Psychometric Evaluation of the SF-36® Health Survey in Medicare Managed Care

Barbara Gandek, M.S., Samuel J. Sinclair, M.Ed., Mark Kosinski, M.A., and John E. Ware, Jr., Ph.D.

Data quality and scoring assumptions for the SF-36® Health Survey were evaluated among the elderly and disabled, using 1998 Cohort I baseline Medicare HOS data (n=177,714). Missing data rates were low, and scoring assumptions were met. Internal consistency reliability was 0.83 to 0.93 for the eight scales and 0.94 and 0.89, respectively, for the physical (PCS) and mental (MCS) component summary measures. Results declined with increased risk factors (e.g., older age, more chronic conditions), but were well above accepted standards for all subgroups. These findings support using standard algorithms for scoring the SF-36® in the HOS and subgroup analyses of HOS data.

INTRODUCTION

The HOS is a longitudinal evaluation of the physical and mental health outcomes of beneficiaries enrolled in MMC plans nationwide. Part of the effectiveness of care component of the HEDIS®, the HOS uses the SF-36® Health Survey as its primary outcomes measure (National Committee for Quality Assurance, 2003). This survey is widely used in monitoring population health, evaluating treatment outcomes, estimating disease burden, and monitoring outcomes in clinical practice. It is the subject of more than 4,000 published studies, with thousands of studies including the elderly, and hundreds of studies limited to the elderly (Turner-Bowker, Bartley, and Ware, 2002).

Inclusion of the SF-36® in the HOS is based on the assumption that data meet minimum psychometric requirements in the Medicare population. The psychometric properties of its scales among the elderly have been examined in a number of studies. Generally, these studies have concluded that the scales are suitable for use among community-dwelling adults age 65 or over. However, several researchers have noted that interview administration may be needed in some elderly populations, and that the scales may have limitations when used with the frail elderly (Hill, Harries, and Popay, 1996). Some studies also have noted higher missing data rates for some items, including limitations in vigorous activities and work/other daily activities.

While these studies generally have demonstrated that the methods used to construct and score scales and summary measures are appropriate for the elderly, small sample sizes have precluded subgroup analyses by characteristics such as ethnicity or educational level. Due to the size of the HOS, we were able to compare data quality and the psychometric performance of the scales and summary measures across multiple groups differing in

Barbara Gandek is with the Health Assessment Lab and Tufts University School of Medicine. Samuel J. Sinclair is with the Health Assessment Lab. Mark Kosinski is with QualityMetric Incorporated and the Health Assessment Lab. John E. Ware, Jr. is with the Health Assessment Lab, QualityMetric Incorporated, Tufts University School of Medicine, and Harvard School of Public Health. The research in this article was supported by the Centers for Medicare & Medicaid Services (CMS) under Contract Number 500-00-0055 with the National Committee for Quality Assurance (NCQA). The views expressed in this article are those of the authors and do not necessarily reflect the views of the Health Assessment Lab, Tufts University School of Medicine, QualityMetric Incorporated, Harvard School of Public Health, NCQA, or CMS. SF-36® is a registered trademark of the Medical Outcomes Trust, which makes the survey available royalty-free to individuals and organizations for academic research.

sociodemographic and clinical characteristics. Thus, in addition to allowing for evaluation of the consistency of psychometric results across diverse groups in the HOS, these results also can serve as reference data for specific subgroups.

METHODS

Data

The HOS has been described in detail elsewhere (Haffer et al., 2003; National Committee for Quality Assurance, 2003). In brief, it is an ongoing longitudinal study that began in 1998. One thousand beneficiaries from each MMC contract market are randomly selected for the HOS annually. Beneficiaries must be continuously enrolled in their plan for at least 6 months prior to sampling. Medicare disabled beneficiaries are included; ESRD patients are excluded. Sampled beneficiaries receive an advance letter from CMS, followed by the questionnaire 1 week later. A second mailing is made to non-respondents 1 month later. Beneficiaries who do not respond to either mailing are contacted up to six times to complete a telephone interview. National Committee for Quality Assurance-certified vendors collect the data, using standard mailing materials and telephone script. Beneficiaries who completed the baseline survey and remained in the same managed care plan are resurveyed at the 2-year followup.

Data reported in this article were collected in 1998 for the Cohort I baseline HOS survey, in which a total of 279,135 beneficiaries were sampled. Beneficiaries primarily were enrolled in M+C HMOs, although a small percent were in continuing cost or demonstration plans. The response rate for the Cohort I baseline survey was 64 percent. Respondents to the survey were slightly younger (mean age=73.1 for respondents versus 73.5 for non-respondents, $p<0.0001$); slightly more likely to be female (56.9 versus 56.5 percent, $p<0.05$); and more likely to be white (88.0 versus 82.6 percent, $p<0.001$).

Health Status Measure

The SF-36® Health Survey is the primary health outcomes measure in the HOS. It contains multi-item scales measuring eight generic health concepts: physical functioning (PF), role limitations due to physical health problems (RP), bodily pain (BP), general health perceptions (GH), vitality (VT), social functioning (SF), role limitations due to emotional problems (RE), and mental health (MH) (Ware and Sherbourne, 1992; Ware et al., 1993). A single-item measure of comparative health (HT) is also included. More information on the development and evaluation of SF-36® scales can be found in (Ware, 2000a).

Summary measures of physical and mental health, PCS and MCS, are calculated from the eight scales using algorithms recommended by the developers (Ware and Kosinski, 2001a). The PCS and MCS were constructed to simplify and improve the analysis of health outcomes by: reducing the number of variables analyzed without much loss of information; measuring across a wider range of score levels than the scales; increasing the reliability of scores by pooling common reliable variance across scales; and improving the validity of scores in discriminating between physical and mental health outcomes by constructing orthogonal component summary scores (Ware et al., 1995). PCS and MCS are the main outcomes measures in the HOS.

Analysis

Assumptions underlying the scoring and construction of the eight scales and two summary measures were evaluated. If assumptions are met, these measures can be scored using standard algorithms (Ware, 1993; 2001a).

Data Quality

A scale score can be more confidently estimated where there are no missing data. Data quality was evaluated by examining the percent of respondents with missing data for each item; the percent with complete data for each scale; and the percent for whom scale scores and summary measures could be calculated using recommended SF-36® scoring algorithms, as in previous studies (McHorney et al., 1994).

Tests of Scaling Assumptions and Scale Properties

A number of scaling assumptions were tested. First, while classical scaling criteria for summated rating scales suggest that item means should be roughly equivalent within a scale, heterogeneity of item content may make it appropriate for means to be non-equivalent in practice (Edwards, 1957). Item means are expected to vary within scales. However, the rough difference between them should be reproducible across samples, if items are consistently spaced along the health continuum (Ware et al., 1993). Hypotheses about the relative position of items were examined, based on item content and previous studies (Gandek and Ware, 1998; McHorney, 1994). All items are scored so a higher value indicates a better health state.

Second, the Likert (1932) method assumes that item standard deviations are roughly equivalent within a scale; items should be standardized if variances vary greatly. In practice, most items in widely used summated rating scales satisfy this standard and there is little to be gained if items are standardized (Ware et al., 1997). However, this assumption was evaluated by visually examining item standard deviations.

Three additional scaling assumptions were tested using a correlation matrix of items and scales. First, each item was examined to see if it was substantially linearly related to the scale score computed from all other items in its hypothesized scale (test of item internal consistency). Item internal consistency generally is considered satisfactory if an item correlates 0.40 or more with its hypothesized scale, after correction for overlap (i.e., correlation with a scale score computed from all other items in that scale) (Howard and Forehand, 1962).

Second, item-scale correlations also were examined to determine if they were approximately equal within a scale, to enable aggregation without weighting. When all items contribute substantially to the total score, this test has been considered satisfied, even if item-scale correlations vary. Only rarely would unequal weighting across items improve the performance of a scale enough to justify the added complexity associated with item weighting within the Likert scaling framework (Ware et al., 1997).

Finally, in addition to demonstrating that an item is measuring what it is supposed to measure, it is important that an item not be a strong measure of other concepts (test of item discriminant validity). Following the logic of the multitrait/multimethod approach (Campbell and Fiske, 1959), the SF-36® was constructed to achieve item discriminant validity as manifested by a significantly ($p<0.05$) higher correlation between each item and its hypothesized scale than

with scales measuring other concepts; this is labeled a definite scaling success (Ware et al., 1997). As in previous studies, the scaling success rate is the ratio of the number of definite scaling successes relative to the total number of item scaling tests for each scale (McHorney et al., 1994).

Internal consistency reliability was estimated using Cronbach's (1951) coefficient alpha. Reliability of measurement refers to the extent to which the measured variance in a score reflects true score, rather than random error. A minimum reliability coefficient of 0.70 has been suggested for group-level analyses (Nunnally and Berstein, 1994).

The percentage of respondents achieving either the highest score (ceiling) or lowest score (floor) also was evaluated. If a high proportion of respondents score at either the ceiling or floor, the ability of the scale to detect change over time in that group is limited.

Scales and summary measures were scored for the psychometric analysis as recommended by the developers (Ware and Kosinski (2001a)), except that the first bodily pain (BP1) and general health (GH1) items were not recalibrated, and the second bodily pain item (BP2) was not rescored. In addition, improved algorithms recommended by the developers were used to score the PCS and MCS, which increased the number of respondents for whom scores could be computed, and reduced score estimation and sampling bias (Ware, 2000a). In brief, this missing data estimation (MDE) approach uses item response theory to score the PF scale, and regression techniques to score the summary measures when scores are not available for all eight scales, as required by the original scoring algorithms.

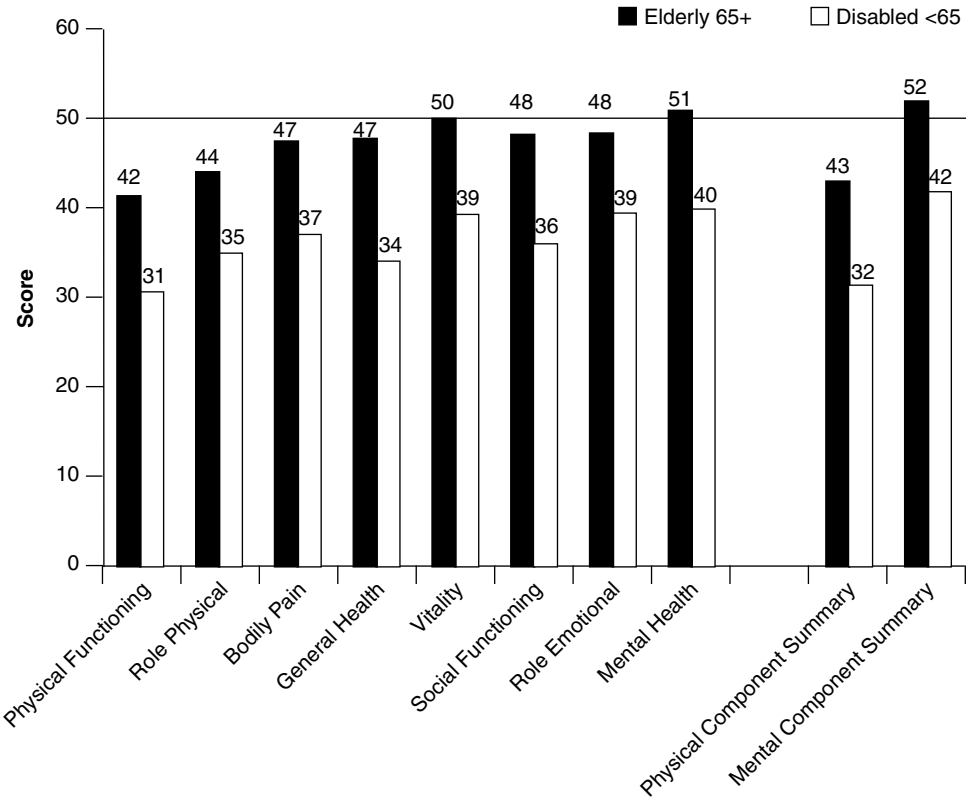
Data quality was analyzed for all respondents who answered any part of the HOS questionnaire. Tests of scaling assumptions and factor analyses used data from respondents for whom all eight scale scores could be calculated. Tests of scaling assumptions used the MAP-R for Windows software (Ware et al., 1997). The exact content of the items and response choices is reproduced in (Ware and Sherbourne, 1992).

Factor Analysis

Two distinct higher-order physical and mental health factors have accounted for 80-85 percent of the reliable variance in the eight scales in the U.S. general population (Ware and Kosinski, 2001a) and among Medical Outcome Study (MOS) patients (McHorney, Ware, and Raczek, 1993). Based on these results and tests of the clinical interpretation of the two factors, psychometrically-based PCS and MCS health summary measures have been constructed by summing the eight-scale scores, after standardizing and applying PCS- and MCS-specific weights to each scale (Ware et al., 1995; Ware and Kosinski, 2001a). A linear transformation is then used to norm the measures, such that a value of 50 is the U.S. general population (1998) mean and 10 is the standard deviation. Norm-based scoring of all eight scales and the summary measures, as recommended by the developers (Ware and Kosinski, 2001a), has the advantage of easily facilitating interpretation of results across measures, as all measures have comparable means and standard deviations. This is illustrated in Figure 1, which presents norm-based results for the Cohort I elderly (age 65 or over) and disabled (under age 65). As can be seen from the figure, disabled beneficiaries scored well below the general population norm on all measures. The elderly scored below national norms on measures that primarily measure physical health (e.g., PCS, PF, RP), but at or near the norms on scales that primarily measure mental health (e.g., MCS, MH, RE).

Figure 1

SF-36® Scores for Elderly and Disabled Cohort I HOS Respondents at Baseline: 1998



NOTE: All summary measures use norm-based scoring (mean=50, standard deviation=10).

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

The same methods of factor extraction and rotation were used to test the appropriateness of the standard PCS and MCS scoring algorithms in the HOS. Two principal components were extracted from the correlations among the scales; components were rotated to orthogonal simple structure using the varimax method to facilitate comparisons with published results and for ease of interpretation. Criteria commonly used to evaluate factor analyses using the principal components method were routinely applied (Harman, 1976). The pattern of correlations between the eight scales and the two rotated components was examined to determine the basis for their interpretation as physical and mental components.

Based on previous studies (McHorney, Ware, and Raczek, 1993; Gandek and Ware, 1998; Ware and Kosinski, 2001a), we hypothesized that: extraction of two components would be supported; more than 60 percent of the total and 80 percent of the reliable variance across scales would be explained by the two components; and more than 50 percent of the total and 70 percent of the reliable variance within each scale would be explained by the two components. The PF scale was hypothesized to correlate highest (lowest) with the physical (mental) component, followed by RP and BP. The MH scale was hypothesized to correlate highest (lowest) with the mental (physical) component, followed by RE and

SF. The GH and VT scales were hypothesized to correlate moderately with both components.

Subgroup Analyses

Data were examined separately for self and proxy administrations. For beneficiaries who completed the survey by self-administration, data also were examined by age, sex, race, education, income, and survey mode. Additional analyses were conducted on subsets of beneficiaries who self-administered the form but might be expected to find the items more difficult to answer. These included beneficiaries on Medicaid or who were disabled (age 18-64), had 3 or more self-reported chronic conditions (out of 17), screened positive for likelihood of depression, had vision problems (mail survey), or had hearing problems (telephone survey). In addition, beneficiaries deemed most likely to have problems answering the survey (less than a 12th grade education, three or more chronic conditions, and a positive screen for likelihood of depression) were examined separately. All respondents completed the survey in English. Chinese and Spanish translations of the HOS were available starting in Cohort I (Chinese) and Cohort II (Spanish). Data on the psychometric performance of Chinese and Spanish translations in the elderly are available elsewhere (Ren et al., 1998; Health Assessment Lab, 2000).

Based on previous studies (McHorney et al., 1994), we expected worse results for beneficiaries who were older, had less education, or were living in poverty. We also expected slightly higher rates of missing data for both role functioning scales (Gandek and Ware, 1998; McHorney et al., 1994). As in other general population samples, floor and ceiling effects were expected to be minimal for the three scales measuring both disability and well-being (GH, VT, MH) (Ware et al., 1993). Due to the relative coarseness of the role scales in Version 1.0, notable floor and ceiling effects were expected for the RP and RE scales. Ceiling effects also were expected for the SF scale. We expected lower ceiling effects and higher floor effects with increased age and among beneficiaries with multiple medical conditions or functional limitations, or who screened positive for likelihood of depression. We did not expect to see substantial differences in the principal components results by age or sex, but did expect to see differences for Asians (Ren et al., 1998).

RESULTS

Data Quality

Twenty-four percent of respondents did not answer one or more of the 36 items. However, the percent of missing data for each item was relatively low overall (median=3.1 percent), ranging from 0.7 to 5.6 percent (Table 1). As expected, missing data rates were somewhat higher for RP and RE items (range=4.5 to 5.6 percent). As noted by Gandek and Ware (1998), and McHorney et al. (1994), PF questions which were ordered as a Guttman scale[1] (i.e., climbing stairs (PF4-PF5) and walking various distances (PF7-PF9)) generally showed higher missing data rates for the last item(s) in the sequence.

Overall, the percent of respondents who answered all items within each scale ranged from approximately 95 percent for the role functioning scales to 97-98 percent for all other scales (Table 2). Data completeness rates were slightly lower for proxy respondents. Among beneficiaries who completed the survey by self-administration, data completeness declined slightly

[1] A Guttman scale is a cumulative scale in which each item consistently increases in extremity or severity.

Table 1

Percent Missing, Means, Standard Deviations, and Correlations Between SF-36® Items and Hypothesized Scales: Medicare HOS, 1998

Item	Abbreviated Content	Percent Missing	Mean	Standard Deviation	Item-Scale Correlation								
					PF	RP	BP	GH	VT	SF	RE	MH	HT
PF1	Vigorous Activities	3.3	1.66	0.72	*0.51	0.46	0.43	0.45	0.47	0.36	0.24	0.25	0.19
PF2	Moderate Activities	2.1	2.24	0.77	*0.81	0.59	0.54	0.58	0.57	0.55	0.38	0.38	0.26
PF3	Lift, Carry Groceries	1.6	2.39	0.73	*0.79	0.55	0.51	0.55	0.53	0.55	0.39	0.38	0.26
PF4	Climb Several Flights	2.4	2.05	0.80	*0.80	0.55	0.50	0.54	0.56	0.48	0.34	0.34	0.24
PF5	Climb One Flight	2.8	2.43	0.73	*0.82	0.54	0.50	0.54	0.53	0.53	0.37	0.36	0.26
PF6	Bend, Kneel	1.6	2.14	0.75	*0.73	0.53	0.54	0.49	0.51	0.47	0.34	0.32	0.23
PF7	Walk Mile	2.6	2.02	0.85	*0.79	0.55	0.50	0.52	0.54	0.47	0.33	0.32	0.23
PF8	Walk Several Blocks	2.5	2.26	0.83	*0.85	0.56	0.51	0.54	0.55	0.52	0.36	0.34	0.25
PF9	Walk One Block	2.8	2.54	0.70	*0.78	0.50	0.47	0.51	0.50	0.52	0.36	0.34	0.26
PF10	Bathe, Dress	1.1	2.75	0.55	*0.53	0.36	0.34	0.41	0.37	0.46	0.33	0.32	0.22
RP1	Cut Down Time	4.5	1.65	0.48	0.55	*0.74	0.55	0.52	0.53	0.59	0.50	0.38	0.29
RP2	Accomplish Less	4.6	1.51	0.50	0.56	*0.78	0.55	0.53	0.58	0.55	0.51	0.38	0.26
RP3	Limited in Kind	5.6	1.55	0.50	0.62	*0.82	0.58	0.55	0.58	0.57	0.47	0.37	0.26
RP4	Had Difficulty	4.8	1.56	0.50	0.60	*0.81	0.59	0.56	0.60	0.58	0.49	0.40	0.28
BP1	Pain Severity	2.6	4.07	1.37	0.54	0.56	*0.79	0.55	0.57	0.54	0.37	0.43	0.30
BP2	Pain Limitations	3.1	3.83	1.19	0.63	0.66	*0.79	0.61	0.63	0.66	0.47	0.48	0.33
GH1	General Health	0.7	3.02	0.98	0.61	0.55	0.55	*0.73	0.64	0.56	0.41	0.47	0.37
GH2	Sick Easier	3.2	4.14	1.06	0.44	0.43	0.44	*0.57	0.50	0.52	0.41	0.50	0.22
GH3	As Healthy as Others	3.3	3.47	1.25	0.50	0.46	0.46	*0.68	0.56	0.50	0.35	0.43	0.28
GH4	Health to Get Worse	3.4	3.40	1.15	0.38	0.36	0.36	*0.50	0.45	0.37	0.27	0.35	0.29
GH5	Health Excellent	3.2	3.22	1.31	0.59	0.56	0.56	*0.77	0.66	0.58	0.41	0.49	0.34
VT1	Full of Pep	3.3	3.41	1.39	0.58	0.57	0.55	0.64	*0.74	0.57	0.41	0.50	0.32
VT2	Lot of Energy	3.4	3.43	1.45	0.59	0.58	0.55	0.66	*0.76	0.58	0.43	0.54	0.32
VT3	Worn Out	3.7	4.34	1.32	0.51	0.52	0.52	0.58	*0.71	0.57	0.42	0.55	0.29
VT4	Tired	2.9	3.96	1.27	0.52	0.52	0.53	0.59	*0.75	0.56	0.42	0.52	0.29
SF1	Social—Extent	2.8	4.11	1.17	0.59	0.63	0.60	0.61	0.62	*0.74	0.57	0.58	0.35
SF2	Social—Frequency	3.1	4.13	1.15	0.57	0.58	0.57	0.61	0.64	*0.74	0.53	0.61	0.32
RE1	Cut Down Time	4.5	1.78	0.42	0.39	0.50	0.40	0.44	0.45	0.55	*0.78	0.55	0.22
RE2	Accomplish Less	4.6	1.69	0.46	0.38	0.52	0.40	0.43	0.46	0.52	*0.77	0.52	0.20
RE3	Not Careful	5.4	1.77	0.42	0.39	0.48	0.39	0.42	0.43	0.52	*0.74	0.50	0.22
MH1	Nervous	2.9	4.96	1.25	0.29	0.30	0.33	0.40	0.41	0.43	0.42	*0.62	0.16
MH2	Down in Dumps	3.1	5.35	1.10	0.35	0.35	0.38	0.45	0.48	0.56	0.52	*0.71	0.22
MH3	Calm and Peaceful	3.6	4.16	1.36	0.36	0.38	0.42	0.49	0.54	0.50	0.45	*0.64	0.23
MH4	Downhearted/Blue	3.3	5.08	1.13	0.33	0.34	0.37	0.44	0.50	0.53	0.50	*0.71	0.21
MH5	Happy	3.0	4.48	1.24	0.31	0.32	0.35	0.46	0.48	0.46	0.39	*0.61	0.22
HT	Change in Health	0.7	2.99	0.76	0.30	0.31	0.33	0.39	0.36	0.36	0.24	0.27	*0.00

*Corrected for item-scale overlap.

NOTES: *N*=177,714. PF is physical functioning. RP is role limitations due to physical health problems. BP is bodily pain. GH is general health. VT is vitality. SF is social functioning. RE is role limitations due to emotional problems. MH is mental health. HT is comparative health. Items GH1, BP1 and BP2 were not recalibrated prior to analysis.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

with increased age and lower education and income. There were no systematic differences by sex or race/ethnicity. Data completeness rates for the role functioning scales were lower among beneficiaries with vision problems, or low education plus multiple health conditions. Nevertheless, 93-94 percent of beneficiaries in these two groups answered all role functioning items.

Scales are scored as long as respondents answer at least one-half of the items within a scale, using a person-specific value to impute for missing data within a scale. After imputing values for missing data, 92.2 percent of respondents had scores for all eight scales. Role functioning scales could be calculated for 95-96 percent of respondents, while scores for all other scales could be calculated for 97-99 percent

Table 2

Percent Complete Items in Each SF-36® Scale, by Characteristics: Medicare HOS, 1998

Characteristic	PF	RP	BP	GH	VT	SF	RE	MH
Total Sample	97.7	95.1	97.2	97.2	96.7	97.1	95.2	96.8
Self/Proxy Completion	96.0-98.0	91.7-95.8	94.7-97.6	96.4-97.6	94.6-97.1	96.2-97.4	91.8-95.9	94.5-97.2
Self-Administered Forms Only (Range Across Groups)								
Age (4 Groups)	95.6-98.7	91.2-96.9	95.6-98.0	95.3-98.1	94.5-97.8	95.3-97.9	91.9-97.0	94.8-97.8
Sex (2 Groups)	97.8-98.4	95.5-96.1	97.5-97.6	97.3-98.0	97.0-97.2	97.4-97.4	95.7-96.2	97.2-97.3
Race/Ethnicity (5 Groups)	96.5-98.2	94.0-96.0	97.2-97.7	95.5-97.9	96.6-97.2	96.8-97.5	94.1-96.1	96.7-97.4
Education (5 Groups)	96.4-98.7	94.0-97.1	97.2-97.9	96.5-98.4	96.5-97.6	96.9-97.8	94.0-97.3	96.5-97.8
Income (5 Groups)	97.1-98.8	94.1-97.1	97.3-97.9	97.4-98.8	96.6-97.7	97.0-97.9	94.4-97.4	96.8-97.8
Survey Mode (Mail/Telephone)	97.4-98.1	95.6-97.2	97.5-97.6	94.0-98.1	97.1-97.1	97.0-97.5	95.7-97.2	97.2-97.5
Other Characteristics								
On Medicaid	97.1	93.9	96.8	96.6	96.3	96.7	94.1	96.6
Disabled Entitlement	98.7	95.5	97.8	97.9	97.8	97.7	94.9	97.9
3+ Chronic Conditions[1]	98.2	95.6	97.6	97.7	97.3	97.5	95.7	97.4
Depression Screener[2]	98.0	95.0	97.3	97.2	96.9	97.1	94.9	96.9
Chronic Ill/Low Education[3]	97.3	94.1	97.3	96.5	96.8	97.1	93.7	96.8
Vision Problems (Mail)	97.1	93.0	96.6	97.2	95.9	96.2	93.1	96.0
Hearing Problems (Telephone)	97.5	97.0	97.5	94.3	97.1	97.3	96.7	97.6

[1] Had 3+ self-report conditions (hypertension, angina, congestive heart failure, myocardial infarction, other heart disease, stroke, respiratory disease, gastrointestinal disorder, arthritis hip/knee, arthritis hand/wrist, sciatica, diabetes, cancer).

[2] Responded yes to at least one of two items about feelings of depression in past year.

[3] Had 3+ chronic conditions, screened positive for possibility of current depression, and <12 years of education.

NOTES: *N*=177,714. PF is physical functioning. RP is role limitations due to physical health problems. BP is bodily pain. GH is general health. VT is vitality. SF is social functioning. RE is role limitations due to emotional problems. MH is mental health.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

(Table 3). PCS and MCS scores could be calculated for approximately 97 percent of respondents using the MDE scoring method and 92.2 percent of respondents using the original algorithms (which require that all eight scales be scored to calculate a summary score). Patterns within various subgroups were similar to those seen for data completeness tests, with somewhat lower rates for proxy respondents, beneficiaries who were age 85 or over, had vision problems, or had low education plus multiple health conditions.

Tests of Scaling Assumptions

Item Means and Standard Deviations

Hypothesized patterns of differences in item means were observed (Table 1). Within the PF scale, the most difficult item (vigorous activities) had the lowest mean (1.66; 1=limited a lot, 3=not limited) and the easiest item (bathing and dressing) had the highest. Item means increased as item difficulty decreased across groups of PF items ordered as Guttman scales (PF4-PF5, PF7-PF9). As expected, VT items that measured energy (VT1-VT2) had lower mean values than items measuring fatigue (VT3-VT4), because well-being items define a higher level of health. Similarly, MH items measuring positive affect (MH3 and MH5) had lower mean values than items measuring negative affect (MH1, MH2, and MH4). Item standard deviations were roughly equivalent within scales, except for the bathing/dressing item.

Item Internal Consistency

All item-hypothesized scale correlations were greater than 0.40 and thus met the test of item internal consistency (Table 1). In general, correlations between items and their hypothesized scales were roughly equivalent within each scale, with some exceptions that have been seen in previous

Table 3

Percent of Respondents for Whom Scales and Summary Measures Were Computable[1], by Characteristics: Medicare HOS, 1998

Characteristic	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS
Total Sample	99.0	96.0	97.8	97.4	97.6	97.9	95.4	97.5	96.7	96.9
Self/Proxy Completion	98.2-99.2	92.9-96.6	96.7-98.1	96.8-97.8	96.0-97.9	97.5-98.1	92.6-96.1	95.6-97.8	93.7-97.2	94.4-97.4
Self-Administered Forms Only (Range Across Groups)										
Age (4 Groups)	98.1-99.5	93.2-97.5	96.7-98.6	95.9-98.2	96.1-98.6	96.8-98.7	92.3-97.1	95.9-98.4	94.9-97.8	95.2-98.0
Sex (2 Groups)	99.2-99.2	96.5-96.7	97.9-98.2	97.6-98.1	97.8-98.0	98.0-98.2	95.9-96.4	97.8-97.9	97.1-97.4	97.3-97.5
Race/Ethnicity (5 Groups)	98.4-99.3	95.6-97.2	97.6-98.7	95.4-98.0	97.5-98.2	97.7-98.6	94.7-96.2	97.2-98.3	96.7-97.9	96.7-97.9
Education (5 Groups)	98.4-99.5	95.4-97.5	98.1-98.2	96.9-98.4	97.8-98.2	98.0-98.3	94.5-97.4	97.6-98.1	96.6-97.7	96.9-97.9
Income (5 Groups)	98.9-99.6	95.5-97.5	98.0-98.2	97.8-98.9	97.8-98.1	98.0-98.3	94.8-97.5	97.7-98.1	96.7-97.7	97.0-97.9
Survey Mode (Mail/Telephone)	99.2-99.2	96.4-98.4	98.0-98.6	93.6-98.4	97.9-98.2	98.1-98.5	95.9-97.8	97.8-98.2	97.1-98.2	97.3-98.3
Other										
On Medicaid	98.7	95.4	97.8	96.9	97.7	98.2	94.7	97.6	96.6	96.8
Disabled Entitlement	99.4	96.3	98.6	98.0	98.6	98.7	95.3	98.4	97.0	97.2
3+ Chronic Conditions[2]	99.3	96.6	98.2	97.9	98.1	98.3	95.9	98.0	97.4	97.6
Depression Screener[3]	99.2	96.2	98.0	97.5	97.9	98.1	95.3	97.7	97.0	97.1
Chronic Ill/Low Education[4]	98.8	95.6	98.2	97.0	98.0	98.3	94.3	97.8	96.5	96.8
Vision Problems (Mail)	98.7	94.1	97.4	97.6	97.3	97.4	93.3	97.0	95.6	95.8
Hearing Problems (Telephone)	99.2	98.3	99.0	93.9	98.5	99.1	97.5	98.3	98.4	98.5

[1] Scores were computed using the half-scale rule for the eight scales and missing data estimation algorithms for summary measures.

[2] Had 3+ self-report conditions (hypertension, angina, congestive heart failure, myocardial infarction, other heart disease, stroke, respiratory disease, gastrointestinal disorder, arthritis hip/knee, arthritis hand/wrist, sciatica, diabetes, cancer).

[3] Responded yes to at least one of two items about feelings of depression in past year.

[4] Had 3+ chronic conditions, screened positive for possibility of current depression, and <12 years of education.

NOTES: *N*=177,714. PF is physical functioning. RP is role limitations due to physical health problems. BP is bodily pain. GH is general health. VT is vitality. SF is social functioning. RE is role limitations due to emotional problems. MH is mental health. PCS is Physical Component Summary. MCS is Mental Component Summary.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

studies (Gandek and Ware, 1998; McHorney et al., 1994). Item-scale correlations for items at the ends of the physical functioning spectrum (PF1, PF10) were lower than correlations for other PF items, as expected; internal consistency tests generally favor items with a mean close to the average (Nunnally and Bernstein, 1994). Also, hypothesized item-scale correlations for the GH items measuring resistance to illness (GH2) and health outlook (GH4) generally were lower than item-scale correlations for the other three GH items, which measure current health.

Item-scale correlations generally were greater than 0.40 and roughly equivalent within a scale (with the exceptions as previously noted) for different subgroups (Table 4). Item-scale correlations were below 0.40 for one general health item (GH4-expect health to get worse) for the hearing impaired group and the group with low education plus multiple health conditions, and for telephone respondents. The item-scale correlation for PF10 (bathing and dressing) also was 0.38 for the subgroup with low education plus multiple health conditions. However, overall scaling assumptions were met within subgroups.

Item Discriminant Validity

Items discriminated well within the total sample, with 100 percent scaling success rates (indicating that items had significantly higher correlations with their hypothesized scales than with other scales) across all scales (data available on request from the authors). Within subgroups, scaling success rates were 100 percent in 257 out of 272 tests (34 subgroups times 8 scales). Success rates for the GH scale ranged from 92.5 to 97.5 percent for three groups (Black, Vision Problems, Hearing Problems). Scaling success rates ranged from 96.9 to 98.8 percent for the PF, VT and MH scales among Asian beneficiaries, due to high correlations among some MH and VT items, and a high correlation between the vigorous activities item (PF1) and other scales. The scaling success rate for MH among Hispanics also was 97.5 percent, again due to high correlations between some MH and VT items. Rates also were lower for some scales in the Proxy-Friend (range 87.5 to 100 percent; median=98 percent) and Proxy-Caregiver (range 37.5 to 100 percent; median=96 percent) groups.

Internal Consistency Reliability

Internal consistency reliability estimates were greater than 0.70, the minimum standard for group comparisons, for all scales and summary measures, across all subgroups (Table 5). Reliability of the PCS and MCS was 0.94 and 0.89 for the total sample, respectively. Internal consistency reliability statistics generally did not vary across age, sex, education, or income groups, with few exceptions. Reliability was lowest in the group with low education plus multiple health conditions, but even within this group, values ranged from 0.73 to 0.91, and thus exceeded the 0.70 minimum recommended for group comparisons.

Floor and Ceiling Effects

As hypothesized, floor and ceiling effects were generally low for the three bipolar scales (GH, VT, MH), ranging from 0.2 to 6.9 percent in the total sample (Table 6). Floor and ceiling effects were modest for the PF and BP scales (floor=2.6 and 1.5 percent, ceiling=8.8 and 18.6 percent, respectively). The RP and RE scales had substantial floor and ceiling effects (floor=29.4 and 16.7 percent, ceiling=43.2 and 65.8 percent, respectively), while the SF scale had a notable ceiling effect (46.8 percent). For the role functioning scales,

Table 4
Range of Item-Scale Correlations, by Characteristics: Medicare HOS, 1998

Characteristic	*N*	PF	RP	BP	GH	VT	SF	RE	MH
Total Sample	163,840	0.53-0.85	0.74-0.82	0.79-0.79	0.50-0.77	0.71-0.76	0.74-0.74	0.74-0.78	0.61-0.71
Survey Completed By									
Self-Adminstration	139,965	0.50-0.84	0.74-0.82	0.79-0.79	0.49-0.77	0.71-0.77	0.73-0.73	0.73-0.76	0.60-0.70
Proxy—Family Member	15,559	0.58-0.86	0.74-0.84	0.80-0.80	0.53-0.77	0.70-0.72	0.74-0.74	0.80-0.85	0.61-0.74
Proxy—Friend	919	0.46-0.85	0.74-0.82	0.79-0.79	0.52-0.78	0.57-0.68	0.67-0.67	0.79-0.83	0.55-0.76
Proxy—Caregiver	478	0.58-0.86	0.78-0.85	0.80-0.80	0.45-0.69	0.60-0.62	0.59-0.59	0.74-0.80	0.51-0.72
Self-Administered Forms Only									
Age									
18-64 Years	7,663	0.42-0.80	0.65-0.76	0.81-0.81	0.47-0.70	0.63-0.71	0.70-0.70	0.75-0.82	0.66-0.77
65-74 Years	81,079	0.48-0.82	0.73-0.81	0.77-0.77	0.47-0.76	0.71-0.77	0.72-0.72	0.71-0.75	0.59-0.68
75-84 Years	43,601	0.46-0.81	0.71-0.79	0.77-0.77	0.44-0.68	0.67-0.72	0.69-0.69	0.71-0.74	0.55-0.66
85 Years or Over	7,622	0.46-0.79	0.70-0.79	0.76-0.76	0.40-0.67	0.62-0.66	0.67-0.67	0.71-0.73	0.51-0.64
Sex									
Male	59,866	0.55-0.84	0.73-0.82	0.78-0.78	0.50-0.77	0.72-0.77	0.73-0.73	0.72-0.76	0.60-0.69
Female	80,099	0.47-0.83	0.74-0.81	0.79-0.79	0.47-0.75	0.70-0.75	0.73-0.73	0.73-0.76	0.59-0.70
Race/Ethnicity									
White (Non-Hispanic)	121,762	0.50-0.84	0.73-0.81	0.79-0.79	0.49-0.77	0.72-0.77	0.75-0.75	0.73-0.76	0.61-0.70
Black (Non-Hispanic)	8,253	0.45-0.80	0.75-0.81	0.76-0.76	0.45-0.69	0.63-0.69	0.65-0.65	0.68-0.76	0.56-0.69
Hispanic	5,565	0.50-0.82	0.78-0.84	0.80-0.80	0.53-0.74	0.63-0.70	0.71-0.71	0.74-0.81	0.56-0.70
Asian/Pacific Islander	1,949	0.44-0.80	0.78-0.84	0.78-0.78	0.52-0.73	0.55-0.67	0.66-0.66	0.76-0.82	0.49-0.66
Other	2,433	0.51-0.83	0.74-0.82	0.78-0.78	0.46-0.72	0.65-0.71	0.67-0.67	0.71-0.75	0.57-0.69
Education									
8th Grade or <	12,057	0.45-0.81	0.74-0.82	0.78-0.78	0.48-0.70	0.62-0.69	0.64-0.64	0.71-0.77	0.54-0.68
Some High School	23,758	0.47-0.82	0.74-0.82	0.78-0.78	0.48-0.73	0.69-0.72	0.70-0.70	0.73-0.77	0.55-0.68
High School Graduate	49,767	0.49-0.83	0.73-0.82	0.79-0.79	0.49-0.76	0.71-0.76	0.74-0.74	0.73-0.76	0.60-0.69
Some College	31,265	0.52-0.83	0.72-0.81	0.79-0.79	0.49-0.77	0.72-0.78	0.77-0.77	0.70-0.75	0.59-0.70
College Graduate	20,981	0.52-0.82	0.71-0.79	0.76-0.76	0.44-0.76	0.70-0.79	0.76-0.76	0.69-0.73	0.58-0.68
Income									
<$10,000	17,673	0.46-0.82	0.74-0.81	0.79-0.79	0.49-0.73	0.67-0.71	0.70-0.70	0.73-0.77	0.57-0.71
$10,000-$19,999	34,288	0.49-0.82	0.73-0.82	0.79-0.79	0.50-0.76	0.71-0.75	0.73-0.73	0.72-0.76	0.61-0.70
$20,000-$29,999	25,228	0.50-0.82	0.72-0.81	0.79-0.79	0.49-0.76	0.71-0.76	0.74-0.74	0.72-0.76	0.60-0.68
$30,000-$39,999	15,693	0.49-0.82	0.71-0.80	0.78-0.78	0.48-0.76	0.70-0.78	0.75-0.75	0.69-0.74	0.58-0.66
$40,000+	20,531	0.50-0.81	0.70-0.80	0.76-0.76	0.45-0.75	0.71-0.79	0.74-0.74	0.68-0.72	0.56-0.66
Other									
On Medicaid	3,506	0.43-0.81	0.71-0.80	0.79-0.79	0.50-0.72	0.62-0.71	0.68-0.68	0.71-0.78	0.57-0.71
Disabled Entitlement	8,155	0.42-0.80	0.65-0.76	0.81-0.81	0.47-0.70	0.63-0.71	0.70-0.70	0.75-0.82	0.66-0.77
3+ Chronic Conditions[1]	60,991	0.46-0.81	0.70-0.78	0.76-0.76	0.43-0.71	0.68-0.71	0.72-0.72	0.73-0.77	0.60-0.72
Depression Screener[2]	30,702	0.49-0.82	0.70-0.79	0.79-0.79	0.43-0.72	0.62-0.67	0.67-0.67	0.67-0.73	0.51-0.67
Chronic Ill/Low Education[3]	6,677	0.38-0.78	0.65-0.73	0.72-0.72	0.37-0.60	0.46-0.58	0.57-0.57	0.64-0.71	0.43-0.63
Vision Problems (Mail)	4,456	0.50-0.83	0.73-0.80	0.80-0.80	0.48-0.74	0.64-0.69	0.73-0.73	0.72-0.78	0.54-0.72
Hearing Problems (Telephone)	1,751	0.43-0.78	0.67-0.74	0.72-0.72	0.35-0.63	0.58-0.65	0.61-0.61	0.71-0.75	0.52-0.70
Survey Mode									
Mail	123,516	0.50-0.84	0.74-0.82	0.80-0.80	0.50-0.78	0.72-0.78	0.76-0.76	0.73-0.76	0.61-0.70
Telephone	16,444	0.48-0.81	0.70-0.79	0.72-0.72	0.38-0.68	0.64-0.69	0.58-0.58	0.71-0.76	0.56-0.69

[1] Had 3+ self-report conditions (hypertension, angina, congestive heart failure, myocardial infarction, other heart disease, stroke, respiratory disease, gastrointestinal disorder, arthritis hip/knee, arthritis hand/wrist, sciatica, diabetes, cancer).

[2] Responded yes to at least one of two items about feelings of depression in past year.

[3] Had 3+ chronic conditions, screened positive for possibility of current depression, and <12 years of education.

NOTES: *N*=177,714. PF is physical functioning. RP is role limitations due to physical health problems. BP is bodily pain. GH is general health. VT is vitality. SF is social functioning. RE is role limitations due to emotional problems. MH is mental health.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

ceiling effects generally declined (and floor effects increased) with increased age (among the elderly), lower education, and lower income. Floor effects were greater for those who were disabled, or who had low education and multiple medical conditions.

Table 5

Internal Consistency Reliability Estimates for Scales and Summary Measures, by Characteristics: Medicare HOS, 1998

Characteristic	*N*	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS
Total Sample	163,840	0.93	0.91	0.88	0.83	0.87	0.85	0.88	0.83	0.94	0.89
Survey Completed By											
Self-Adminstration	139,965	0.93	0.90	0.87	0.82	0.87	0.84	0.87	0.83	0.94	0.89
Proxy—Family Member	15,559	0.95	0.92	0.89	0.84	0.86	0.85	0.92	0.86	0.95	0.92
Proxy—Friend	919	0.94	0.91	0.88	0.84	0.81	0.80	0.91	0.83	0.94	0.91
Proxy—Caregiver	478	0.95	0.92	0.89	0.80	0.80	0.74	0.88	0.82	0.94	0.89
Self-Administered Forms Only											
Age											
18-64 Years	7,663	0.92	0.87	0.89	0.80	0.84	0.82	0.90	0.88	0.93	0.93
65-74 Years	81,079	0.92	0.90	0.86	0.81	0.87	0.84	0.86	0.82	0.94	0.88
75-84 Years	43,601	0.92	0.89	0.86	0.80	0.86	0.82	0.85	0.80	0.93	0.87
85 Years or Over	7,622	0.92	0.89	0.86	0.76	0.82	0.80	0.85	0.79	0.92	0.86
Sex											
Male	59,866	0.93	0.91	0.87	0.83	0.88	0.84	0.87	0.83	0.94	0.89
Female	80,099	0.93	0.90	0.88	0.82	0.87	0.84	0.87	0.83	0.94	0.89
Race/Ethnicity											
White (Non-Hispanic)	121,762	0.93	0.90	0.87	0.82	0.88	0.85	0.86	0.83	0.94	0.88
Black (Non-Hispanic)	8,253	0.92	0.90	0.86	0.80	0.84	0.79	0.86	0.83	0.93	0.88
Hispanic	5,565	0.93	0.92	0.88	0.83	0.84	0.83	0.89	0.83	0.94	0.90
Asian/Pacific Islander	1,949	0.92	0.91	0.86	0.82	0.80	0.80	0.90	0.78	0.94	0.89
Other	2,433	0.93	0.90	0.87	0.81	0.85	0.80	0.86	0.83	0.94	0.88
Education											
8th Grade or <	12,057	0.93	0.90	0.87	0.80	0.83	0.78	0.86	0.82	0.93	0.89
Some High School	23,758	0.93	0.91	0.87	0.81	0.86	0.82	0.87	0.82	0.94	0.89
High School Graduate	49,767	0.93	0.90	0.87	0.83	0.88	0.85	0.87	0.83	0.94	0.89
Some College	31,265	0.93	0.90	0.88	0.83	0.89	0.87	0.85	0.84	0.94	0.89
College Graduate	20,981	0.92	0.89	0.86	0.82	0.89	0.86	0.84	0.83	0.94	0.89
Income											
<$10,000	17,673	0.93	0.90	0.88	0.82	0.85	0.82	0.87	0.84	0.93	0.87
$10,000-$19,999	34,288	0.93	0.90	0.88	0.83	0.87	0.85	0.87	0.84	0.94	0.88
$20,000-$29,999	25,228	0.93	0.90	0.87	0.83	0.88	0.85	0.86	0.83	0.94	0.89
$30,000-$39,999	15,693	0.92	0.89	0.87	0.82	0.88	0.86	0.85	0.82	0.94	0.89
$40,000+	20,531	0.92	0.89	0.85	0.81	0.89	0.85	0.83	0.81	0.94	0.88
Other											
On Medicaid	3,506	0.92	0.89	0.88	0.82	0.84	0.81	0.87	0.84	0.93	0.90
Disabled Entitlement	8,155	0.92	0.87	0.89	0.80	0.84	0.82	0.90	0.88	0.93	0.93
3+ Chronic Conditions[1]	60,991	0.92	0.89	0.87	0.80	0.85	0.83	0.87	0.84	0.93	0.89
Depression Screener[2]	30,702	0.93	0.89	0.88	0.81	0.82	0.80	0.84	0.80	0.93	0.87
Chronic Ill/Low Education[3]	6,677	0.91	0.85	0.84	0.73	0.74	0.73	0.82	0.75	0.89	0.84
Vision Problems (Mail)	4,456	0.93	0.90	0.89	0.82	0.83	0.84	0.87	0.83	0.93	0.89
Hearing Problems (Telephone)	1,751	0.91	0.87	0.83	0.74	0.81	0.75	0.85	0.81	0.91	0.87
Survey Mode											
Mail	123,516	0.93	0.91	0.88	0.83	0.88	0.86	0.87	0.83	0.94	0.89
Telephone	16,444	0.92	0.88	0.83	0.78	0.84	0.73	0.86	0.82	0.93	0.87

[1] Had 3+ self-report conditions (hypertension, angina, congestive heart failure, myocardial infarction, other heart disease, stroke, respiratory disease, gastrointestinal disorder, arthritis hip/knee, arthritis hand/wrist, sciatica, diabetes, cancer).

[2] Responded yes to at least one of two items about feelings of depression in past year.

[3] Had 3+ chronic conditions, screened positive for possibility of current depression, and <12 years of education.

NOTES: *N*=177,714. PF is physical functioning. RP is role limitations due to physical health problems. BP is bodily pain. GH is general health. VT is vitality. SF is social functioning. RE is role limitations due to emotional problems. MH is mental health. PCS is Physical Component Summary. MCS is Mental Component Summary.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

Table 6

Percent of Respondents Scoring at the Floor and Ceiling of Each SF-36® Scale, by Characteristics: Medicare HOS, 1998

Characteristic	*N*	PF		RP		BP		GH		VT		SF		RE		MH	
		Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling
Total Sample	163,840	2.6	8.8	29.4	43.2	1.5	18.6	0.4	2.9	1.9	1.7	1.6	46.8	16.7	65.8	0.2	6.9
Survey Completed By																	
Self-Adminstration	139,965	1.4	9.3	27.0	45.4	1.2	19.1	0.3	3.2	1.4	1.7	0.9	49.8	14.4	68.1	0.1	7.3
Proxy—Family Member	15,559	12.6	5.5	46.3	29.0	3.8	14.3	1.7	1.2	6.2	1.0	7.3	26.4	30.2	54.1	0.6	4.1
Proxy—Friend	919	9.5	5.8	48.1	26.7	5.3	14.9	2.2	1.4	4.5	1.7	6.5	26.1	35.6	47.0	1.3	5.0
Proxy—Caregiver	478	13.8	4.8	50.4	26.6	5.4	16.3	2.1	0.4	5.4	1.3	6.3	19.7	37.4	41.0	0.8	4.0
Self-Administered Forms Only																	
Age																	
18-64 Years	7,663	4.7	1.7	64.2	11.4	7.2	5.2	2.5	0.3	7.8	0.3	4.9	11.0	41.2	38.7	0.9	2.0
65-74 Years	81,079	0.8	12.2	19.9	54.3	0.7	21.3	0.1	4.1	0.9	2.1	0.5	56.5	10.2	75.1	0.1	7.6
75-84 Years	43,601	1.5	6.1	30.8	38.2	0.9	17.9	0.1	2.4	1.1	1.4	0.8	46.3	16.1	63.3	0.1	7.6
85 Years or Over	7,622	3.4	3.3	43.1	25.7	1.4	17.3	0.2	1.6	2.0	1.0	1.3	37.4	23.2	51.1	0.1	7.8
Sex																	
Male	59,866	1.4	10.9	25.5	47.8	1.0	21.2	0.3	3.3	1.3	2.1	0.8	51.4	13.1	70.5	0.1	8.5
Female	80,099	1.4	8.1	28.1	43.6	1.3	17.6	0.2	3.1	1.5	1.5	0.9	48.6	15.4	66.3	0.1	6.4
Race/Ethnicity																	
White (Non-Hispanic)	121,762	1.3	9.2	26.5	45.7	1.1	18.9	0.2	3.4	1.4	1.6	0.8	50.9	13.5	69.4	0.1	7.2
Black (Non-Hispanic)	8,253	2.1	8.4	33.2	38.8	2.1	19.2	0.4	1.4	1.4	2.4	1.3	40.4	21.8	55.3	0.2	7.8
Hispanic	5,565	1.8	11.3	30.7	45.0	1.9	21.3	0.6	3.0	1.6	2.9	1.3	42.6	21.7	60.1	0.2	8.2
Asian/Pacific Islander	1,949	0.8	12.1	20.8	57.1	0.4	28.1	0.2	3.2	0.3	2.8	0.4	51.2	13.7	73.4	0.0	7.5
Other	2,433	2.5	10.3	31.3	41.4	2.7	20.6	0.4	2.3	2.2	2.6	2.0	42.0	17.8	61.6	0.1	9.5
Education																	
8th Grade or<	12,057	2.8	7.7	37.1	35.0	2.2	17.9	0.6	1.5	2.5	1.7	1.5	38.7	23.9	53.0	0.3	7.0
Some High School	23,758	2.0	7.6	34.1	38.5	1.4	17.8	0.3	1.7	1.7	1.6	1.0	43.0	20.3	59.1	0.1	6.3
High School Graduate	49,767	1.2	8.7	26.5	45.8	1.1	19.0	0.2	2.8	1.3	1.6	0.8	51.1	13.9	69.2	0.1	7.1
Some College	31,265	1.0	9.3	24.7	47.2	1.1	18.7	0.3	4.0	1.3	1.7	0.9	51.3	11.4	72.1	0.1	7.5
College Graduate	20,981	0.8	13.0	17.2	55.8	0.6	22.7	0.2	5.8	0.6	2.2	0.6	58.7	7.6	78.9	0.1	8.7
Income																	
<$10,000	17,673	2.9	6.1	40.3	31.7	2.4	15.4	0.7	1.7	2.6	1.6	1.7	34.7	25.7	52.0	0.2	6.4
$10,000-$19,999	34,288	1.6	6.9	32.7	38.5	1.4	16.5	0.3	2.3	1.8	1.3	1.1	43.5	17.8	62.4	0.1	6.5
$20,000-$29,999	25,228	1.0	8.8	25.1	46.8	0.9	18.4	0.2	3.0	1.1	1.5	0.6	51.2	12.4	71.0	0.1	7.0
$30,000-$39,999	15,693	0.7	10.3	20.1	52.4	0.6	19.9	0.2	4.0	0.8	1.5	0.4	56.2	9.0	76.2	0.0	7.5
$40,000+	20,531	2.5	10.3	31.3	41.4	2.7	20.6	0.4	2.3	2.2	2.6	2.0	42.0	17.8	61.6	0.1	9.5

See footnotes at end of table.

Table 6—Continued

Percent of Respondents Scoring at the Floor and Ceiling of Each SF-36® Scale, by Characteristics: Medicare HOS, 1998

Characteristic	*N*	PF		RP		BP		GH		VT		SF		RE		MH	
		Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling	Floor	Ceiling
Other																	
On Medicaid	3,506	4.4	3.6	50.7	20.9	4.1	11.0	1.5	1.1	4.3	1.3	3.6	22.6	34.9	40.8	0.6	4.2
Disabled Entitlement	8,155	4.7	1.7	63.7	11.6	7.0	5.2	2.5	0.3	7.6	0.3	4.8	11.5	40.7	39.2	0.8	2.1
3+ Chronic Conditions[1]	60,991	2.2	2.4	41.6	27.0	2.2	7.7	0.5	0.6	2.5	0.4	1.5	33.3	21.9	57.1	0.2	4.3
Depression Screener[2]	30,702	3.4	3.1	49.9	20.7	3.5	8.4	1.0	0.7	4.2	0.2	2.9	15.0	41.2	31.0	0.5	0.5
Chronic Ill/Low Education[3]	6,677	5.6	1.0	64.4	10.0	5.3	4.0	1.4	0.1	6.0	0.0	3.7	8.7	52.5	20.4	0.8	0.3
Vision Problems (Mail)	4,456	4.1	4.6	45.2	26.3	3.1	12.9	1.3	1.5	3.5	0.8	2.5	29.2	29.8	46.9	0.3	4.0
Hearing Problems (Telephone)	1,751	2.5	6.5	30.5	34.4	3.0	19.1	0.5	1.0	3.4	1.3	3.1	37.1	19.1	59.3	0.6	5.9
Survey Mode																	
Mail	123,516	1.3	8.9	27.6	45.1	1.1	18.0	0.3	3.3	1.3	1.6	0.8	49.8	14.8	67.6	0.1	7.0
Telephone	16,444	2.1	11.7	22.7	47.1	1.9	27.4	0.3	2.4	1.9	2.8	1.4	49.4	11.9	72.0	0.2	9.7

[1] Had 3+ self-report conditions (hypertension, angina, congestive heart failure, myocardial infarction, other heart disease, stroke, respiratory disease, gastrointestinal disorder, arthritis hip/knee, arthritis hand/wrist, sciatica, diabetes, cancer).

[2] Responded yes to at least one of two items about feelings of depression in past year.

[3] Had 3+ chronic conditions, screened positive for possibility of current depression, and <12 years of education.

NOTES: *N*=177,714. PF is physical functioning. RP is role limitations due to physical health problems. BP is bodily pain. GH is general health. VT is vitality. SF is social functioning. RE is role limitations due to emotional problems. MH is mental health. PCS is Physical Component Summary. MCS is Mental Component Summary.

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

Floor effects for the RE scale also were notable for those who screened positive for likelihood of current depression.

Factor Analysis

Although eigenvalues for the first two factors were 5.08 and 0.77, extraction of two factors was supported by an examination of scree plots, variance explained by the second factor (10 percent), and substantial unrotated loadings (>0.48) for two scales (MH, RE) on the second factor. Seventy-three percent of the total and 85 percent of the reliable variance in scale scores was accounted for by the two components. The total and reliable variance explained in each scale by the two components was substantial, ranging from 68 to 80 percent (median=73 percent) across scales for the total variance and 79 to 97 percent (median=84 percent) for the reliable variance. The pattern of correlations observed between scales and the two rotated principal components supported their interpretation as physical and mental health components (Table 7). As hypothesized: the PF scale correlated strongest (0.85) with the physical component and weakest (0.18) with the mental component; both RP and BP scales had stronger correlations (RP=0.77, BP=0.78) with the physical component than the mental component (RP=0.33, BP=0.26); the MH scale correlated highest (0.85) with the mental component and along with RE correlated lowest (0.27) with the physical component; and the GH and VT scales had noteworthy correlations with both components.

The range of correlations observed between scales and the two rotated principal components across the subgroups supported the interpretation of the two components as physical and mental (results for five subgroups in Table 7; other results available on request from the author). The greatest departures were seen in the results across racial/ethnic groups. Within the Asian group, the BP, GH, and VT scales had higher loadings on the mental component and somewhat lower loadings on the physical component. The RE scale also had a higher loading on the physical component (0.48) and lower loading on the mental component (0.54) among Asians. This pattern has been seen in other studies of Asian populations (Gandek and Ware, 1998; Ren et al., 1998).

DISCUSSION

To the best of our knowledge, this study is the largest evaluation of data quality and psychometric performance of the SF-36® scales and summary measures among the elderly and disabled. Overall, data quality was satisfactory and scoring assumptions were met, across the total sample and within subgroups. Missing data rates generally were low, although somewhat higher for role physical and role emotional items. Tests of assumptions underlying the construction and scoring of scales generally were satisfied, although some analyses confirmed hypothesized poorer performance among the more disadvantaged subgroups. Regardless, internal consistency reliability estimates for the scales were above 0.70, the accepted standard for group comparisons, for all scales in all subgroups studied. Factor analysis confirmed the two hypothesized physical and mental components and strongly supported the construction and scoring of the PCS and MCS, as recommended by their developers (Ware et al., 1995). Reliability estimates were above 0.90 for the PCS in all but one subgroup and generally ranged from 0.87-0.90 for the MCS. The quality and consistency of these results is noteworthy given that the data were collected by six different survey vendors who used both mail and telephone administrations of surveys.

Table 7

Range of Correlations Between SF-36® Scales and Rotated Principal Components, Total Sample and Subgroups: Medicare HOS, 1998

	Total Sample (*N*=163,840)			Subgroups (Self-Administered Forms Only)									
				Age		Sex		Race/Ethnicity		Education		Income	
Scale	Physical	Mental	h^2/r_{tt}	Physical	Mental	Physical	Mental	Physical	Mental	Physical	Mental	Physical	Mental
PF	0.85	0.18	0.82	0.83-0.85	-0.02-0.16	0.84-0.86	0.15-0.20	0.83-0.87	0.17-0.20	0.85-0.86	0.13-0.17	0.84-0.87	0.13-0.15
RP	0.77	0.33	0.79	0.74-0.77	0.24-0.32	0.77-0.79	0.31-0.33	0.76-0.81	0.31-0.40	0.77-0.80	0.28-0.33	0.77-0.79	0.28-0.34
BP	0.78	0.26	0.79	0.74-0.79	0.22-0.26	0.77-0.80	0.25-0.26	0.68-0.79	0.24-0.40	0.73-0.79	0.20-0.38	0.77-0.79	0.19-0.31
GH	0.72	0.42	0.86	0.64-0.73	0.36-0.46	0.72-0.74	0.40-0.40	0.47-0.73	0.38-0.64	0.62-0.74	0.35-0.53	0.67-0.74	0.31-0.47
VT	0.72	0.46	0.85	0.60-0.72	0.42-0.51	0.71-0.73	0.44-0.47	0.45-0.72	0.45-0.71	0.62-0.73	0.44-0.56	0.66-0.74	0.42-0.51
SF	0.62	0.60	0.89	0.54-0.60	0.58-0.64	0.60-0.62	0.59-0.62	0.51-0.62	0.59-0.70	0.55-0.64	0.57-0.64	0.57-0.62	0.56-0.64
RE	0.27	0.81	0.84	0.14-0.28	0.74-0.83	0.26-0.27	0.81-0.81	0.25-0.48	0.54-0.82	0.20-0.41	0.63-0.84	0.18-0.29	0.77-0.84
MH	0.27	0.85	0.97	0.15-0.25	0.84-0.89	0.23-0.28	0.84-0.86	0.11-0.26	0.85-0.92	0.15-0.27	0.84-0.92	0.20-0.27	0.82-0.88

NOTES: PF is physical functioning. RP is role limitations due to health problems. BP is bodily pain. GH is general health. VT is vitality. SF is social functioning. RE is role limitations due to emotional problems. MH is mental health. h^2/r_{tt} is the variance in each scale explained by the two principal components (h^2) divided by the reliabiity of the scale (r_{tt}).

SOURCE: Centers for Medicare & Medicaid Services: Data from the Medicare Health Outcomes Survey, Cohort I Baseline, 1998.

Since most of the 36 items have their roots in instruments that were originally developed for use with the elderly and disabled (Ware et al., 1993), the strength of these results should not be surprising, and our results are consistent with those reported in the literature. Across five studies reporting data completeness (Andresen et al., 1996; Brazier et al., 1996; Lyons, Perry, and Littlepage, 1994; McHorney et al., 1994; Parker et al., 1998), the median percentage of respondents for whom scale scores could be calculated was 92-93 percent for RP and RE; 95 percent for PF, GH, and VT; and 97 percent for BP, SF, and MH. Hobson and Meara (1997) also reported that 24 percent of elderly Parkinson's disease patients missed one or more items (Hobson and Meara, 1997). Across 10 studies reporting internal consistency reliability statistics for the elderly (available from the author on request), median internal consistency reliability statistics ranged from 0.76 (GH) to 0.91 (PF), with other statistics ranging between 0.80 and 0.89. Floor and ceiling effects, which refer to concentrations of scores at the lowest and highest possible levels, have been reported frequently (Anderson, Laubscher, and Burns, 1996; Beusterien, Steinwald, and Burns, 1996; McHorney et al., 1994; McHorney, 1996; Reuben et al., 1995) although results have varied depending on the populations studied. However, median floor effects were low for most scales (0-2 percent for six scales across all five studies), but were substantial for the RP (30 percent) and RE (14 percent) scales. Median ceiling effects were 7 percent or below for four scales (PF, GH, VT, MH), but were higher for RP (31 percent), BP (20 percent), SF (49 percent), and RE (62 percent) in those studies.

Floor and/or ceiling effects are inevitable in the Version 1.0 role functioning scales because the role functioning items have two response choices (yes/no) and thus, the scales define only four or five distinct levels (Ware et al., 1993). To address this issue, Version 2.0 of the SF-36® uses five response options (all of the time, most of the time, some of the time, a little of the time, none of the time) for both role scales. This improvement has yielded a five-fold increase in the range of levels measured by both Version 2.0 role functioning scales and has substantially reduced ceiling and floor effects (Ware, Kosinski, and Dewey, 2000b). In a general population sample of adults age 65 or over who completed both SF-36® versions, floor effects for the role physical scale declined from 26 percent in Version 1.0 to 3 percent in Version 2.0, while ceiling effects declined from 38 to 17 percent. Similar results were seen for the role emotional scale, with a reduction in floor effects from 14 to 2 percent and in ceiling effects from 67 to 47 percent (Kosinski, 2004). The fact that the floor has been lowered the most is particularly relevant to the HOS because the elderly are more likely to decline than improve in health over time and it is not possible to measure a decline in functioning below the floor of any particular scale. Other solutions, including the targeting of role functioning items to each respondent's level of functioning using computerized dynamic health assessments, are currently being evaluated (Ware, 2003).

Among the advantages of the PCS and MCS summary measures, the primary HOS outcome measures, is their greater reliability over a much wider range of scores, their virtual elimination of floor and ceiling effects, and their greater validity in discriminating between physical and mental health outcomes (Ware et al., 2004). The eight scales are substantially intercorrelated (most >0.50 in the HOS) and substantially intercorrelated summary components would offer little advantage when it comes to interpreting outcomes. To

achieve these advantages, the HOS has adopted the orthogonal scoring of the PCS and MCS summary measures as recommended by the developers (Ware et al., 1995; Ware and Kosinski, 2001a). Specifically, the use of principal components analysis rather than principal factor analysis and orthogonal over oblique rotations has a number of advantages, including: a simple additive model of factor content, thereby facilitating the interpretation of each scale; summary measures that explain as much of the variance in the eight scales as possible; summary scales that are easy to score; and summary scales that have good discriminant validity and are interpretable as physical and mental dimensions of health (McHorney, Ware, and Raczek, 1993; Ware, 2000a).

Among the practical advantages of the PCS and MCS is the virtual elimination of ceiling and floor effects. In contrast to the results for the eight scales, only 2 percent of respondents scored within 15 points (1.5 standard deviations) of either the PCS or MCS ceiling or floor at baseline, within the first two HOS cohorts (Ware et al., 2004). Concerns expressed regarding extreme PCS and MCS scores (Taft, Karlsson, and Sullivan, 2001) have prompted examination of HOS and other data relevant to the validity and interpretation of very high and very low scores (Ware and Kosinski, 2001b). For example, in support of their validity in the HOS, beneficiaries with very low (i.e., PCS below 20) scores at baseline have one-fourth higher 2-year death rates, compared to those scoring 21-30 (21.4 versus 17.3 percent, respectively) at baseline (Ware and Kosinski, 2001b). A growing number of peer-reviewed randomized controlled trials and other health outcomes studies also support the validity of the PCS and MCS summary measures, including more than 120 summarized for 1994 to 2000 (Turner-Bowker, Bartley, and Ware, 2002).

As seen in other studies of the elderly, missing data rates were slightly higher for the role functioning items. It has been suggested that these items are problematic for the elderly because they include the word "work" (Hill, Harries, and Popay, 1996), and one researcher (Hayes et al., 1995) suggested rewording the role functioning items to ask about limitations in "regular daily activities (or work)". A study using the reworded items only found a slight improvement in missing data rates, and Hobson and Meara (1997) concluded that the disadvantage of losing comparability with other SF-36® studies far outweighed the minor gain in data completeness. We also note that in spite of somewhat higher missing data rates, role functioning scales could be scored for 95-96 percent of respondents overall, and for at least 92-93 percent of respondents in all subgroups.

This study used classical test theory to evaluate the SF-36® data. Studies of other data using item response theory (IRT) have shown strong linear associations between summated ratings scores for scales and those derived from IRT models except at the extremes, as would be expected (Haley, McHorney, and Ware, 1994). IRT results have also suggested that improvements in scale scoring algorithms are possible, especially for the PF scale. This improvement has been incorporated into the improved MDE scoring algorithms for the PCS and MCS.

Establishing that the scales and summary measures meet scaling assumptions is a necessary but insufficient prerequisite for their use. Examination of their validity is necessary to evaluate how well the measures have captured the underlying health constructs. Tests of clinical validity also have been used in conjunction with factor analysis to evaluate the assumptions underlying the construction of the summary measures. These validity tests have been reported in detail

elsewhere for other samples (McHorney, Ware, and Raczek, 1993; Ware et al., 1995; Ware and Kosinski, 2001a). Evidence of the clinical validity of the SF-36® in the HOS is available in other publications (Haffer et al., 2003) and interpretation guidelines for the scales and summary measures in the HOS have been published (Ware et al., 2004).

Finally, we note that while similar to response rates in other mail and telephone surveys of the elderly (Andresen et al., 1996), the baseline Cohort I response rate was only 64 percent. Non-respondents to the survey may have been more likely to have cognitive deficits, substantial vision or hearing impairments, or other characteristics that might result in poorer data quality. In addition, research has demonstrated that MMC beneficiaries in the 1990s were healthier than FFS beneficiaries (Aber and McCormick, 2000). Thus, our results may not generalize to all elderly and disabled populations. However, we note that the psychometric methods used are assumed to be robust across samples, and do not assume that the sample analyzed is representative of the underlying population. In addition, mean scores on the PCS and MCS for the HOS elderly in this sample were within 0.5 points of the U.S. general population age 65 or over in 1998 (Ware and Kosinski, 2001a). Thus, these results provide support for the continuing analysis and interpretation of scores among the elderly and disabled who are able to complete the SF-36® Health Survey, including those who by reasons of advanced age, lower education, or substantial illness may be most at risk for poor health outcomes.

REFERENCES

Aber, M. and McCormick, C.: Risk Adjustment and the Health of the Medicare HMO Population. *Health Care Financing Review* 21(3):275-280, Spring 2000.

Anderson, C., Laubscher, S., and Burns, R.: Validation of the Short Form 36 (SF-36) Health Survey Questionnaire Among Stroke Patients. *Stroke* 27:1812-186, 1996.

Andresen, E.M., Bowley, N., Rothenberg, B.M., et al.: Test-Retest Performance of a Mailed Version of the Medical Outcomes Study 36-Item Short-Form Health Survey among Older Adults. *Medical Care* 34(12):1165-1170, December 1996.

Beusterien, K.M., Steinwald, B., and Ware Jr., J.E.: Usefulness of the SF-36 Health Survey in Measuring Health Outcomes in the Depressed Elderly. *Journal of Geriatric Psychiatry and Neurology* 9(1):13-21, 1996.

Brazier, J.E., Walters, S.J., Nicholl, J.P., and Kohler, B.: Using the SF-36 and Euroqol on an Elderly Population. *Quality of Life Research* 5(2):195-204, April 1996.

Campbell, D.T. and Fiske, D.W.: Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin* 56(2):81-105, 1959.

Cronbach, L.J.: Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16(3):297-334, 1951.

Edwards, A.L.: *Techniques of Attitude Scale Construction*. Appleton-Century-Crofts, Inc., New York. 1957.

Gandek, B. and Ware, J.E. (eds.): Translating Functioning and Well-Being: International Quality of Life Assessment (IQOLA) Project Studies of the SF-36 Health Survey. *Journal of Clinical Epidemiology* 51(11):891-1214, November 1998.

Haffer, S.C., Bowen, S.E., Shannon, E.D., and Fowler, B.M.: Assessing Beneficiary Health Outcomes and Disease Management Initiatives in Medicare. *Disease Management and Health Outcomes* 11:111-124, 2003.

Haley, S.M., McHorney, C.A., and Ware, J.E.: Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): I. Unidimensionality and Reproducibility of the Rasch Item Scale. *Journal of Clinical Epidemiology* 47(6):671-684, June 1994.

Harman, H.H.: *Modern Factor Analysis*. (3rd ed. rev.) University of Chicago Press. Chicago, IL. 1976.

Hayes, V., Morris, J., Wolfe, C., and Morgan, M.: The SF-36 Health Survey Questionnaire: Is It Suitable for Use with Older Adults? *Age and Ageing* 24(2):120-125, March 1995.

Health Assessment Lab: *Psychometric Analysis of the Spanish-Language Version of the Medicare Health Outcomes Survey*. Boston, MA. Health Assessment Lab, 2000. Internet address: http://www.cms.hhs.gov/surveys/hos/download/HOS_Psychometrics_Spanish_HOS.pdf (Accessed 2004.)

Hill, S., Harries, U., and Popay, J.: Is the Short Form 36 (SF-36) Suitable for Routine Health Outcomes Assessment in Health Care for Older People? Evidence from Preliminary Work in Community Based Health Services in England. *Journal of Epidemiology and Community Health* 50:94-98, 1996.

Hobson, J.P. and Meara, R.J.: Is the SF-36 Health Survey Questionnaire Suitable as a Self-Report Measure of the Health Status of Older Adults with Parkinson's Disease? *Quality of Life Research* 6(3):213-216, January 1997.

Howard, K.I. and Forehand, G.G.: A Method for Correcting Item-Total Correlations for the Effect of Relevant Item Inclusion. *Educational and Psychological Measurement* 22:731-735, 1962.

Kosinski, M.: QualityMetric Incorporated, personal communication. Lincoln, RI. February 26, 2004.

Likert, R.: A Technique for the Measurement of Attitudes. *Archives of Psychology* 140(5):1-55, 1932.

Lyons, R.A., Perry, H.M., and Littlepage, B.N.: Evidence for the Validity of the Short-Form 36 Questionnaire (SF-36) in an Elderly Population. *Age and Ageing* 23:182-184, 1994.

McHorney, C.A.: Measuring and Monitoring General Health Status in Elderly Persons: Practical and Methodological Issues in Using the SF-36 Health Survey. *The Gerontologist* 36(5):571-583, October 1996.

McHorney, C.A., Ware, J.E., Lu, J.F.R., and Sherbourne, C.D.: The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions, and Reliability Across Diverse Patient Groups. *Medical Care* 32(1):40-66, January 1994.

McHorney, C.A., Ware, J.E., and Raczek, A.E.: The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs. *Medical Care* 31(3):247-263, March 1993.

National Committee for Quality Assurance.: *HEDIS® 2003 (Volume 6): Specifications for the Medicare Health Outcomes Survey*. National Committee for Quality Assurance. Washington DC. 2003.

Nunnally, J.C. and Bernstein, I.R.: *Psychometric Theory* (3rd Edition). McGraw-Hill. NewYork. 1994.

Parker, S.G., Peet, S.M., Jagger, C., et al.: Measuring Health Status in Older Patients. The SF-36 in Practice. *Age and Ageing* 27(1):13-18, January 1998.

Ren, X.S., Amick, B., Zhou, L., and Gandek, B.: Translation and Psychometric Evaluation of a Chinese Version of the SF-36 Health Survey in the United States. *Journal of Clinical Epidemiology* 51(11):1129-1138, November 1998.

Reuben, D.B., Valle, L.A., Hays, R.D., and Siu, A.L.: Measuring Physical Function in Community-Dwelling Older Persons: A Comparison of Self-Administered, Interviewer-Administered, and Performance-Based Measures. *Journal of the American Geriatric Society* 43:17-23, 1995.

Taft, C., Karlsson, J., and Sullivan, M.: Do SF-36 Summary Component Scores Accurately Summarize Subscale Scores? *Quality of Life Research* 10(5):395-404, January 2001.

Turner-Bowker, D.M., Bartley, B.J., and Ware, J.E.: *SF-36® Health Survey and "SF" Bibliography: Third Edition (1988-2000)*. QualityMetric Incorporated. Lincoln, RI. 2002.

Ware, J.E.: SF-36 Health Survey Update. *Spine* 25(24):3130-3139, 2000a.

Ware, J.E.: Conceptualization and Measurement of Health-Related Quality of Life: Comments on an Evolving Field. *Archives of Physical Medicine and Rehabilitation* 84(4 Suppl 2):S43-51, 2003.

Ware, J.E., Gandek, B., Sinclair, S.J, and Kosinski, M.: *Measuring and Improving Health Outcomes: An SF-36® Primer for the Medicare Health Outcomes Survey*. Health Assessment Lab and QualityMetric Incorporated. Waltham, MA. 2004.

Ware, J.E., Harris, W.J., Gandek, B., et al.: *MAP-R for Windows: Multitrait/Multi-Item Analysis Program—Revised User's Guide*. Health Assessment Lab. Boston, MA. 1997.

Ware, J.E. and Kosinski, M.: *SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version 1, Second Edition*. QualityMetric Incorporated. Lincoln, RI. 2001a.

Ware, J.E. and Kosinski, M.: Interpreting SF-36 Summary Health Measures: A Response. *Quality of Life Research* 10(5):405-413, January 2001b.

Ware, J.E., Kosinski, M., Bayliss, M.S., et al.: Comparison of Methods for Scoring and Statistical Analysis of SF-36 Health Profiles and Summary Measures: Summary of Results from the Medical Outcomes Study. *Medical Care* 33(Suppl 4):AS264-AS279, 1995.

Ware, J.E., Kosinski, M., and Dewey, J.E.: How to Score Version 2 of the SF-36® Health Survey. QualityMetric Incorporated. Lincoln, RI. 2000b.

Ware, J.E. and Sherbourne, C.D.: The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection. *Medical Care* 30:473-483, 1992.

Ware, J.E., Snow, K.K., Kosinski, M., and Gandek, B.: *SF-36 Health Survey Manual and Interpretation Guide*. The Health Institute. Boston, MA. 1993.

Reprint Requests: Barbara Gandek, M.S., Health Assessment Lab, 235 Wyman Street, Suite 130, Waltham, MA 02451. E-mail: bgandek@hal-health.org