Wolters Kluwer Health | Lippincott Williams & Wilkins

MEDICAL CARE
Official Journal of the Medical Care Section, American Public Health Association

 Volume 34(3), March 1996, pp 220-233

# A 12-Item Short-Form Health Survey: Construction of Scales and Preliminary Tests of Reliability and Validity

[Original Article]

**WARE, JOHN E. JR. PHD[*,†,‡]; KOSINSKI, MARK MA[*]; KELLER, SUSAN D. PHD[*]**

[*]From The Health Institute, New England Medical Center, Boston, Massachusetts.

[†]From the School of Medicine, Tufts University, Boston, Massachusetts.

[‡]From the School of Public Health, Harvard University, Boston, Massachusetts.

Address correspondence to: John E. Ware, Jr., PhD, The Health Institute, New England Medical Center, #345, 750 Washington Street, Boston, MA 02111.

## Abstract

**Regression methods were used to select and score 12 items from the Medical Outcomes Study 36-Item Short-Form Health Survey (SF-36) to reproduce the Physical Component Summary and Mental Component Summary scales in the general US population (n = 2,333). The resulting 12-item short-form (SF-12) achieved multiple *R* squares of 0.911 and 0.918 in predictions of the SF-36 Physical Component Summary and SF-36 Mental Component Summary scores, respectively. Scoring algorithms from the general population used to score 12-item versions of the two components (Physical Component Summary and Mental Component Summary) achieved *R* squares of 0.905 with the SF-36 Physical Component Summary and 0.938 with the SF-36 Mental Component Summary when cross-validated in the Medical Outcomes Study. Test-retest (2-week) correlations of 0.89 and 0.76 were observed for the 12-item Physical Component Summary and the 12-item Mental Component Summary, respectively, in the general US population (n = 232). Twenty cross-sectional and longitudinal tests of empirical validity previously published for the 36-item short-form scales and summary measures were replicated for the 12-item Physical Component Summary and the 12-item Mental Component Summary, including comparisons between patient groups known to differ or to change in terms of the presence and seriousness of physical and mental conditions, acute symptoms, age and aging, self-reported 1-year changes in health, and recovery from depression. In 14 validity tests involving physical criteria, relative validity estimates for the 12-item Physical Component Summary ranged from 0.43 to 0.93 (median = 0.67) in comparison with the best 36-item short-form scale. Relative validity estimates for the 12-item Mental Component Summary in 6 tests involving mental criteria ranged from 0.60 to 1.07 (median = 0.97) in relation to the best 36-item short-form scale. Average scores for the 2 summary measures, and those for most scales in the 8-scale profile based on the 12-item short-form, closely mirrored those for the 36-item short-form, although standard errors were nearly always larger for the 12-item short-form.**

---

Although the 36-item short-form (SF-36) health survey has proved to be useful for a variety of purposes,[1,2] it is too long for inclusion in some large-scale health measurement and monitoring efforts. Can an even shorter form yield satisfactory results? Two developments have led to a strategy for constructing a shorter version of the SF-36 Health Survey. First, physical and mental health factors have been found to account for 80% to 85% of the reliable variance in the eight SF-36 scales in both patient and general populations in the United States [2,3] and in other countries.[4] Second, in cross-sectional and longitudinal tests, the SF-36 Physical Component Summary (PCS) has detected hypothesized differences in nearly all tests based on physical criteria (such as the severity of heart failure or the age-related decline in physical health) and the Mental Component Summary (MCS) has detected hypothesized differences 100% of the time in tests using mental criteria (such as the impact of clinical depression and change in severity over time).[2,5]

The results observed for the PCS and MCS measures show that it is possible to use psychometric methods to reduce the number of health dimensions assessed without substantial loss of information. More important, summary measures make it possible to construct an even shorter health survey because the number of items in a survey is a function of the number of health dimensions for which separate scores are to be estimated with precision.[5,6] Thus, in those applications where two summary scores (physical and mental health) are sufficient, a shorter survey may prove to be valid and practical enough for more wide-spread use.

This article documents the methods used to select and evaluate a subset of 12 items from the SF-36 Health Survey. Our objectives, which were achieved, were to develop a form that: 1) could be scored to explain at least 90% of the variance in SF-36 physical and mental health summary measures; 2) would reproduce the average scores for the summary measures and eight-scale profile with a high degree of comparability; and 3) could be printed on one to two pages of a self-administered questionnaire or administered by an interviewer in less than 2 minutes, on average. Also, we present here results from 2 studies of reliability and 20 tests of the empirical validity of the 12-item short-form (SF-12) health survey summary measures and 8-scale profile in comparison with SF-36 summary measures and scales.

## Methods

### Data Sources

Data for these studies came from two sources. The first source was the National Survey of Functional Health Status (NSFHS), a cross-sectional survey used to gather norms for the SF-36 Health Survey. Sampling methods and sample characteristics are well-documented elsewhere.[1,7] The NSFHS database was used to select and score 12 items from the SF-36 Health Survey. The NSFHS also was previously used to derive and to develop norm-based scoring for SF-36 PCS and MCS measures in the general population.[2,5]

The second source was the Medical Outcomes Study (MOS), an observational study of adult patients with chronic conditions.[8,9] The MOS data were used to cross-validate population-based scoring algorithms for summary measures and the eight-scale profile based on the SF-12 and to perform empirical tests of validity. Details on the design of the MOS are well-documented elsewhere.[8-11] Briefly, the MOS sampled patients with hypertension, congestive heart failure, survivors of a recent myocardial infarction, and Type II diabetes using criteria and sampling methods published elsewhere.[1,3,5,11] Those with depressive disorder, the fifth condition studied, were selected on the basis of a patient-completed form and results from a subsequent interview.[12]

### Health Status Measures

Results for two summary measures based on the SF-12 were compared with those derived from the SF-36 [2,5] as well as the eight-scale profile.[1,2] To simplify our presentation, we adopt the following conventions in labeling measures. In previous publications, the SF-36 PCS and MCS measures were abbreviated as PCS and MCS, respectively.[2,5] However, we label them as PCS-36 and MCS-36, here to indicate that the 36-item form was used to score them. The PCS and MCS based on the SF-12 are referred to here as PCS-12 and MCS-12. The construction and scoring of the eight SF-36 scales and PCS-36 and MCS-36, including tests of scaling assumptions, are summarized elsewhere.[1-3,5]

The two SF-12 summary measures were constructed independently to reproduce the SF-36 physical and mental summary measures. Forward-step regression analysis was used to identify a subset of 12 or fewer items from the SF-36 and 2 weighting algorithms for estimating PCS-36 and MCS-36. On the basis of previous experience, we were confident that a 12-item short-form printed on a single questionnaire page could be completed by the great majority of respondents and that a less compressed 2-page version would be satisfactory for virtually all respondents capable of self-administration. Further, from published estimates [13] of response times from several general population studies, we expected that the 12 items could be self-administered in 2 minutes or less by most respondents. The latter expectation

was confirmed in a small pilot test in which 26 of 32 adults (81.3%) completed the SF-12 in less than 2 minutes. A second objective in choosing items was the representation of the eight SF-36 health concepts (Fig. 1). Ten items were sufficient to reproduce both the PCS-36 and MCS-36 scores with an $R^2$ above 0.90. Two additional items were selected to represent all eight concepts.

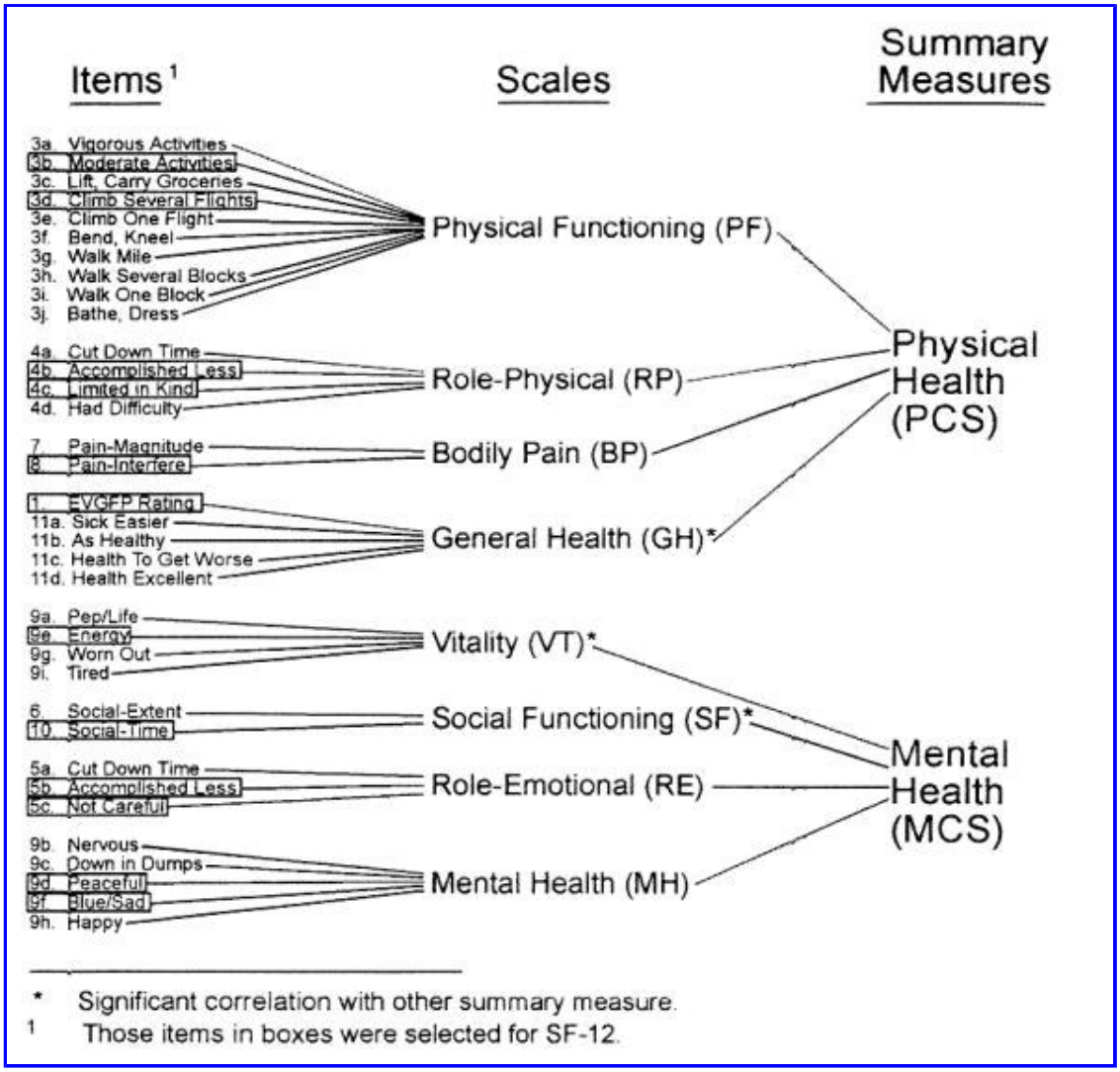



FIG. 1. Measurement model.

Two scoring strategies were compared: 1) equal-interval scoring of response categories for 10 items and recalibration of response choices for 2 items to better meet scaling assumptions (the standard SF-36 method)[1] and (2) unequal interval scoring with item weights for response categories empirically derived in the general US population. In the first method, weights were derived by estimating PCS-36 and MCS-36 scales using items scored according to the standard SF-36 scoring method. In the second method, item response categories were defined as "dummy" variables, which were used to estimate PCS-36 and MCS-36 scales in the general US population. As with the PCS-36 and MCS-36,[2,5] norm-based standardized scores were computed for the PCS-12 and MCS-12 scales to have means of 50 and standard deviations of 10 in the general US population.

## Reliability

Data from repeated administrations of the SF-36 2 weeks apart were analyzed to estimate the test-retest reliability of PCS-12 and MCS-12 scores. Test-retest reliability coefficients were estimated using productmoment correlations between scores in the only two English language datasets with test-retest administrations of the SF-36. These included a subset of the general US population participating in the NSFHS survey (n = 232)[7] and a sample from the UK general population (n = 187).[14]

## Empirical Validation

Tests of validity were designed to address issues involved in the many intended uses of the forms and study conditions that might affect interpretations. For example, the SF-12 measures of physical and mental health should discriminate between groups of patients who differ in physical and mental health according to proven clinical measures. This standard method of construct validation follows the logic of "known groups" validity.[15] The performance of PCS-12 and MCS-12 in discriminating between groups was compared with the SF-36 summary measures and eight scales. For these tests, patients from the MOS were categorized into four groups known to differ in physical and/or mental health as defined clinically. The criteria used to define these groups are identical to those reported previously for studies of SF-36 scales and summary measures.[2,3,5] Briefly, 10 categories of comparisons were performed, involving groups of patients differing in: 1) the seriousness of a physical condition (serious versus minor physical diagnosis); 2) the presence/severity of a mental condition (serious mental condition versus minor medical); 3) the incremental impact of a serious physical condition on a mental condition and the incremental impact of a serious mental condition on a serious physical condition; 4) specific physical diagnoses (four groups); 5) severity of hypertension (two levels), diabetes (four levels), and congestive heart failure (two levels); 6) the presence of 16 comorbid conditions; 7) the frequency of acute symptoms; 8) cross-sectional and longitudinal comparisons of age effects among the most well group of patients (uncomplicated hypertension); 9) longitudinal comparisons among groups of patients classified at 1-year follow-up according to self-reported changes in physical, mental, and general health status; and 10) cross-sectional and longitudinal (2-year) comparisons of patients with clinical depression.

We tested the same hypotheses for the SF-12 in relation to the above variables as were previously tested for the SF-36.[2,3,5] For example, we expected the PCS-12 would be most valid in distinguishing groups differing in the presence and severity of physical conditions and would perform less well than did the MCS-12 in distinguishing groups differing in the presence and severity of mental conditions. We hypothesized the reverse pattern of results for the MCS-12 relative to the PCS-12. Based on results with the SF-36, we expected the SF-12 to pass these tests of validity and that our hypotheses would be supported. We were particularly interested in using these tests to gauge the validity of the SF-12 relative to the SF-36.

## Analytic Plan

The analytic plan was identical to that used in previous studies so that results for the SF-12 could be unambiguously compared with results for the SF-36. The first three analyses of "criterion groups" were performed using analysis of variance methods.[1-3] Analyses of criterion variables in categories 4 to 6 used ordinary least squares multiple regression techniques with the same statistical adjustments for differences in age, gender, race, poverty, study site, health care setting, and season of the year used in previous MOS analyses.[2,5] Longitudinal analyses and other cross-sectional analyses for criterion variables in categories 7 to 10 used least squares regression methods but without adjustments for baseline patient characteristics to maintain comparability with previous tests.[2,5] All multivariate analyses of scales used multivariate analysis of variance to test whether any SF-36 or SF-12 scales differed across any of the groups being compared. For those tests that yielded a significant multivariate analysis of variance F-ratio, regression models were estimated to test the relative validity (RV) of each scale. Thus, RV was estimated only for those scales that met two statistical criteria: 1) significant overall multivariate analysis of variance F for the set of criterion variables (defining patient groups) in relation to all scales and 2) significant univariate analysis of variance F for the

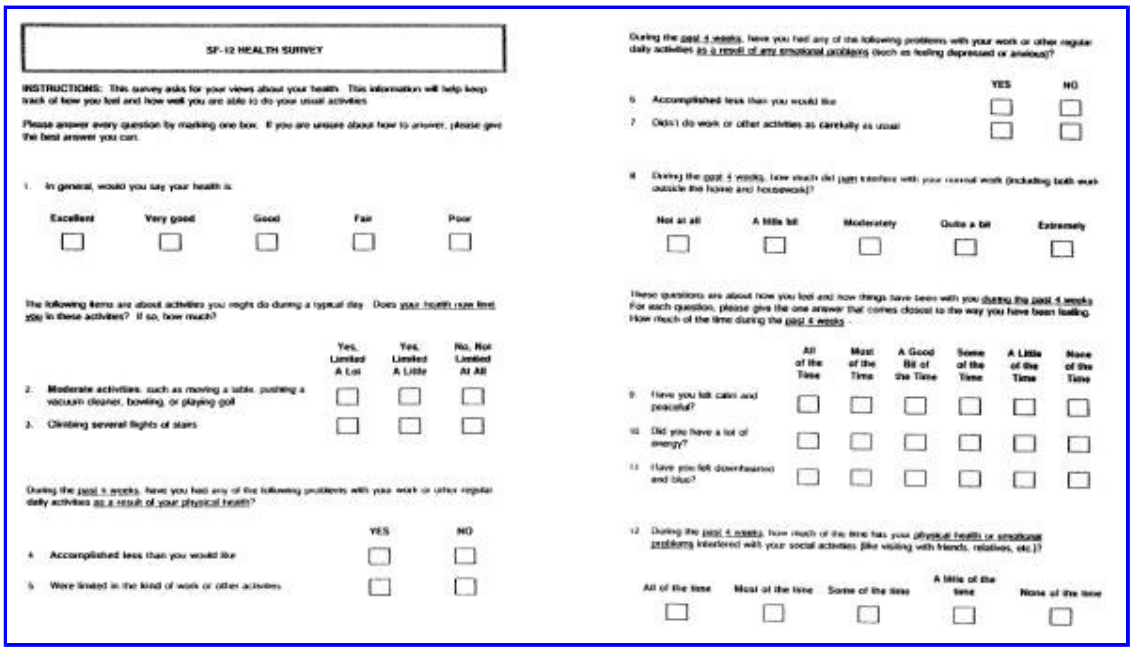
same set of criterion variables and the scale in question.

To estimate the validity of each SF-12 summary component measure relative to summary measures and scales based on the SF-36, ratios of F-statistics were compared [16,17] as in previous studies.[2,3,5] The F-statistic for each measure in each test is a ratio of the amount of separation in scores between groups or between assessments over time relative to the within group (error) variance. The F-statistic is larger when the average separation between groups or change over time is larger and/or the error variance is smaller. The RV estimate for each SF-12 measure in each test indicates, in proportional terms, its empirical validity relative to summary measures and scales based on the SF-36. When one measure performs exceptionally well, estimates (based on RV) sometimes are low to the point of being misleading.[2,5] Therefore, standardized estimates of effect size also were computed by dividing the average difference or change observed for each measure by the standard deviation for that measure estimated in the general US population, as in previous studies.[2,5]

## Results

### Construction of SF-12

The 12 items chosen (Figure 2) achieved a multiple $R^2$ of 0.911 in the prediction of PCS-36 and 0.918 in the prediction of MCS-36 in the general US population (n = 2,474). The SF-12 items that were the best predictors of PCS-36 were those from proven physical health scales (Physical Functioning, Role Physical, or Bodily Pain scales), whereas those items that best predicted the MCS-36 were those from proven mental health scales (Mental Health, Role Emotional, and Social Functioning scales), as would be expected.[19]


SF-12 HEALTH SURVEY

INSTRUCTIONS: 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.

Please answer every question by marking one box. If you are unsure about how to answer, please give the best answer you can.

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

Excellent | Very good | Good | Fair | Poor

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 A Lot | Yes, Limited A Little | No, Not Limited At All |
|---|---|---|---|
| 2. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf | | | |
| 3. Climbing several flights of stairs | | | |

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?

| | YES | NO |
|---|---|---|
| 4. Accomplished less than you would like | | |
| 5. Were limited in the kind of work or other activities | | |

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)?

| | YES | NO |
|---|---|---|
| 6. Accomplished less than you would like | | |
| 7. Didn't do work or other activities as carefully as usual | | |

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 | A little bit | Moderately | Quite a bit | Extremely

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 -

| | 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 Time |
|---|---|---|---|---|---|---|
| 9. Have you felt calm and peaceful? | | | | | | |
| 10. Did you have a lot of energy? | | | | | | |
| 11. Have you felt downhearted and blue? | | | | | | |

12. 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.)?

All of the time | Most of the time | Some of the time | A little of the time | None of the time

FIG. 2. SF-12 Health Survey, two-page format.

Significant improvements in $R^2$ were observed using the second scoring method (which weights response categories to better reflect the unequal intervals between them), in predictions of the two summary measures in the general US population, as documented elsewhere. Briefly, the $R^2$ for PCS-12 using the second scoring method represented an improvement from 0.842 to 0.911 ($P < 0.01$).[19] Similarly, the $R^2$ for MCS-12 using the second scoring method represented an improvement from 0.846 to 0.918 ($P < 0.01$).

Selected items and weights derived from the general US population were used to score PCS-12 and MCS-12 for purposes of cross-validation in the MOS sample (n = 2,293). The PCS-12 and MCS-12 scored using general population weights were very highly correlated with PCS-36 and MCS-36 in the MOS sample ($r$ = 0.951 and 0.969, respectively). These correlations represent $R^2$ values of 0.904 and 0.939, respectively. The PCS-12 and MCS-12 were very weakly correlated ($r$ = 0.06) with each other as in 39 other analyses of the SF-36 (0.01-0.07) documented to date.2 Thus, the independence of PCS and MCS scales scored using the SF-36 is maintained with the SF-12.

## Descriptive Statistics for SF-12 Scales

Table 1 lists the means and standard errors for SF-12 and SF-36 forms of the two summary scales and eight-scale profiles for the four-group comparison (criteria, 1-3). Summary scale scores based on SF-12 averaged within 1 point of summary scale scores based on the SF-36 across all groups. The great majority of mean scores for the eight scales estimated from SF-12 were within 3 points of those for SF-36; noteworthy exceptions were observed for the General Health scale (refer to Table 1).

| Scale | Comparison Groups | | | |
|---|---|---|---|---|
| | Minor Medical (n = 599) | Serious Physical (n = 126) | Mental Only (n = 131) | Serious Physical and Mental (n = 33) |
| Physical Summary (PCS) | | | | |
| SF-36 PCS | 46.60 ± 0.4 | 37.49 ± 1.0 | 49.26 ± 1.0 | 34.78 ± 1.6 |
| SF-12 PCS | 47.42 ± 0.4 | 38.75 ± 1.0 | 49.32 ± 0.9 | 36.34 ± 1.6 |
| Mental Summary (MCS) | | | | |
| SF-36 MCS | 54.29 ± 0.3 | 54.46 ± 0.8 | 36.37 ± 1.1 | 42.51 ± 1.8 |
| SF-12 MCS | 53.82 ± 0.3 | 52.51 ± 0.8 | 37.03 ± 1.1 | 43.18 ± 1.7 |
| Physical Functioning (PF) | | | | |
| SF-36 PF | 80.99 ± 0.9 | 59.67 ± 2.5 | 82.87 ± 1.6 | 49.86 ± 4.7 |
| SF-12 PF | 80.25 ± 0.9 | 59.82 ± 2.6 | 81.76 ± 1.7 | 53.10 ± 4.8 |
| Role Physical (RP) | | | | |
| SF-36 RP | 70.35 ± 1.4 | 43.25 ± 3.6 | 58.78 ± 3.4 | 24.24 ± 5.5 |
| SF-12 RP | 70.68 ± 1.4 | 45.86 ± 3.3 | 64.11 ± 3.1 | 29.99 ± 5.3 |
| Bodily Pain (BP) | | | | |
| SF-36 BP | 76.24 ± 0.9 | 67.14 ± 2.2 | 65.50 ± 2.1 | 51.73 ± 4.1 |
| SF-12 BP | 77.73 ± 0.8 | 68.23 ± 2.1 | 65.32 ± 2.1 | 55.03 ± 4.3 |
| General Health (GH) | | | | |
| SF-36 GH | 67.31 ± 0.7 | 49.96 ± 2.0 | 58.93 ± 2.0 | 41.48 ± 2.5 |
| SF-12 GH | 70.71 ± 0.5 | 58.23 ± 1.4 | 68.85 ± 1.3 | 53.67 ± 2.6 |
| Vitality (VT) | | | | |
| SF-36 VT | 62.06 ± 0.8 | 48.86 ± 1.9 | 45.37 ± 1.8 | 39.19 ± 3.7 |
| SF-12 VT | 60.25 ± 0.7 | 51.55 ± 1.7 | 49.15 ± 1.5 | 48.02 ± 3.2 |
| Social Functioning (SF) | | | | |
| SF-36 SF | 91.91 ± 0.6 | 80.46 ± 2.2 | 66.03 ± 2.2 | 66.29 ± 4.0 |
| SF-12 SF | 90.68 ± 0.6 | 82.86 ± 1.9 | 68.25 ± 2.1 | 66.98 ± 3.9 |
| Role Emotional (RE) | | | | |
| SF-36 RE | 84.56 ± 1.3 | 74.07 ± 3.4 | 40.97 ± 3.4 | 46.46 ± 7.1 |
| SF-12 RE | 84.08 ± 1.2 | 76.06 ± 3.1 | 44.71 ± 3.2 | 49.88 ± 6.7 |
| Mental Health (MH) | | | | |
| SF-36 MH | 82.23 ± 0.6 | 77.68 ± 1.3 | 52.80 ± 1.8 | 55.76 ± 3.5 |
| SF-12 MH | 81.40 ± 0.5 | 77.48 ± 1.3 | 54.17 ± 1.7 | 60.98 ± 3.0 |

MOS, Medical Outcomes Study.

TABLE 1. Comparison of Means ± Standard Errors for SF-36 and SF-12 Scales and Summary Measures, MOS Patients Differing in Physical and Mental Conditions

## Reliability

The test-retest reliability of the PCS-12 summary measures was 0.890 in the United States and 0.864 in the United Kingdom. Coefficients of 0.760 and 0.774 were observed for the MCS-12 in the United States and the United Kingdom, respectively. Although these reliability estimates for PCS-12 and MCS-12 are slightly below those for PCS-36 and MCS-36, they compare favorably with those for the eight SF-36 scales, which ranged from 0.63 to 0.89 (median, 0.80) in these studies.[1] They are also slightly higher than those for eight-scale scores for the SF-12, which ranged from 0.63 to 0.91 (median, 0.76). In addition, changes in scores between test and retest averaged less than 1 point for the two summary measures in both samples, and 85.3% scored at the second administration within the 95% confidence interval of the scores at the first administration for both PCS-12 and MCS-12.

## Four-Group Test Validity

Tables 2 and 3 list RV coefficients and effect size estimates from tests of the validity of SF-12 summary measures and scales in discriminating among groups known to differ in physical and mental conditions (criterion variables, 1-3). In the first two tests for physical differences (Table 2), the PCS-12 yielded RVs of 0.93 and 0.63 relative to the best SF-36 scale. These values are only slightly lower than the RVs observed for PCS-36 (0.97 and 0.72). Effect size estimates were 0.87 and 1.30 for PCS-12 and 0.91 and 1.45 for PCS-36 in these tests. As hypothesized, in tests of physical differences, MCS-12 yielded very low RV coefficients as did MCS-36. The eight scales scored from SF-12 had lower RV coefficients than did SF-36 versions, with few exceptions.

| | Serious Physical vs. Minor Medical | | | | Serious Physical and Mental vs. Mental Only | | | |
|---|---|---|---|---|---|---|---|---|
| Scale | Mean Difference | ES | F | RV | Mean Difference | ES | F | RV |
| Physical Component Summary (PCS) | | | | | | | | |
| SF-36 PCS | -9.11 | 0.91 | 89.30[a] | 0.97 | -14.48 | 1.45 | 48.02[a] | 0.72 |
| SF-12 PCS | -8.67 | 0.87 | 85.38[a] | 0.93 | -12.98 | 1.30 | 41.99[a] | 0.63 |
| Mental Component Summary (MCS) | | | | | | | | |
| SF-36 MCS | -1.83 | 0.18 | 5.43[c] | 0.06 | 6.14 | 0.61 | 6.50[c] | 0.10 |
| SF-12 MCS | -1.31 | 0.13 | 3.03 | 0.03 | 6.15 | 0.61 | 7.24[b] | 0.11 |
| Physical Functioning (PF) | | | | | | | | |
| SF-36 PF | -21.32 | 0.91 | 92.16[a] | 1.00 | -33.01 | 1.42 | 66.58[a] | 1.00 |
| SF-12 PF | -20.43 | 0.88 | 76.21[a] | 0.83 | -28.66 | 1.33 | 45.83[a] | 0.69 |
| Role Physical (RP) | | | | | | | | |
| SF-36 RP | -27.10 | 0.80 | 57.15[a] | 0.62 | -34.54 | 1.02 | 22.09[a] | 0.33 |
| SF-12 RP | -24.82 | 0.73 | 52.27[a] | 0.57 | -34.12 | 1.05 | 25.20[a] | 0.38 |
| Bodily Pain (BP) | | | | | | | | |
| SF-36 BP | -9.10 | 0.40 | 17.14[a] | 0.19 | -13.77 | 0.58 | 8.88[b] | 0.13 |
| SF-12 BP | -9.50 | 0.40 | 21.62[a] | 0.23 | -10.29 | 0.49 | 4.75[c] | 0.07 |
| General Health (GH) | | | | | | | | |
| SF-36 GH | -17.34 | 0.85 | 92.16[a] | 1.00 | -17.45 | 0.86 | 17.98[a] | 0.27 |
| SF-12 GH | -12.48 | 0.61 | 91.97[a] | 0.99 | -15.19 | 0.93 | 26.73[a] | 0.40 |
| Vitality (VT) | | | | | | | | |
| SF-36 VT | -13.20 | 0.63 | 44.62[a] | 0.48 | -6.18 | 0.29 | 2.34 | 0.03 |
| SF-12 VT | -8.70 | 0.41 | 25.91[a] | 0.28 | -1.13 | 0.06 | 0.11 | 0.00 |
| Social Functioning (SF) | | | | | | | | |
| SF-36 SF | -11.44 | 0.50 | 45.70[a] | 0.50 | 0.26 | 0.01 | 0.00 | 0.00 |
| SF-12 SF | -7.82 | 0.34 | 23.81[a] | 0.26 | -1.27 | 0.05 | 0.07 | 0.00 |
| Role Emotional (RE) | | | | | | | | |
| SF-36 RE | -10.48 | 0.32 | 11.22[a] | 0.12 | 5.49 | 0.17 | 0.50 | 0.00 |
| SF-12 RE | -8.02 | 0.24 | 7.18[b] | 0.08 | 5.17 | 0.16 | 0.50 | 0.00 |
| Mental Health (MH) | | | | | | | | |
| SF-36 MH | -4.55 | 0.25 | 10.43[b] | 0.11 | 2.96 | 0.16 | 0.56 | 0.00 |
| SF-12 MH | -3.62 | 0.20 | 9.00[b] | 0.10 | 6.81 | 0.41 | 3.28 | 0.05 |

MOS, Medical Outcomes Study; ES, effect size = mean difference/SD: where SD comes from the general US population; RV, relative validity.
[a] $P < 0.001$.
[b] $P < 0.01$.
[c] $P < 0.05$.

TABLE 2. Summary of Group Comparisons Involving Physical Conditions, MOS Patients

| Scale | Mental vs. Minor Medical | | | | Serious Mental and Physical vs. Serious Physical Only | | | |
|---|---|---|---|---|---|---|---|---|
| | Mean Difference | ES | F | RV | Mean Difference | ES | F | RV |
| Physical Component Summary (PCS) | | | | | | | | |
| SF-36 PCS | 2.66 | 0.27 | 8.01[b] | 0.02 | -2.71 | 0.27 | 1.56 | 0.03 |
| SF-12 PCS | 1.90 | 0.19 | 4.33[c] | 0.01 | -2.41 | 0.24 | 1.28 | 0.03 |
| Mental Component Summary (MCS) | | | | | | | | |
| SF-36 MCS | -17.92 | 1.79 | 433.06[a] | 1.12 | -9.95 | 0.99 | 29.48[a] | 0.62 |
| SF-12 MCS | -16.79 | 1.68 | 414.53[a] | 1.07 | -9.33 | 0.93 | 28.62[a] | 0.60 |
| Physical Functioning (PF) | | | | | | | | |
| SF-36 PF | 1.88 | 0.08 | 0.86 | 0.00 | -9.80 | 0.42 | 3.31 | 0.07 |
| SF-12 PF | 1.52 | 0.06 | 0.50 | 0.00 | -6.72 | 0.31 | 1.42 | 0.03 |
| Role Physical (RP) | | | | | | | | |
| SF-36 RP | -11.57 | 0.34 | 10.89[a] | 0.03 | -19.01 | 0.56 | 6.40[c] | 0.13 |
| SF-12 RP | -6.57 | 0.19 | 3.84[c] | 0.01 | -15.87 | 0.49 | 5.06[c] | 0.11 |
| Bodily Pain (BP) | | | | | | | | |
| SF-36 BP | 10.74 | 0.45 | 25.20[a] | 0.06 | -15.14 | 0.64 | 10.18[b] | 0.21 |
| SF-12 BP | 12.41 | 0.52 | 37.94 | 0.10 | -13.20 | 0.63 | 7.78[b] | 0.16 |
| General Health (GH) | | | | | | | | |
| SF-36 GH | -8.37 | 0.41 | 21.90[a] | 0.06 | -8.48 | 0.42 | 4.33[c] | 0.09 |
| SF-12 GH | -1.85 | 0.09 | 2.16 | 0.01 | -4.56 | 0.28 | 2.13 | 0.04 |
| Vitality (VT) | | | | | | | | |
| SF-36VT | -16.69 | 0.80 | 74.99[a] | 0.19 | -9.67 | 0.46 | 5.24[c] | 0.11 |
| SF-12VT | -11.10 | 0.53 | 45.43[a] | 0.12 | -3.53 | 0.20 | 0.92 | 0.02 |
| Social Functioning | | | | | | | | |
| SF-36 SF | -25.87 | 1.13 | 234.09[a] | 0.61 | -14.17 | 0.62 | 9.00[b] | 0.19 |
| SF-12 SF | -22.43 | 0.98 | 180.63[a] | 0.47 | -15.88 | 0.65 | 14.29[a] | 0.30 |
| Role Emotional (RE) | | | | | | | | |
| SF-36 RE | -43.58 | 1.32 | 196.28[a] | 0.51 | -27.61 | 0.84 | 13.18[a] | 0.28 |
| SF-12 RE | -39.37 | 1.19 | 172.66[a] | 0.45 | -26.18 | 0.82 | 14.29[a] | 0.30 |
| Mental Health (MH) | | | | | | | | |
| SF-36 MH | -29.43 | 1.63 | 386.51[a] | 1.00 | -21.93 | 1.21 | 47.61[a] | 1.00 |
| SF-12 MH | -27.23 | 1.51 | 375.97[a] | 0.97 | -16.50 | 0.99 | 30.91[a] | 0.65 |

MOS, Medical Outcomes Study; ES, effect size = mean difference/SD: where SD comes from the general US population; RV, relative validity.
[a] $P < 0.001$.
[b] $P < 0.01$.
[c] $P < 0.05$.

TABLE 3. Summary of Group Comparisons Involving Mental Conditions, MOS Patients

Table 3 lists results from the two validity tests for differences in mental health. The RV coefficients of 1.07 and 0.60 were observed for MCS-12, relative to the best SF-36 scale; effect size was 1.68 and 0.93 in these tests. These results are comparable with those observed for MCS-36 (RV = 1.12 and 0.62). As hypothesized, PCS-12 yielded very low RV estimates in tests for mental health differences. The same three SF-12 scales (Mental Health, Role Emotional, Social Functioning) were most valid in these tests as in previous SF-36 studies, although SF-12 scales yielded lower RV coefficients than did SF-36 scales.

## Other Tests' Validity

Table 4 summarizes 192 RV coefficients for SF-12 and SF-36 forms of the summary measures and scales across 16

tests of validity. Detailed results from these tests are reported elsewhere.[19] The first 12 columns include criterion variables defining differences in physical health, and the last 3 columns are tests for differences in mental health (an exception is the 9th column, GI/GU symptom cluster, shown previously to impact most on mental health).[5] Statistical conclusions based on PCS-12 agreed 10 of 12 times with the 3 best SF-36 physical scales (Physical Functioning, Role Physical, Bodily Pain) in comparisons involving physical health criteria. The RV coefficients for PCS-12 ranged from 0.43 to 0.78, and the median equals 0.67 in these tests. The PCS-36 performed better than did PCS-12 in all but one of these tests.

| | Severity of Disease | | | | | Symptom Clusters | | | | Age Differences | | Self-Reported Change | | | Clinical Depression | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Measures | Chronic Conditions | Hypertension | Diabetes | CHF | Comorbid Conditions | Ear, Nose and Throat | Central Nervous System | Musculoskeletal | GI/GU[b] | Cross-sectional | Longitudinal | Physical | General | Mental | Cross-sectional | Longitudinal |
| PF | 0.76[c] | | | 1.00[c] | 0.58[c] | 1.00[c] | 0.72[c] | 0.28[c] | 0.39[c] | 1.00[c] | 0.34[c] | 1.00[c] | 0.79[c] | 0.16[c] | | |
| RP | 0.36[c] | 1.00[c] | | 0.84[c] | 0.53[c] | 0.32[c] | 0.39[c] | 0.26[c] | | 0.31[c] | | 0.36[c] | 0.38[c] | 0.11[c] | 0.05[c] | 0.19[d] |
| BP | 0.12[c] | | | | 1.00[c] | | 0.21[c] | 1.00[c] | 0.39[c] | 0.14[d] | | 0.20[c] | 0.21[c] | 0.08[c] | 0.06[c] | |
| GH | 1.00[c] | | | 0.74[c] | 0.70[c] | 0.42[d] | 0.37[c] | 0.11[c] | | | 1.00[c] | 0.72[c] | 1.00[c] | 0.38[c] | 0.07[c] | 0.15[c] |
| VT | 0.33[c] | | | 0.72[c] | 0.48[c] | | 1.00[c] | 0.13[c] | | | 0.74[c] | 0.29[c] | 0.54[c] | 0.38[c] | 0.23[c] | 0.80[c] |
| SF | 0.28[c] | | | | 0.33[c] | 0.28[c] | 0.47[c] | 0.08[c] | 0.66[c] | | | 0.45[c] | 0.61[c] | 0.51[c] | 0.59[c] | 0.52[c] |
| RE | 0.12[c] | | | 0.57[d] | 0.16[d] | | 0.34[c] | 0.03[c] | 0.39[c] | | | 0.11[c] | 0.08[d] | 0.46[c] | 0.42[c] | 0.77[c] |
| MH | 0.11[c] | | | 0.21[c] | 0.25[c] | 0.97[c] | 0.83[c] | 0.01[c] | 1.00[c] | 0.18[c] | | 0.20[c] | 0.41[c] | 1.00[c] | 1.00[c] | 1.00[c] |
| Manova F 8 Scales | 6.61[d] | 2.17[c] | NS | 3.42[c] | 5.14[c] | 3.57[c] | 9.96[c] | 3.07[c] | 28.14[c] | 7.65[c] | 2.60[d] | 9.35[c] | 12.91[c] | 10.97[c] | 165.85[c] | 4.89[c] |
| SF36 PCS | 0.59[c] | | | 0.68[c] | 0.94[c] | 0.78[c] | 0.60[c] | 0.55[c] | | 0.71[c] | | 0.81[c] | 0.78[c] | 0.06[c] | 0.01[c] | |
| SF36 MCS | | | | 0.27[c] | 0.14[c] | 0.43[d] | 0.82[c] | 0.03[c] | 0.92[c] | 0.33[c] | | 0.12[c] | 0.24[c] | 1.06[c] | 1.03[c] | 1.38[c] |
| SF12 PCS | 0.65[c] | | | 0.58[c] | 0.77[c] | 0.67[c] | 0.51[c] | 0.42[c] | | 0.78[c] | | 0.73[c] | 0.70[c] | 0.08[d] | | |
| SF12 MCS | | | | 0.39[c] | 0.14[c] | | 0.67[c] | 0.03[c] | 0.98[c] | 0.29[c] | | 0.11[d] | 0.21[c] | 0.93[c] | 0.98[c] | 0.97[c] |

CHF, congestive heart failure; GI/GU, gastrointestinal/genitourinary; NS, not significant; PF, Physical Functioning; RP, Role Physical; BP, Bodily Pain; GH, General Health; VT, Vitality; SF, Social Functioning; RE, Role Emotional; MH, Mental Health.
[a]For each test, the "best" of the eight SF-36 scales (with highest F ratio) is labeled RV = 1.00.
[b]Considered mental health criteria for purposes of summarizing trends in results.
[c]$P < 0.001$.
[d]$P < 0.01$.
[e]$P < 0.05$.

TABLE 4. Summary of Relative Validity Coefficients[a] for SF-36 Profiles, SF-36 Summary Measures, and SF-12 Summary Measures; Sixteen Comparisons Used to Test Validity

Differences in either or both of the two SF-36 "general" scales (General Health and Vitality) were significant in 12 of the 16 tests, including 4 with an RV equaling 1.00. The PCS-12 differences were significant in 9 of the 12 tests (range of RV coefficients from 0.08 to 0.77; median, 0.65); the MCS-12 also detected 9 of 12 significant differences (range of RV coefficients from 0.03 to 0.98; median, 0.39). With few exceptions, RV coefficients for PCS-12 and MCS-12 were lower than those for PCS-36 and MCS-36 (medians of 0.65 and 0.39 for SF-12 versus medians of 0.78 and 0.27 for SF-36 summary measures).

For all four comparisons among groups differing in the presence and severity of mental health conditions that yielded significant differences for any of the three best mental health scales (Mental Health, Role Emotional, Social Functioning), conclusions based on MCS-12 always agreed with those based on MCS-36 (refer to the three right-most tests and GI/GU columns in Table 4). The RV coefficients for MCS-12 ranged from 0.93 to 0.98 across the four tests and were below those for MCS-36 in all but one test (GI/GU).

## Discussion

The SF-12 Health Survey represents another step in the "downsizing" of measures from the MOS. These 12 questionnaire items and norm-based SF-12 scoring algorithms appear to accomplish three objectives: 1) reproduction of more than 90% of the variance in SF-36 PCS and MCS measures in the general US population and on cross-validation in the MOS; 2) accurate reproduction of average scores for both SF-36 summary measures, but less accurately for the eight-scale profile; and 3) reduction in length sufficient to print the form on one to two questionnaire

pages and sufficient for self-administration in 2 minutes or less.

The challenge in constructing a short-form measure is one of balancing the number of questionnaire items against other important considerations, such as the comprehensiveness of content and the statistical precision of scores. The two scales (Physical Functioning, Role Physical) that best predict physical health and the two scales (Role Emotional, Mental Health) that best predict mental health are the only scales reproduced in the SF-12 using two items each, owing to their proven usefulness and to the lack of precision of estimates of these concepts based on a single item.[21] The remaining four scales (Bodily Pain, General Health, Vitality, and Social Functioning) are estimated from only one item each. The 12 items provide a representative sampling of the content of the 8 health concepts and the various operational definitions of those concepts, including what respondents are able to do, how they feel, and how they evaluate their health status.[9]

Whereas the original MOS scales were constructed to be highly, internally consistent, items for the SF-36, and even more so items for the SF-12, were selected with heterogeneity in mind. Each selected item has unique reliable variance that is of proven value in prediction. In the case of the SF-12, we have chosen to predict summary measures for two clusters of highly related SF-36 scales. By pooling the reliable variance in physical and mental health across measures, we have been able to maintain satisfactory reliability while reducing the number of items. By summarizing measures shown empirically to produce the same result, the SF-12 and SF-36 summary measures simplify the analysis of health data while minimizing information loss.[2,5]

Although PCS-12 and MCS-12 always reached the same statistical conclusions about group differences as did PCS-36 and MCS-36, they did so with relative validity coefficients that were typically 10% below those observed for the SF-36. The SF-12 versions define fewer levels and pool less reliable variance and should, therefore, be expected to yield less reliable assignments of individuals to those levels. However, for large group studies (eg, n = 500), the differences in measurement reliability between SF-12 and SF-36 are not as important, because confidence intervals around group averages are determined largely by the sample size. Therefore, this tradeoff between precision and questionnaire length is likely to prove worthwhile for purposes of monitoring general and specific populations based on large sample sizes.

A remarkably high degree of correspondence was achieved in reproducing the SF-36 PCS and MCS measures using SF-12 items. Correlations between SF-12 and SF-36 versions of PCS and MCS were 0.951 and 0.969, respectively, on cross-validation, and estimates of group means were consistently within 1 point. This high degree of correspondence between means for SF-36 and SF-12 summary measures has also been demonstrated for general population groups differing in age and gender.[19] These results suggest that norms and other interpretation guidelines published for the SF-36 summary measures will be useful in interpreting SF-12. These guidelines include cross-sectional norms and norms for 1-year change scores in general and specific populations, content-related interpretation guidelines, as well as criterion-based guidelines for the two summary measures.[2,5] Criterion-based interpretation guidelines include results from prospective predictions of inpatient and outpatient utilization of health care services, subsequent job loss due to health problems among employed adults, 5-year survival probabilities at various levels of scale scores, and cutoff scores for screening for psychiatric and physical conditions. However, large differences in mean scores were observed for SF-12 and SF-36 versions of the eight-scale profile, suggesting that SF-36 norms for the eight scales should be used cautiously in interpreting SF-12, pending further evaluation of scoring algorithms that might increase their comparability.

The fact that the SF-12 is entirely a subset of the SF-36 will greatly increase its usefulness in comparing results across studies that use either form. The SF-12 also was constructed to maintain comparability with other widely used MOS short-form measures, including four of the six global items in the MOS Short-Form Health Survey,[9,19] which is widely used by academic medical centers and in general population surveys. (Physical and role functioning items in that survey are not included in either the SF-36 or SF-12.) Finally, the SF-12 includes five of the six items used in scoring a

health utility index based on the SF-36, which is forth-coming from the International Quality of Life Assessment Project.[4,20]

Translations of the SF-12 are available from the International Quality of Life Assessment Project in the five non-English languages most widely spoken in the United States (Spanish, French, German, Italian, and Japanese).[22] Translations in Chinese, Korean, and Vietnamese, three of the fastest growing non-English speaking populations in the United States, are currently being evaluated.[22] In total, translations and adaptations of the SF-12 in 30 language/country combinations are available or forthcoming from the International Quality of Life Assessment Project.[4]

Equal-interval (linear) scoring algorithms proved satisfactory for all but two of the items in SF-36 studies to date.[4] However, the information value of each item is even more important when there are fewer items, leading us to look for other potential gains. Scoring based on weighted item response categories increased the variance explained in both summary measures by more than 7% and yielded mean scores that more closely approximated those based on the SF-36, which was our second objective. Therefore, to maintain maximum comparability with the interpretation guidelines for SF-36 versions of PCS and MCS, the more complicated unequal interval scoring was adopted for SF-12 in tests of validity reported here. To facilitate easy and accurate estimations of these scores, a computer diskette with scoring algorithms, a test data set, and written documentation are included in the SF-12 Scoring Manual, available at cost from The Health Institute, New England Medical Center.[19]

It should be noted that conclusions about the SF-12 were based on analyses of items interspersed within the SF-36. We assume that the same or better results will be obtained when the SF-12 items are administered alone. In support of this assumption, results from tests of scaling assumptions and conclusions about reliability for the SF-36, when it was embedded with other items measuring the same concepts, were replicated when it was administered alone.[1] Tests of these assumptions are forthcoming from numerous studies of general and specific populations that fielded the SF-12 without the other SF-36 items.

Results from this study are encouraging about the feasibility of further downsizing short-form surveys for purposes of monitoring the health of both general and specific populations. The PCS-12 and MCS-12 reached the same statistical conclusions about hypothesized group differences as did the PCS-36 and MCS-36, respectively. Thus, the SF-12 represents a plausible alternative to the SF-36 for measuring health status. Questionnaire length was reduced by two thirds with minimal loss in measurement precision. This difference between 12 and 36 questionnaire items is important because it may determine whether health status is measured in some large-scale studies. For example, the new "report cards" that will be based on the Member Health Care Survey required for accreditation by the National Committee for Quality Assurance include the SF-12, whereas the SF-36 was deemed too long and too costly to administer on a large scale.[18]

In choosing between forms, it is important to consider that the SF-36 defines more levels of health and better represents the content of health measures than does the SF-12. Consequently, SF-36 summary measures, particularly the eight-scale SF-36 profile, yield more reliable estimates of individual levels of health, giving the SF-36 a decided advantage over the SF-12 in smaller studies. Therefore, the choice of the SF-12 over the SF-36 is most justified in studies with large sample sizes having severe constraints on questionnaire length and in studies focusing on patient-based assessments of physical and mental health.

## References

1. Ware JE, Snow KK, Kosinski M, et al. SF-36 health survey: Manual and interpretation guide. Boston, MA: The Health Institute, New England Medical Center, 1993. [Context Link]

2. Ware JE, Kosinski M, Keller SD. SF-36 physical and mental health summary scales: A user's manual. Boston, MA: The Health

Institute, New England Medical Center, 1994. [Context Link]

3. McHorney CA, Ware JE, Raczek AE. The MOS 36-item short-form health status survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247. **ExternalResolverBasic** | **Bibliographic Links** | **Library Holdings** | [Context Link]

4. Ware JE, Keller SD, Gandek B, et al. Evaluating translations of health status surveys: Lessons from the IQOLA project, International Journal of Technology Assessment in Health Care, 1995;11:525. [Context Link]

5. Ware JE, Kosinski M, Bayliss MS, et al. Comparison of methods for the scoring and statistical analysis of SF-36 health profiles and summary measures: Results from the medical outcomes study. Med Care 1995;33(4):AS264. [Context Link]

6. Nunnally JC, Bernstein IR. Psychometric theory. 3rd ed. New York, NY: McGraw-Hill, 1994. [Context Link]

7. McHorney CA, Kosinski M, Ware JE. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: Results from a national survey. Med Care 1994;32:551. **ExternalResolverBasic** | **Bibliographic Links** | **Library Holdings** | [Context Link]

8. Tarlov AR, Ware JE, Greenfield S, et al. The medical outcomes study: An application of methods for monitoring the results of medical care. Journal of the American Medical Association 1989;262:925. [Context Link]

9. Stewart AL, Ware JE. Measures for a new era of health assessment. In: Stewart AL, Ware JE, eds. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Durham, NC: Duke University Press, 1992. [Context Link]

10. Stewart AL, Greenfield S, Hays RD, et al. Functional status and well-being of patients with chronic conditions: Results from the medical outcomes study. JAMA 1989;262:907. **ExternalResolverBasic** | **Bibliographic Links** | **Library Holdings** | [Context Link]

11. Kravitz RL, Greenfield S, Rogers W, et al. Differences in the mix of patients among medical specialties and systems of care: Results from the medical outcomes study. JAMA 1992;267:1617. **ExternalResolverBasic** | **Bibliographic Links** | **Library Holdings** | [Context Link]

12. Wells KB, Hays RD, Burnam MA, et al. Detection of depressive disorder for patients receiving prepaid or fee-for-service care: Results from the medical outcomes study. JAMA 1989;262:3298. **ExternalResolverBasic** | **Bibliographic Links** | **Library Holdings** | [Context Link]

13. Ware JE, Karmos AH. Scales for measuring general health perceptions. Health Services Research 1976;11:396. [Context Link]

14. Brazier JE, Harper R, Jones NMB, et al. Validating the SF-36 health survey questionnaire: New outcome measure for primary care. Br Med J 1992;305:160. [Context Link]

15. Kerlinger FN. Foundations of behavioral research. New York, NY: Holt, Rinehart and Winston, 1964. [Context Link]

16. Winer BJ, Brown DR, Michels KM. Statistical principles in experimental design. 3rd ed. New York, NY: McGraw-Hill, Inc., 1991. [Context Link]

17. Liang MH, Larson MG, Cullen KE, et al. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542. **ExternalResolverBasic** | **Bibliographic Links** | **Library Holdings** | [Context Link]

18. NCQA, Annual Member Health Care Survey Manual. Washington, DC: National Committee for Quality Assurance, June 1995. [Context Link]

19. Ware JE, Kosinski M, Keller SD. SF-12: How to score the SF-12 physical and mental health summary scales. 2nd ed. Boston, MA: The Health Institute, New England Medical Center, 1995. [Context Link]

20. Brazier J, Usherwood T, Harper R, et al. Deriving a preference-based single index measure from the UK SF-36 health survey (in preparation). [Context Link]

21. McHorney CA, Ware JE, Rogers WH, et al. 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):MS253. **ExternalResolverBasic** | **Bibliographic Links** | **Library Holdings** | [Context Link]

22. The World Almanac and Book of Facts, 1995. New Jersey: Funk & Wagnalls, 1994. [Context Link]

Key words: health survey; SF-12; SF-36; Medical Outcomes Study; health status assessment; health-related quality of life

---

*Accession Number: 00005650-199603000-00003*

---