM. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: results from the IQOLA project. J Clin Epidemiol 1998; 51:1179-1188.


(51) Turner-Bowker DM, Bayliss MS, Ware JE, Jr., Kosinski M. Usefulness of the SF-8 Health Survey for comparing the impact of migraine and other conditions. Qual Life Res 2003; 12:1003-1012.


(53) Rand Health Sciences Program. Rand 36-Item Health Survey 1.0. Santa Monica, California: Rand Corporation, 1992.


(69) McHorney CA, Ware JE, Jr. 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; 33:15-28.


(71) Ware JE, Jr., Kosinski M, Keller SD. How to score the SF-12 physical and mental health summary scales. Boston, MA: The Health Institute, New England Medical Center, 1995.


(78) Lessler JT. Choosing questions that people can understand and answer. Med Care 1995; 33(suppl):AS203-AS208.


Exhibit 10.29  The Short-Form-36 Health Survey, Version 2

Your Health and Well-Being

This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. Thank you for completing this survey!

For each of the following questions, please mark an X in the one box that best describes your answer.

1. **In general, would you say your health is:**

   Excellent ▼
   Very Good ▼
   Good ▼
   Fair ▼
   Poor ▼

   G1 G2 G3 G4 G5

2. **Compared to one year ago, how would you rate your health in general now?**

   Much better than one year ago ▼
   Somewhat 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 ▼

   G1 G2 G3 G4 G5

3. **The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?**

   Yes, Limited ▼
   Yes, A Lot ▼
   No, Not Limited ▼
   A Little ▼
   At All ▼

a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports

b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf

c. Lifting or carrying groceries

d. Climbing several flights of stairs

e. Climbing one flight of stairs

f. Bending, kneeling, or stooping

g. Walking more than a mile

h. Walking several blocks

i. Walking one block

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?

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>Cut down on the amount of time you spent on work or other activities</td>
<td>G1</td>
</tr>
<tr>
<td>b.</td>
<td>Accomplished less than you would like</td>
<td>G1</td>
</tr>
<tr>
<td>c.</td>
<td>Were limited in the kind of work or other activities</td>
<td>G1</td>
</tr>
<tr>
<td>d.</td>
<td>Had difficulty performing the work or other activities (for example, it took extra effort)</td>
<td>G1</td>
</tr>
</tbody>
</table>

5. During the past 4 weeks, have you had any of the following problems with your work or other regular activities as a result of any emotional problems (such as feeling depressed or
THE SHORT-FORM-36 HEALTH SURVEY

anxious)?

Yes \hspace{1cm} No
\begin{itemize}
  \item[G_1] Cut down on the amount of time you spent on work or other activities
  \item[G_1] Accomplished less than you would like
  \item[G_1] Did work or other activities less carefully than usual
\end{itemize}

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

Not at all \hspace{1cm} A little bit \hspace{1cm} Moderately \hspace{1cm} Quite a bit \hspace{1cm} Extremely
\begin{itemize}
  \item[G_1] \item[G_2] \item[G_3] \item[G_4] \item[G_5]
\end{itemize}

7. How much bodily pain have you had during the past 4 weeks?

None \hspace{1cm} Very mild \hspace{1cm} Mild \hspace{1cm} Moderate \hspace{1cm} Severe \hspace{1cm} Very Severe
\begin{itemize}
  \item[G_1] \item[G_2] \item[G_3] \item[G_4] \item[G_5] \item[G_6]
\end{itemize}

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

Not at all \hspace{1cm} A little bit \hspace{1cm} Moderately \hspace{1cm} Quite a bit \hspace{1cm} Extremely
\begin{itemize}
  \item[G_1] \item[G_2] \item[G_3] \item[G_4] \item[G_5]
\end{itemize}

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...

<table>
<thead>
<tr>
<th>Question</th>
<th>All of the Time</th>
<th>Most of the Time</th>
<th>A Good Bit of the Time</th>
<th>Some of the Time</th>
<th>A Little of the Time</th>
<th>None of the Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Did you feel full of pep?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>b. Have you been a very nervous person?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>c. Have you felt so down in the dumps that nothing could cheer you up?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>d. Have you felt calm and peaceful?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>e. Did you have a lot of energy?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>f. Have you felt downhearted and blue?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>g. Did you feel worn out?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>h. Have you been a happy person?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
<tr>
<td>i. Did you feel tired?</td>
<td>G1</td>
<td>G2</td>
<td>G3</td>
<td>G4</td>
<td>G5</td>
<td>G6</td>
</tr>
</tbody>
</table>
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 friends, relatives, etc.)?  

<table>
<thead>
<tr>
<th></th>
<th>All of the time</th>
<th>Most of the time</th>
<th>Some of the time</th>
<th>A little of the time</th>
<th>None of the time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>▼</td>
<td>▼</td>
<td>▼</td>
<td>▼</td>
<td>▼</td>
</tr>
<tr>
<td></td>
<td>G₁</td>
<td>G₂</td>
<td>G₃</td>
<td>G₄</td>
<td>G₅</td>
</tr>
</tbody>
</table>

11. How TRUE or FALSE is each of the following statements for you?

<table>
<thead>
<tr>
<th></th>
<th>Definitely True</th>
<th>Mostly True</th>
<th>Don't Know</th>
<th>Mostly False</th>
<th>Definitely False</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. I seem to get sick a little easier than other people</td>
<td>G₁</td>
<td>G₂</td>
<td>G₃</td>
<td>G₄</td>
<td>G₅</td>
</tr>
<tr>
<td>b. I am as healthy as anybody I know</td>
<td>G₁</td>
<td>G₂</td>
<td>G₃</td>
<td>G₄</td>
<td>G₅</td>
</tr>
<tr>
<td>c. I expect my health to get worse</td>
<td>G₁</td>
<td>G₂</td>
<td>G₃</td>
<td>G₄</td>
<td>G₅</td>
</tr>
<tr>
<td>d. My health is excellent</td>
<td>G₁</td>
<td>G₂</td>
<td>G₃</td>
<td>G₄</td>
<td>G₅</td>
</tr>
</tbody>
</table>

Thank you for completing these questions!