SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales

John E Ware
John Ware Research Group

SF-12

How To Score the SF-12 Physical & Mental Health Summary Scales (Second Edition)

PCS-12®

MCS-12

PCS-12

MCS-12

PCS-12

MCS-12

SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales

John E. Ware, Jr., Ph.D.
Mark Kosinski, M.A.
Susan D. Keller, Ph.D

The Health Institute, New England Medical Center,
Boston, Massachusetts

Second Edition: December, 1995
First Edition: March, 1995

Permission to reproduce the SF-12™ Health Survey is routinely granted by the Medical Outcomes Trust (MOT) to individuals and organizations for their own use. The MOT is a non-profit corporation created to ensure the availability of the SF-12 Health Survey and other instruments, while preserving the standardization of their content, scoring, and labeling.

Requests for permission to reproduce the SF-12 Health Survey should be sent to Medical Outcomes Trust, 20 Park Plaza, Suite 1014, Boston MA 02116-4313.

Requests for permission to reproduce parts of this manual should be sent to John E. Ware, Jr., Ph.D., The Health Institute, NEMC-345, 750 Washington St., Boston MA 02111.

Suggested citation:

Ware JE, Kosinski M, Keller SD. SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. Boston, MA: The Health Institute, New England Medical Center, Second Edition, 1995.

SF-12™
Health Survey
Standard Scoring

This symbol identifies scoring documents, computer software, and data processing services that produce results that are comparable with SF-12 results reported in this manual. Look for this symbol for assurance that standard SF-12 scoring has been reproduced.

❦ Table of Contents

❦ Acknowledgements

Development and validation of the SF-12 Health Survey was supported by The Health Institute at New England Medical Center from its own research funds and by unrestricted research grants for the International Quality of Life Assessment (IQOLA) Project from Glaxo Research Institute, Research Triangle Park, North Carolina, and Schering-Plough Corporation, Kenilworth, New Jersey (J.E Ware, Jr., Ph.D., Principal Investigator).

We gratefully acknowledge the contributions of our colleagues at The Health Institute including, Martha Bayliss, M.S. and Barbara Gandek, M.S. who provided valuable assistance in studies of the reliability and validity of the SF-12 Health Survey; John E. Brazier, M.Sc., who provided the test-retest data from the U.K.; the Bay Area Business Group on Health for providing data on administrations of the SF-12 survey; William H. Rogers, Ph.D., who provided statistical consultation.

Several documents about the SF-36 Health Survey proved useful in preparing scoring documentation for the SF-12, including: SF-36 scoring manuals published by The Health Institute, New England Medical Center, entitled *How to Score the MOS 36-Item Short-Form Health Survey (SF-36)* (Ware, 1988; Medical Outcomes Trust, 1991); *Scoring Exercise for the SF-36 Health Survey* (Medical Outcomes Trust, 1994); the *SF-36 Health Survey Manual and Interpretation Guide* (Ware, Snow, Kosinski, and Gandek, 1993); and the *SF-36 Physical and Mental Health Summary Scales: A User's Manual* (Ware, Kosinski, and Keller, 1994).

❦ Preface to the Second Edition

This manual documents the development of the SF-12, explains each step in the scoring process in detail, explains how to test the accuracy of scores, and presents detailed tables of results of tests of reliability and validity not reported elsewhere. The equivalence of SF-12 and SF-36 norms is also documented.

There has been much progress since efforts to develop the SF-12 Health Survey began at The Health Institute in the Spring of 1994. A peer-reviewed journal article entitled "A 12-Item Short-Form Health Survey (SF-12): Scale Construction and Preliminary Tests of Reliability and Validity," has been accepted for publication in *Medical Care*. This article explains how and why the SF-12 was developed and presents studies of the reliability and validity of the SF-12 physical and mental health summary scales and the eight-scale SF-12 health profile. The tentative publication date for the *Medical Care* article is March, 1996.

The Scientific Advisory Committee (SAC) of the Medical Outcomes Trust has completed its own independent peer review of the SF-12 Health Survey on the basis of studies to date. In September, 1995 the SAC approved the SF-12 for distribution by the Trust. Many investigators and organizations have adopted the SF-12 for large-scale health outcomes monitoring efforts including the National Commission on Quality Assurance (NCQA), which chose the SF-12 for the Annual Member Health Care Survey required for accreditation.

Because of the complexity of the SF-12 scoring algorithms, which are composites of weighted item responses, a computer diskette with the code for scoring algorithms and a test dataset are included for use in computing SF-12 summary scale scores and for checking the accuracy of computations.

1. How to Use This Manual

Suggestions regarding how to use this manual and how to find specific information quickly are offered below.

Background

What is the SF-12? *Chapter 2* briefly summarizes the logic and development of the SF-12 Health Survey.

Permission to Use

The one step process involved in getting royalty free permission to reproduce and use the SF-12 is explained in *Chapter 2*

Development

Chapter 3 explains the selection of 12 questionnaire items for the SF-12 Standard version, which uses a 4-week recall, and the Acute version, which uses a 1-week recall.

Administration Guidelines

The SF-12 can be self-administered, administered by face-to-face personal interview, by telephone personal interview, or by computer. A new "left-to-right" format, which improves the consistency of responses to the SF-12, is explained in *Chapter 3*. Copies of SF-12 forms and the script for personal interviews are in *Appendix C*.

Scoring Algorithms

Chapter 4 explains how to score SF-12 items as well as physical and mental health summary scales.

Scoring Checks

To test the accuracy of scoring, the use of a test dataset and the results that should be observed are explained in *Chapter 5*.

Reliability of Scores

Chapter 6 presents estimates of the reliability of SF-12 scores based on the test-retest method. Average changes in individual scores are also documented.

Norms

For tables documenting equivalence of norms for the SF-12 and SF-36 surveys in a representative sample of the general U.S. population, go to *Chapter 7*.

Validity

Chapter 8 documents results from empirical tests of validity.

Future Directions

Translations and other work in progress are briefly discussed in *Chapter 9*. Advantages and disadvantages of the SF-12 are discussed.

❦ 2. Introduction

What is the SF-12?

The SF-12 is a multipurpose short-form (SF) generic measure of health status. It was developed to be a much shorter, yet valid, alternative to the SF-36 for use in large surveys of general and specific populations as well as large longitudinal studies of health outcomes. The SF-12 is rapidly becoming an instrument of choice for purposes of monitoring the health of both general and specific populations because it is substantially shorter than the SF-36. It has been adopted for many large population outcomes monitoring efforts that did not include the SF-36 because of its length. It appears that more than 1 million SF-12 surveys will be administered in 1995 and the SF-12 has been selected for inclusion in the National Committee for Quality Assurance (NCQA) *Annual Member Health Care Survey* (Version 1.0), which NCQA and many large employers require for accreditation. These trends confirm the expected practical advantage of the SF-12.

The 12 items in the SF-12 are a subset of those in the SF-36; SF-12 includes one or two items from each of the eight health concepts (see Table 2.1). Thus, the SF-12 measures eight concepts commonly represented in widely used surveys: physical functioning, role limitations due to physical health problems, bodily pain, general health, vitality (energy/fatigue), social functioning, role limitations due to emotional problems, and mental health (psychological distress and psychological well being). Both standard (4-week) and acute (1-week) recall versions are available (see Appendix C).

TABLE 2.1 NUMBER OF SF-36 AND SF-12 HEALTH SURVEY ITEMS PER CONCEPT

Concept	SF-12	SF-36
Physical Functioning (PF)	2	10
Role-Physical (RP)	2	4
Bodily Pain (BP)	1	2
General Health (GH)	1	5
Energy/Fatigue (VT)	1	4
Social Functioning (SF)	1	2
Role-Emotional (RE)	2	3
Mental Health (MH)	2	5
Change in Health (HT)	-	1

Objectives

The SF-36 Health Status Survey is widely used throughout the world because it is brief, readily available without charge, psychometrically sound, and of proven usefulness in measuring and interpreting health status and outcomes in both general and specific populations. However, many large-scale projects omit health status because their databases are limited to the information that can be collected using a single-page scannable questionnaire or only a few minutes of interviewing time.

Our admittedly ambitious goal for the SF-12 Project was to develop a one-page, two-minute, questionnaire module that would reproduce the SF-36 physical and mental health summary measures with at least 90% accuracy while representing all eight SF-36 health concepts with one or more questionnaire items.

SF-12 forms for self-administration and scripts for personal interviews can be administered to most people in two minutes or less and have been used with a high degree of acceptability and data quality. Because the SF-12 is a subset of the SF-36, translations and adaptations of the SF-36 currently being evaluated in over 30 countries (Ware, Keller, Gandek, et al., 1995) will yield translations of the SF-12.

PCS and MCS Summary Scales

The construction of the SF-36 Physical Component Summary (PCS) and Mental Component Summary (MCS) scales and evidence to date supporting their usefulness in cross-sectional and longitudinal tests provided the foundation for the construction of a health survey that is much shorter than the SF-36. The number of items in a survey is, at least in part, a function of the number of health dimensions for which separate scores are to be estimated with precision. If two summary scores are useful for many purposes, as suggested by results reported to date for the PCS and MCS scales from the SF-36 (Ware, Kosinski, and Keller, 1994; Ware, Kosinski, Bayliss, et al., 1995), it is likely that the SF-36 can be further shortened for use in monitoring health outcomes in general and specific populations and possibly for other purposes.

Following the discovery of the PCS and MCS summary scales based on the SF-36, we began evaluating how well those summary scales could be reproduced from a much shorter questionnaire. We identified ten items from six of the eight SF-36 scales that reproduced at least 90% of the variance in both PCS and MCS, as defined using the SF-36 scales. Addition of two more items created a 12-item short-form that yielded more than satisfactory estimates of the PCS and MCS and made it possible to reproduce the profile of eight SF-36 health concepts. We labelled the new short form the SF-12 Health Survey (Ware, Kosinski, and Keller, 1996).

The SF-12 uses two items each to estimate scores for four of the eight health concepts (physical functioning, role-physical, role-emotional, and mental health). Scores for the remaining four health concepts (bodily pain, general health, vitality, and social functioning) are estimated using one item each. Results from empirical studies to date indicate that 12-item versions of PCS and MCS correlate very highly with the SF-36 versions (see Figures 2.1 through 2.4).

Limitations of the SF-12

The summary scales based on the SF-12, which we refer to as PCS-12 and MCS-12, are based on fewer items, define fewer levels, and should be expected to yield less reliable assignments of individuals to those levels than the SF-36 based summary scales. For large group studies these differences in precision are not as important, because confidence intervals for group averages are largely determined by sample size. Scoring algorithms for an 8-scale profile based on the SF-12 are currently being evaluated and will be the subject of a forthcoming report.

The eight-concept profile based on SF-12 items (calibrated to reproduce the original SF-36 scales) appears to be very similar, on average, to the original SF-36 profile, although each score is estimated with less precision. This disadvantage of single-item and two-item scales relative to longer multi-item scales is well documented (McHorney, Ware, Rogers, et al., 1992). Advantages and disadvantages of the SF-12 are further discussed in Chapter 9.

Registered users of the SF-12 will be sent a scoring update as soon as work on the SF-12 8-scale profile is completed. They also have the option of using the "SF-12 Hotline" for assistance with scoring (see Chapter 4). A registration form is contained in Appendix D.

Permission to Use the SF-12

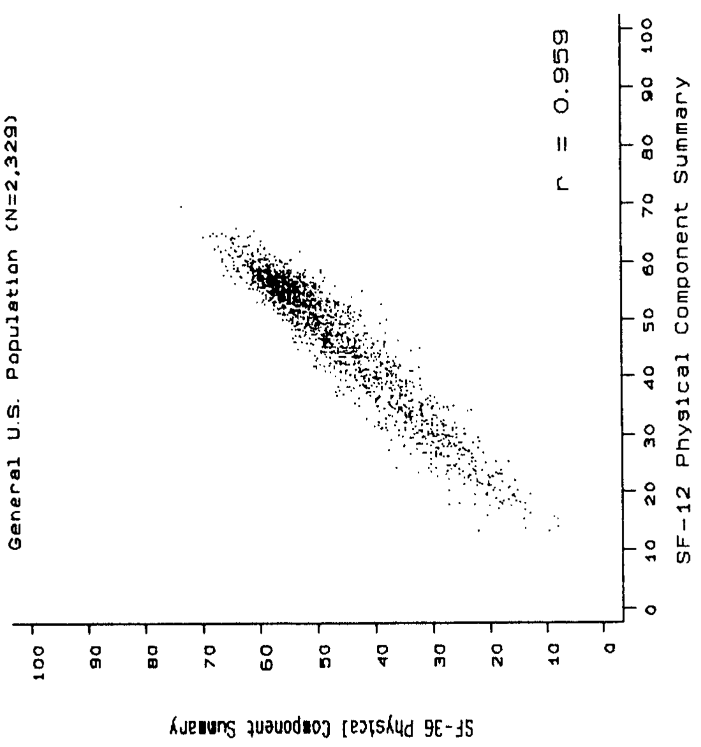
Permission to use and reproduce the SF-12 is routinely granted by the Medical Outcomes Trust (MOT) without charge. The trust is a nonprofit clearinghouse for widely used patient-based measures. Permission to reproduce SF-12 items and scoring algorithms has also been granted to computer software vendors and commercial survey and data processing firms offering a wide range of services based on standard SF-12 scoring algorithms and interpretation guidelines. Computer software products and scoring services that are comparable with this manual carry the SF-12 label (see Chapter 4).

Information about the SF-12 and permission to reproduce it can be obtained from The Medical Outcomes Trust, 20 Park Plaza, Suite 1014, Boston, MA 02116-4313. The telephone number is 617-426-4046; the fax number is 617-426-4131.

Information about translations of the SF-12 can be obtained from the IQOLA Project, The Health Institute, NEMC-345, 750 Washington St., Boston, MA 02111.

Figure 2.1

Plot of SF-36 and SF-12 Physical Component Summary Scores General U.S. Population (N=2,329)



Source: 1990 NSFHS

Figure 2.2

Plot of SF-36 and SF-12 Mental Component Summary Scores General U.S. Population (N=2,329)

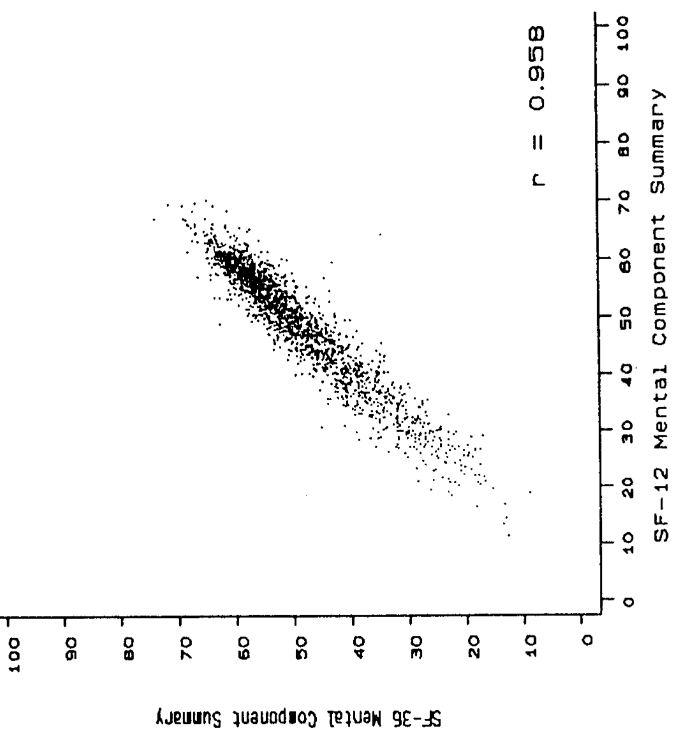
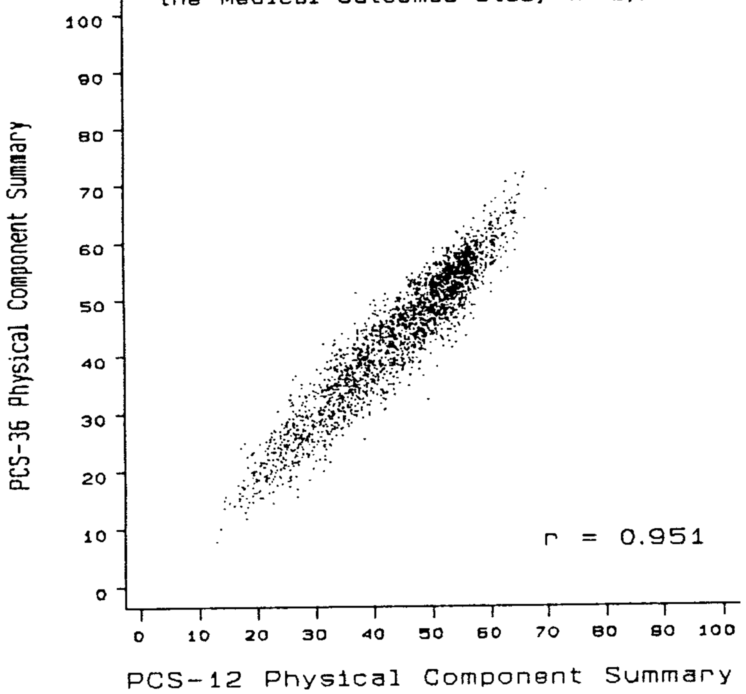


Figure 2.3

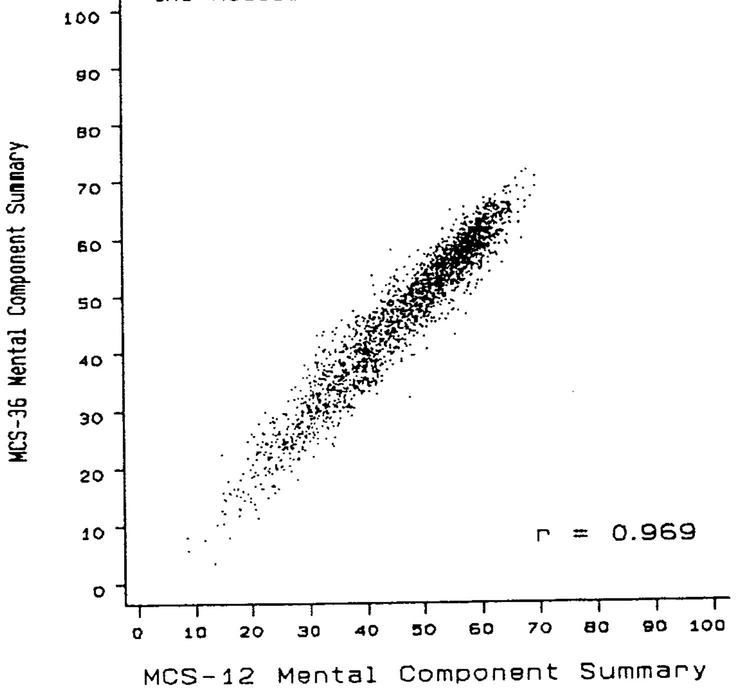
Plot of SF-12 and SF-36 Physical Component Summary Scores in the Medical Outcomes Study (N=2,833)



Source: Medical Outcomes Study (MOS)

Figure 2.4

Plot of SF-12 and SF-36 Mental Component Summary Scores in the Medical Outcomes Study (N=2,833)



❦ 3. Construction of the SF-12

Important Developments

Two developments set the stage for the construction of a much shorter version of the SF-36 Health Survey. The first was the finding that physical and mental health factors accounted for 80-85% of the reliable variance in the eight SF-36 scales in both patient and general populations in the U.S. and in other countries (Ware, Snow, Kosinski, and Gandek, 1993; McHorney, Ware, and Raczek, 1993; Ware, Keller, Gandek, et al., 1995). The second was the finding that SF-36 Physical and Mental Component Summary scales rarely missed a hypothesized difference in cross-sectional and longitudinal tests based on independent physical and mental criterion variables. These results suggested that it may be possible to further reduce the number of items in the SF-36 without substantial loss of information.

The successes of summary measures set the stage for constructing 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. A shorter survey is possible if summary measures can be reproduced well with fewer questionnaire items.

Shorter Summary Measures

In the Spring of 1994, we began evaluating how well scores for the two SF-36 summary scales could be reproduced using a much smaller subset of items. Our objectives were to develop a shorter survey that: a) could be scored to explain at least 90% of the variance in SF-36 physical and mental health summary measures, b) could fit on a single page of a scannable form, c) could be administered in less than two minutes on average, and d) would yield comparable average scores for the eight-scale health status profile in population studies. The first three objectives were achieved as documented here and elsewhere (Ware, Kosinski, and Keller, 1996).

The last objective is currently being pursued. Scoring advances that will improve the correspondence between 8-scale health profiles estimated from SF-12 and SF-36 Health Surveys are being evaluated (Ware, Kosinski, and Keller, 1996).

As documented in detail elsewhere (Ware, Kosinski, and Keller, 1996), summary measures were constructed independently to reproduce corresponding SF-12 physical and mental health summary measures. Regression methods were used to identify a subset of SF-12 items from the SF-36 and robust weighting algorithms for reproducing the SF-36 Physical Component Summary (PCS-36) and the SF-36 Mental

Component Summary (MCS-36). We also sought to represent all eight SF-36 health concepts with one or more items (Figure 3.1). Ten items representing six of the eight SF-36 concepts were sufficient to achieve our objective of reproducing PCS-36 and MCS-36 scores with an R^2 above 0.90 or greater. Two additional items were selected to represent all eight concepts.

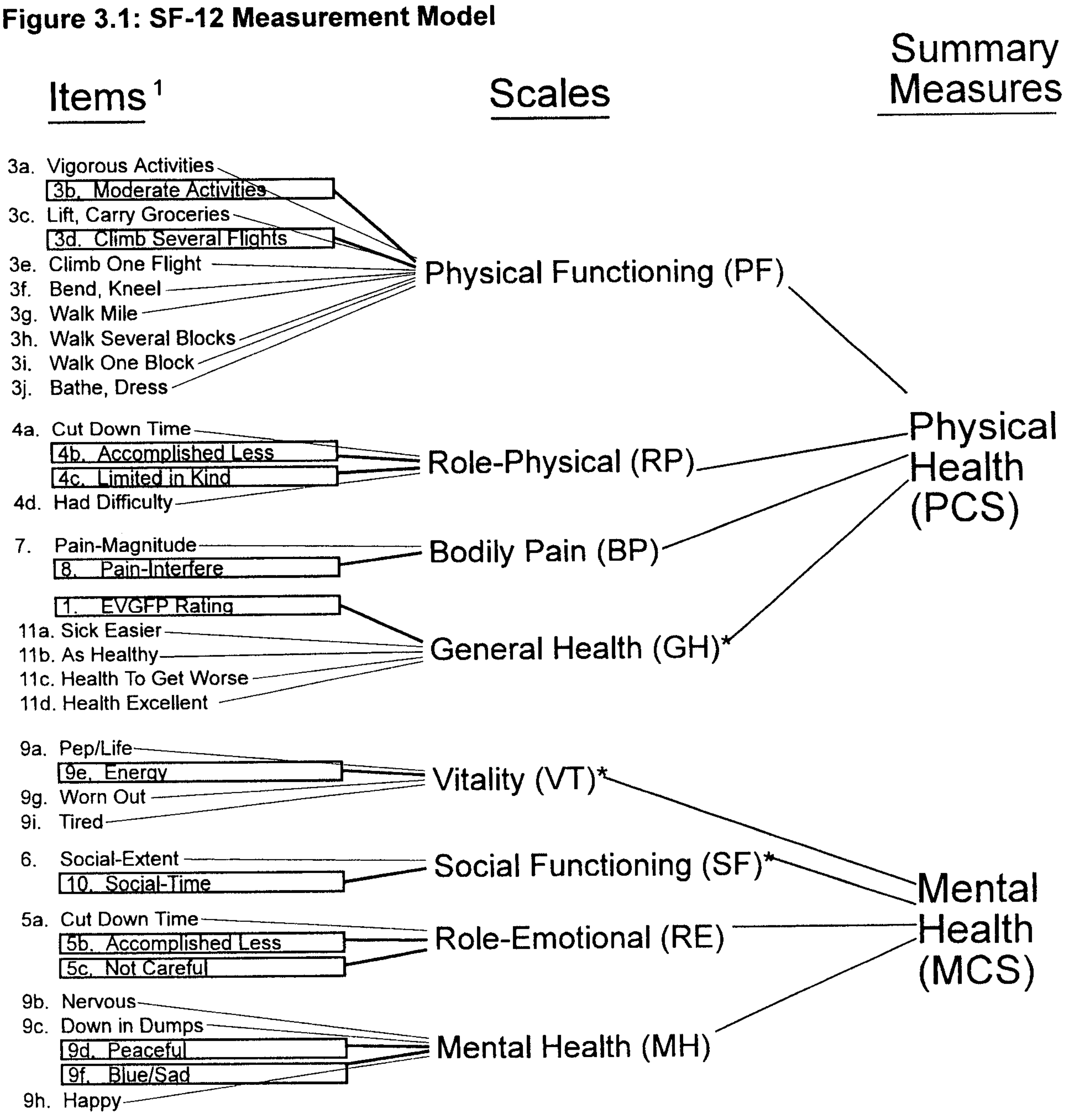
The accuracy of reproductions of physical and mental health summary scores proved, not surprisingly, to be dependent on the scoring of item response choices. Item response categories defined as "dummy" variables, which allowed the intervals between those categories to vary, improved R^2 values by about 8-9% ($p < 0.01$) (Ware, Kosinski, and Keller, 1996).

Comment

The SF-12 Health Survey represents another attempt to balance the number of questionnaire items against other important considerations that determine the usefulness of the resulting scores. The strategy of trying to predict summary measures for two clusters of substantially related SF-36 scales appears to have been successful with the SF-12. By pooling the reliable variance across physical health measures and the reliable variance across mental health measures, we have achieved satisfactory reliability while reducing the number of items. By summarizing measures that have been shown previously to produce the same or very similar results, summary measures simplify the analysis of health outcomes while minimizing information loss (Ware, Kosinski, Bayliss, et al., 1995).

The high degree of correspondence between SF-36 PCS and MCS summary scores estimated using SF-12 items suggests that PCS-12 and MCS-12 scores will have much the same interpretation as scores estimated using the full SF-36. Can norms and other interpretation guidelines published for the SF-36 summary measures be used in interpreting SF-12? These issues are addressed in chapters seven and eight.

Figure 3.1: SF-12 Measurement Model



* Significant correlation with other summary measure.

[1] Abbreviated item content (see Appendix C). Items in boxes were selected for SF-12.

Source: Ware, Kosinski, and Keller, 1994; Ware, Kosinski, and Keller, in press

❦ 4. How to Score the SF-12 Summary Measures

Overview

The SF-12 physical (PCS-12) and mental (MCS-12) component summary scales are scored using norm-based methods. Physical and mental regression weights and a constant for both measures come from the general U.S. population. Both the PCS-12 and MCS-12 scales are transformed to have a mean of 50 and a standard deviation of 10 in the general U.S. population.

The advantages of the standardization and norm-based scoring of the PCS-12 and MCS-12 is that results for one can be meaningfully compared with the other and their scores have a direct interpretation in relation to the distribution of scores in the general U.S. population. Specifically, all scores above and below 50 are above and below the average, respectively, in the general U.S. population. Because the standard deviation is 10 for both PCS-12 and MCS-12 scales, each one point difference in scores also has a direct interpretation. A one-point difference is one-tenth of a standard deviation.

Scoring the SF-12 physical (PCS-12) and mental (MCS-12) summary scales involves four steps:

(1) "cleaning" out-of-range values for item response choices and reverse scoring four items so that a higher score indicates better health,

(2) indicator variables (scored 1/0) are created for the item response choice categories,

(3) indicator variables are weighted (using regression coefficients from the general U.S. population), are aggregated, and

(4) by adding a constant (regression intercept), aggregate PCS-12 and MCS-12 scores are standardized to have the same mean as SF-36 versions in the general U.S. population.

General U.S. population statistics used in the standardization and in the aggregation of PCS-12 and MCS-12 scales are presented in Table 4.1. Detailed information including formulas for scale aggregation and transformation of scores are presented below. These steps ensure the standardization of scoring PCS-12 and MCS-12 for purposes of comparing results across studies. Any changes in scoring may compromise such comparisons as well as the reliability and validity of the resulting scale scores.

Steps in Scoring

Step 1: Data Cleaning and Item Recoding

All 12 items should be checked for out-of-range values, prior to assigning the final item value. Out-of-range values are those values that are lower than an item's precoded minimum value or higher than an item's precoded maximum value. All out-of-range values should be recoded as missing data.

Reverse scoring of four items is required so that a higher item value indicates better health for all SF-12 items and summary scales. Four SF-12 items are reverse scored because higher precoded item values for these items indicate a poorer health state. The four items that are reversed scored are: GH1 (item #1), BP2 (item #8), MH3 (item #9), and VT2 (item #10). For example, the highest precoded value for the item "How much of the time did you feel calm and peaceful" is "6-None of the time", which indicates a poor health state.

Step 2: Creating Indicator Variable for Item Response Choices

The second step in scoring PCS-12 and MCS-12 summary scales consists of creating indicator variables (1/0) for all but one of the response choice categories of each item. A one is assigned to the response choice if endorsed and a zero is assigned if it is not endorsed. Note that an indicator variable is not created for the response choice category indicating the highest health state for each item. Therefore, out of 47 total possible response choice categories among the 12 items, only 35 indicator variables are created. For example, the physical functioning item assessing limitations in moderate activities (PF02) has three response choice categories, 1=yes, limited a lot; 2=yes, limited a little; and 3=no, not limited at all. Indicator variables are derived for response choices 1 and 2, a total of two indicator variables. No indicator variable would be derived for the third response choice category, that which represents the highest health state for moderate activities. The same logic is repeated for each of the other 11 items.

Step 3: Weighting and Aggregation of Indicator Variables

The third step involves the weighting of indicator variables and computation of aggregate scores for physical and mental summary scales. It should be noted that two sets of regression weights, physical and mental, from the general U.S. population are utilized (see Table 4.1). Computation of PCS-12 is achieved by multiplying each indicator variable by its respective *physical* regression weight and summing the 35 products. Similarly, MCS-12 is computed by multiplying each indicator variable by its respective *mental* regression weight and summing the 35 products.

Pending results from ongoing evaluations of other scoring options for handling missing data, it is recommended that SF-12 summary scale scores be set to missing if the respondent is missing any one of the SF-12 items in the survey.

Step 4: Norm-Based Standardization of Scale Scores

The fourth step involves transforming each summary scale score to the norm-based scoring, which is referred to as "50/10" scoring because means of 50 and standard deviations of 10 are achieved in the general U.S. population. For the PCS-12, this is accomplished by adding the respective constant (Table 4.1) to the sum of the 35 products (physical) from step two. Similarly, for the MCS-12, this is accomplished by adding the respective constant (Table 4.1) to the sum of the 35 products (mental) from step two.

Scoring Checks

Because errors can lead to inaccurate summary scale scores, we strongly recommend formal checks of the accuracy of scoring. First, it is important to check each of the 12 items for out of range values for the response choices. All out of range values should be set to missing. Second, the indicator (dummy) variables derived for the item response choices should only have values of zero and one.

The following checks for item-scale correlations are also strongly recommended once the SF-12 summary scales have been scored. First, the physical functioning (PF02, PF04), role physical (RP2, RP3), and bodily pain (BP2) items should correlate highest with the PCS-12 scale and lowest with the MCS-12 scale. The social functioning (SF2), role emotional (RE2, RE3), and mental health (MH3, MH4) items should correlate highest with the MCS-12 scale and lowest with the PCS-12 scale. In most instances, the general health item (GH1) will correlate higher with the PCS-12 scale and the vitality item (VT2) will correlate higher with the MCS-12 scale. Second, the correlation between the PCS-12 and MCS-12 should be very low. Any discrepancies should be investigated for scoring errors.

Comment

Whereas simple equal-interval (linear) scoring has proven satisfactory for all but two of the SF-36 questionnaire items in studies to date (Ware, Keller, Gandek, et al., 1995), more complicated scoring yielded significant gains in reproducibility for the SF-12. The reason is that the information value of each questionnaire item is crucial when there are many fewer items.

Weighted item response categories improved the "linear fit" between physical and mental health levels and questionnaire item scores by more than seven percent. Accordingly, average SF-12 scores more closely approximated those based on the SF-36 in all groups studied. Therefore, to maximum comparability with the interpretation guidelines for SF-36 versions of PCS and MCS, the more complicated item response weighting documented in this chapter was adopted for the SF-12. To facilitate use of these scoring algorithms a computer diskette with scoring algorithms, a test data set, and written documentation are included in this manual.

SF-12 algorithms have been made available to computer software vendors and to other organizations providing scoring and analysis services for the SF-12. Look for the symbol to the right:

SF-12™
Health Survey
Standard Scoring

This symbol is your assurance that computer software products and data processing services produce results that are comparable with this Manual and with other normative data and interpretation guidelines for the SF-12 Health Survey.

Registered users of the SF-12 Health Survey who encounter any problems with scoring PCS-12 and MCS-12 have the option of calling the "SF-12 Hotline" at The Health Institute at (617) 636-8098, or FAX (617) 636-8077.

TABLE 4.1. WEIGHTS USED TO SCORE PHYSICAL (PCS-12) AND MENTAL (MCS-12) SCALES

Item **Response Choice(s)**	**Indicator Variable (i/o)**	**Physical Weight**	**Mental Weight**
Moderate Activities (PF02)			
Limited a lot	PF02_1	-7.23216	3.93115
Limited a little	PF02_2	-3.45555	1.86840
Climbing Several Flights of Stairs (PF04)			
Limited a lot	PF04_1	-6.24397	2.68282
Limited a little	PF04_2	-2.73557	1.43103
Accomplish less than you would like (RP2)			
Yes	RP2_1	-4.61617	1.44060
Limited in the kind of activities (RP3)			
Yes	PR3_1	-5.51747	1.66968
Pain interferes with normal work (BP2)			
Extremely	BP2_1	-11.25544	1.48619
Quite a bit	BP2_2	-8.38063	1.76691
Moderately	BP2_3	-6.50522	1.49384
A little bit	BP2_4	-3.80130	0.90384
In general, would you say your health is (GH1)			
Poor	GH1_1	-8.37399	-1.71175
Fair	GH1_2	-5.56461	-0.16891
Good	GH1_3	-3.02396	0.03482
Very good	GH1_4	-1.31872	-0.06064
Have a lot of energy (VT2)			
None of the time	VT2_1	-2.44706	-6.02409
A little of the time	VT2_2	-2.02168	-4.88962
Some of the time	VT2_3	-1.61850	-3.29805
A good bit of the time	VT2_4	-1.14387	-1.65178
Most of the time	VT2_5	-0.42251	-0.92057
Health interferes w/social activities (SF2)			
All the time	SF2_1	-0.33682	-6.29724
Most of the time	SF2_2	-0.94342	-8.26066
Some of the time	SF2_3	-0.18043	-5.63286
A little of the time	SF2_4	0.11038	-3.13896
Accomplish less than you would like (RE2)			
Yes	RE2_1	3.04365	-6.82672
Didn't do activities as carefully as usual RE3)			
Yes	RE3_1	2.32091	-5.69921
Felt calm and peaceful (MH3)			
None of the time	MH3_1	3.46638	-10.19085
A little of the time	MH3_2	2.90426	-7.92717
Some of the time	MH3_3	2.37241	-6.31121
A good bit of the time	MH3_4	1.36689	-4.09842
Most of the time	MH3_5	0.66514	-1.94949
Felt downhearted and blue (MH4)			
All of the time	MH4_1	4.61446	-16.15395
Most of the time	MH4_2	3.41593	-10.77911
A good bit of the time	MH4_3	2.34247	-8.09914
Some of the time	MH4_4	1.28044	-4.59055
A little of the time	MH4_5	0.41188	-1.95934
Constant	-	56.57706	60.75781

❦ 5. Scoring Exercise and Test Dataset

The purpose of this scoring exercise is to help SF-12 users evaluate results from each step in the process of calculating SF-12 physical (PCS-12) and mental (MCS-12) summary scales.

A test dataset and SAS code for scoring PCS-12 and MCS-12 scales has been provided on a computer diskette (see inside of back cover). The test dataset, which is called "SF12RAW.DAT" on the diskette contains data from 50 administrations of the SF-12.

The enclosed diskette also provides the user with the SAS code used to derive the SF-12 physical and mental summary scales. The SAS program is called "SF12SUMM.SCR" on the diskette. The SAS code begins with procedures designed to clean out of range values for the 12 items and to reverse the scoring of four items so that a higher score indicates better health. Next, the SAS code derives the 35 indicator variables for the item response choices. Lastly, the SAS code provides the algorithms for combining items in scoring the SF-12 physical and mental summary scales.

Table 5.1 presents descriptive statistics for the PCS-12 and MCS-12 summary scales from the test data set. After scoring the test dataset, the means, standard deviations, and minimum and maximum observed values observed should agree with those presented in Table 5.1. Table 5.2 presents correlations between SF-12 items and SF-12 physical and mental summary scales. Correlations between the 12 items and the two summary scales documented in Table 5.1, should be observed in analyses of the test dataset.

Registered users of the SF-12 Health Survey who encounter any problems with scoring PCS-12 and MCS-12 have the option of calling the "SF-12 Hotline" at The Health Institute at (617) 636-8098, or FAX (617) 636-8077.

TABLE 5.1. TEST DATASET DESCRIPTIVE STATISTICS: SF-12 PHYSICAL AND MENTAL SUMMARY MEASURES

	Number of Cases	Mean	Standard Deviation	Minimum Observed Score	Maximum Observed Score
Physical Component Summary (PCS-12)	50	43.9	11.0	18.4	57.8
Mental Component Summary (MCS-12)	50	48.8	10.6	18.7	65.2

TABLE 5.2. TEST DATASET CORRELATIONS: SF-12 ITEMS AND SUMMARY MEASURES (N=50)

SF-12 Items	PCS-12	MCS-12
PF02	.87	.03
PF03	.73	.29
RP2	.56	.24
RP3	.82	.15
BP2 (reversed scored)	.77	.23
GH1 (reversed scored)	.63	.34
VT2 (reversed scored)	.44	.31
SF2	.51	.54
RE2	.14	.78
RE3	-.03	.80
MH3 (reversed scored)	-.14	.85
MH4	.04	.85

❦ 6. Reliability of Scores

This chapter presents reliability estimates for the PCS-12 and MCS-12 scales and summarizes the methods used in calculating those estimates. A reliability coefficient estimates the proportion of variance in a scale score that is real as opposed to random error. Reliability coefficients of 0.70 or greater are generally satisfactory for scales used in group-level analyses and coefficients of 0.90 or greater for scales used in decisions at the individual level (Nunnally and Bernstein, 1994).

By pooling common reliable variance in physical or mental health across questionnaire items, SF-12 summary measures were expected to achieve satisfactory reliability. However, in comparison with most MOS health status scales, which were developed to be highly internally consistent (Stewart and Ware, 1992), SF-12 items are relatively heterogeneous. Each SF-12 item was selected, at least in part, because it contained unique reliable variance of proven value in estimating physical or mental health. A practical implication is that internal consistency estimates of reliability under-estimate the reliability of SF-12 summary measures and are not applicable to single-item measures. As described below, test-retest estimation methods were relied upon.

Data Sources and Methods

Reliability estimates for the PCS-12 and MCS-12 scales were calculated using data from general population surveys in the U.S. (McHorney, Kosinski, and Ware, 1994) and the U.K. (Brazier, Jones, and Kind, 1993). The reliability of the PCS-12 and MCS-12 scales was estimated by correlating scale scores from repeated administrations of the SF-36 two-weeks apart using product-moment correlations between scale scores. Changes in PCS-12 and MCS-12 scale scores were assessed by calculating the percentage of scores from the second administration that remained with the 95% confidence interval (CI) of scores from the first administration and by calculating the average difference in PCS-12 and MCS-12 scale scores between administrations.

Results

Table 6.1 summarizes results for two studies of the reliability of PCS-12 and MCS-12 scale scores. The test-retest reliability for the PCS-12 scale was 0.89 in the U.S. and 0.86 in the U.K. Reliability coefficients of 0.76 and 0.77 were observed for the MCS-12 scale in the U.S. and the U.K., respectively.

Differences in PCS-12 and MCS-12 scale scores between administrations averaged one-point or less in both U.S. and U.K. populations, and greater than 85.3% scored at the second administration within the 95% CI of the scores at the first administration for both PCS-12 and MCS-12.

TABLE 6.1 TEST-RETEST RELIABILITY COEFFICIENTS, AVERAGE CHANGES, AND THE PERCENT SCORING WITHIN THE 95% CONFIDENCE INTERVAL (CI) BETWEEN ADMINISTRATIONS, PCS-12 AND MCS-12 SCALE SCORES IN THE GENERAL U.S. (N=232) AND U.K. (N=187) POPULATIONS

	U.S. Population			U.K. Population		
Scale	Reliability	Average Change	(%) w/in 95% CI	Reliability	Average Change	(%) w/in 95% CI
Physical Summary (PCS)						
SF-12 PCS	0.89	0.94	85.3	0.88	0.22	89.9
Mental Summary (MCS)						
SF-12 MCS	0.76	0.81	85.3	0.78	1.00	87.6

Comment

Although these reliability estimates are almost certain to be biased downward because of changes in health status between test and retest, they exceed accepted standards for measures used in measuring and monitoring health at the group level (Nunnally and Bernstein, 1994).

❦ 7. Equivalence of SF-12 and SF-36 Summary Scales in the General U.S. Population

The equivalence of physical (PCS) and mental (MCS) summary scales estimated from the SF-12 and SF-36 health surveys and their relative validity in discriminating among patient groups known to differ in health status have been evaluated in the MOS. Results are summarized in a forthcoming article (Ware, Kosinski, and Keller, 1996), and detailed tables of results are presented in Chapter 8.

In addition, we have evaluated the equivalence of scores for summary measures based on the SF-12 and SF-36 Health Survey in the general US population and across groups differing in age and sex sampled from that population. Sampling methods used in the National Survey of Functional Health Status, which provided the data analyzed, are documented in the SF-36 user's manuals (Ware, Snow, Kosinski, and Gandek, 1993; Ware, Kosinski, and Keller, 1994).

Descriptive statistics, including means, standard deviations and scores at selected percentiles, for summary measures based on SF-12 and SF-36 Health Surveys were compared across population subgroups to determine whether norms published for SF-36 summary measures (Ware, Kosinski, and Keller, 1994) could be used in interpreting summary scores computed from the SF-12. Results summarized below support the use of norms and other interpretation guidelines published for the SF-36 summary measures (Ware, Kosinski, and Keller, 1994) in interpreting the SF-12 summary measures.

Preliminary Results

Table 7.1 compares descriptive statistics for SF-36 Physical Component Summary (PCS-36) and Mental Component Summary (MCS-36) scores with the summary scores estimated from the SF-12 Health Survey for the general US population (N = 2,329). Tables 7.2 and 7.3 compare the same descriptive statistics for SF-12 and SF-36 summary scores in sub-groups of males (N = 997) and females (N = 1,332). Tables 7.4 through 7.9 present results for six U.S. population groups differing in age (males and females combined).

Comparisons of results for PCS-12 versus PCS-36 and for MCS-12 versus MCS-36 show a very high level of agreement for all descriptive statistics. Across the seventeen total population and subgroup comparisons, Tables

7.1 through 7.9 show that average scores differed by less than one point. Further, estimates of scores at the 25th, 50th, and 75th percentiles for the total population (Table 7.1) and for all 16 subgroups (Tables 7.2 through 7.9) were within a fraction of one point for both the PCS and MCS summary scores estimated from SF-12 and SF-36 surveys. Ranges of scores, which are unstable because they are determined by only two subjects (the highest and lowest), showed less agreement between surveys.

Comment

Although further study of the interchangability of summary measures based on the SF-12 and SF-36 health surveys is required before conclusions are drawn, results to date suggest that the numerous tables of norms and interpretation guidelines published for the SF-36 summary measures (Ware, Kosinski, and Keller, 1994; 1995) can be used in interpreting the SF-12. This conclusion is based on the cross-validation of the algorithms from the general U.S. population used to score the SF-12 physical and mental summary scales in the MOS (Chapter 3) and the tables documenting the equivalence of means and other descriptive statistics in this chapter.

TABLE 7.1. COMPARISON OF SF-36 AND SF-12 NORMS FOR THE GENERAL U.S. POPULATION, TOTAL SAMPLE

Total Sample (N=2,329)		**SF-36 PCS**	**SF-12 PCS**
	Mean	50.12	50.12
	25th Percentile	46.27	46.53
	50th Percentile	53.22	53.55
	75th Percentile	56.82	56.49
	Standard Deviation	9.90	9.45
	Range	8-73	13-69

Total Sample (N=2,329)		**SF-36 MCS**	**SF-12 MCS**
	Mean	50.04	50.04
	25th Percentile	45.36	45.13
	50th Percentile	53.05	52.85
	75th Percentile	57.18	57.30
	Standard Deviation	10.00	9.59
	Range	9-74	10-70

TABLE 7.2. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, MALES

Males (N=997)		SF-36 PCS	SF-12 PCS
	Mean	51.18	51.22
	25th Percentile	48.25	48.79
	50th Percentile	53.95	54.30
	75th Percentile	57.11	56.61
	Standard Deviation	9.28	8.80
	Range	9-73	14-69

Males (N=997)		SF-36 MCS	SF-12 MCS
	Mean	50.73	50.72
	25th Percentile	46.86	46.16
	50th Percentile	53.40	53.53
	75th Percentile	57.43	57.82
	Standard Deviation	9.58	9.31
	Range	13-74	14-70

TABLE 7.3. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, FEMALES

Females (N=1,332)		**SF-36 PCS**	**SF-12 PCS**
	Mean	49.15	49.11
	25th Percentile	44.22	44.32
	50th Percentile	52.52	52.76
	75th Percentile	56.50	56.02
	Standard Deviation	10.35	9.92
	Range	8-69	13-65

Females (N=1,332)		**SF-36 MCS**	**SF-12 MCS**
	Mean	49.41	49.42
	25th Percentile	43.63	43.78
	50th Percentile	52.69	51.94
	75th Percentile	56.80	56.85
	Standard Deviation	10.33	9.80
	Range	9-71	11-70

TABLE 7.4. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, AGE 18-34 YEARS

Age 18-34 (N=636)		SF-36 PCS	SF-12 PCS
	Mean	53.62	53.33
	25th Percentile	51.17	51.56
	50th Percentile	55.20	55.18
	75th Percentile	58.22	57.21
	Standard Deviation	7.30	6.73
	Range	18-73	18-68

Age 18-34 (N=636)		SF-36 MCS	SF-12 MCS
	Mean	48.81	49.18
	25th Percentile	44.29	44.48
	50th Percentile	51.56	51.81
	75th Percentile	56.67	56.43
	Standard Deviation	10.23	9.74
	Range	13-67	11-62

TABLE 7.5. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, AGE 35-44 YEARS

Age 35-44 (N=487)		**SF-36 PCS**	**SF-12 PCS**
	Mean	52.19	52.18
	25th Percentile	49.38	50.22
	50th Percentile	54.11	54.30
	75th Percentile	57.21	56.82
	Standard Deviation	7.73	7.30
	Range	10-67	14-64

Age 35-44 (N=487)		**SF-36 MCS**	**SF-12 MCS**
	Mean	49.96	50.10
	25th Percentile	45.64	45.67
	50th Percentile	52.57	52.24
	75th Percentile	56.66	56.83
	Standard Deviation	9.23	8.62
	Range	18-66	20-65

TABLE 7.6. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, AGE 45-54 YEARS

Age 45-54 (N=324)		SF-36 PCS	SF-12 PCS
	Mean	49.52	49.71
	25th Percentile	45.71	46.54
	50th Percentile	52.78	52.76
	75th Percentile	55.86	56.24
	Standard Deviation	9.69	9.50
	Range	14-67	14-65

Age 45-54 (N=324)		SF-36 MCS	SF-12 MCS
	Mean	50.61	50.45
	25th Percentile	45.85	45.30
	50th Percentile	53.91	53.30
	75th Percentile	57.24	57.83
	Standard Deviation	10.10	9.55
	Range	10-68	18-67

TABLE 7.7. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, AGE 55-64 YEARS

Age 55-64 (N=250)		SF-36 PCS	SF-12 PCS
	Mean	46.12	46.55
	25th Percentile	39.45	41.43
	50th Percentile	50.43	50.22
	75th Percentile	54.23	54.78
	Standard Deviation	11.05	10.63
	Range	13-62	16-63

Age 55-64 (N=250)		SF-36 MCS	SF-12 MCS
	Mean	51.19	50.57
	25th Percentile	47.34	46.39
	50th Percentile	54.37	53.14
	75th Percentile	57.88	57.49
	Standard Deviation	9.56	9.82
	Range	13-65	14-65

TABLE 7.8. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, AGE 65-74 YEARS

Age 65-74 (N=408)		SF-36 PCS	SF-12 PCS
	Mean	43.49	43.65
	25th Percentile	35.05	35.83
	50th Percentile	46.58	46.36
	75th Percentile	52.52	53.18
	Standard Deviation	11.15	11.02
	Range	8-59	13-59

Age 65-74 (N=408)		SF-36 MCS	SF-12 MCS
	Mean	52.59	52.10
	25th Percentile	48.34	47.06
	50th Percentile	55.52	55.31
	75th Percentile	59.13	58.91
	Standard Deviation	9.31	9.53
	Range	21-74	19-70

TABLE 7.9. COMPARISON OF NORMS FOR THE GENERAL U.S. POPULATION, AGE 75+ YEARS

Age 75+ (N=217)		**SF-36 PCS**	**SF-12 PCS**
	Mean	38.04	38.68
	25th Percentile	28.99	29.37
	50th Percentile	38.35	38.68
	75th Percentile	47.53	47.77
	Standard Deviation	11.19	11.04
	Range	13-59	17-57

Age 75+ (N=217)		**SF-36 MCS**	**SF-12 MCS**
	Mean	50.75	50.06
	25th Percentile	41.91	40.48
	50th Percentile	54.36	53.53
	75th Percentile	59.53	58.89
	Standard Deviation	11.71	10.94
	Range	18-71	22-69

❦ 8. Validity

Because of the widespread use of the SF-12 across a variety of populations and purposes, evidence of all types of validity is relevant. The content validity of the SF-12 compares favorably with that of the SF-36 Health Survey (Chapter 2, Table 2.1), which has been shown to represent the health concepts most frequently included in widely-used health measures (Ware, Snow, Kosinski, and Gandek, 1993; Ware, Kosinski, and Keller, 1994; Ware, 1995).

Data Sources and Methods

Data analyzed in testing the validity of the SF-12 came from two sources, which are documented in detail elsewhere. Included were the National Survey of Functional Health Status (NSFHS), which normed the SF-36 Health Survey (McHorney, Kosinski, and Ware, 1994; Ware, Kosinski, and Keller, 1994), and the Medical Outcomes Study (MOS), an observational study of health outcomes for patients with chronic conditions (Tarlov, Ware, Greenfield, et al., 1989; Stewart and Ware, 1992).

Our approach to the empirical validation of the SF-12 followed very closely the logic and methods of previous MOS studies. Tests were designed to closely parallel the intended uses of the SF-12 and to address study design issues that might affect interpretations. We used the method of construct validation referred to as "known groups" validity (Kerlinger, 1973) as in previous MOS studies (McHorney, Ware, and Raczek, 1993; Ware, Kosinski, Bayliss, et al., 1995). For example, the SF-12 measure of physical health (PCS-12) was expected to discriminate between groups of patients who differ in physical condition according to proven clinical measures. The empirical validity of PCS-12 and MCS-12 was compared to the SF-36 summary measures and to the best SF-36 scale. The criteria used in forming patient groups were the same as those reported previously for studies of SF-36 measures and other measures (McHorney, Ware, and Raczek, 1993; Ware, Kosinski, and Keller, 1994; Ware, Kosinski, Bayliss, et al., 1995).

As documented in Appendix B, and briefly in table footnotes in this chapter, "criteria" included: 1) physical condition (e.g., serious vs. minor physical diagnosis); 2) mental condition (e.g., serious mental condition vs. minor medical); 3) the additional impact of a serious physical or mental condition; 4) differences among physical diagnostic groups; 5) severity of disease within a diagnosis; 6) the impact of comorbid conditions; 7) the impact of acute symptoms during the past month; 8) differences in age effects on health based on cross-sectional comparisons;

9) longitudinal comparisons among patients who self-reported an improvement or a worsening in health; and 10) estimates of the impact of clinical depression based on cross-sectional and longitudinal analyses.

PCS-12 and MCS-12 summary scales were expected to differ markedly in their validity on the basis of factor analytic studies of construct validity (McHorney, Ware, and Raczek, 1993; Ware, Kosinski, Bayliss, et al., 1995; Ware, Kosinski, and Keller, 1995). Specifically, MCS-12 was expected to be the best mental measure while the PCS-12 was expected to be the best physical measure (Ware, Kosinski, and Keller, 1996).

We summarize below published and unpublished results from the first studies of the empirical validity of SF-12 scales and summary measures and we compare results for the SF-12 with results for the SF-36 Health Survey. Some results, which are published elsewhere (Ware, Kosinski, and Keller, 1996), are only summarized. Results previously unpublished are detailed in tables at the end of the chapter.

Four-Group Test

We begin with a summary of results from the MOS "four group" test of validity, which tests validity in discriminating among four clinically defined groups of patients differing in the severity of physical and/or mental conditions. The "four group" test of the SF-12 is a replication of empirical tests previously published for the SF-36 scales and summary measures (McHorney, Ware, and Raczek, 1993; Ware, Kosinski, and Keller, 1994; Ware, Kosinski, Bayliss, et al., 1995). The four groups that were compared included patients with (1) only minor medical conditions (e.g., uncomplicated hypertension), (2) a serious physical condition (e.g., congestive heart failure), (3) a mental condition only (e.g., clinical depression), and (4) both serious physical and mental conditions.

For all tests, the validity of PCS-12 and MCS-12 was compared with the validity of summary measures and scales based on the SF-36. These comparisons were based on ratios of F- statistics as in previous MOS studies (McHorney, Ware, and Raczek, 1993; Ware, Kosinski, and Keller, 1994; Ware, Kosinski, Bayliss, et al., 1995). The F- statistic is a ratio of the amount of separation in scores between groups or between assessments over time (validity) relative to the within-group variance (error). The F-statistic is larger when the separation between groups or change over time is larger and/or the error variance is smaller. The RV coefficient for each measure in each test indicates, in proportional terms, its empirical validity relative to the best SF-36 measure.

In the "four group" tests, the PCS-12 reached the same statistical conclusions (e.g., serious worse than minor conditions) and yielded relative validity (RV) coefficients of 0.93 and 0.63 relative to the best SF-36 scale (results reported elsewhere; see Ware, Kosinski, and Keller, 1996). The best SF-36 scales were Physical Functioning and General Health for comparisons involving the seriousness of physical conditions. RV values observed for PCS-12 in these tests were only slightly below those observed for the PCS-36 in the same tests (0.97 and 0.72). As expected, the MCS-12 yielded very low RV coefficients in tests for *physical* differences, as did the MCS-36.

In comparisons between groups known to differ in mental condition, RV coefficients of 1.07 and 0.60 were observed for the MCS-12 in comparisons with the best SF-36 scale, which was the Mental Health scale. Again, results for MCS-12 were very similar to previously published results for MCS-36, which achieved RV coefficients of 1.12 and 0.62 in these tests. Finally, as expected, PCS-12 yielded very low RV coefficients in tests of validity involving groups differing in <u>mental</u> health.

These results suggest that, in comparisons between groups differing in the severity of their physical conditions, the PCS-12 is likely to reach the same statistical conclusion as PCS-36 but with a relative validity that is 5-10% or lower than that achieved by PCS-36. In the "four group" test the MCS-12 performed about the same as the MCS-36 and both were equal to or better than the best SF-36 scale.

Specific Chronic Conditions

Mean scores for groups of MOS patients with four chronic conditions (hypertension, congestive heart failure, recent myocardial infarction, and Type II diabetes) are compared in Table 8.1. This table also presents F-ratios for the comparison of group means as well as RV coefficients for SF-12 and SF-36 measures. As expected for these diagnoses, the physical summary measures performed much better than the mental summary measures, which showed no significant differences among the four groups. Surprisingly, PCS-12 performed slightly better than PCS-36 in these tests (RV = 0.65 vs. 0.59, respectively), relative to the best SF-36 scale (the General Health scale). The ordering of the four groups from hypertension, which scored the best, to congestive heart failure, which scored the worse, was the same for PCS-12, PCS-36, and for the SF-36 scales (PF and GH) that best discriminated among these groups.

These results suggest that summary measures based on the SF-12 are likely to reach the same conclusions as those based on the SF-36 in comparisons among groups differing in the four physical conditions studied. Results also confirm that both SF-12 and SF-36 summary physical measures perform less well than the best SF-36 scale, but that the PCS-12 represents about the same compromise, in this regard, as the PCS-36.

Severity of Disease

Comparisons of groups of patients differing in the severity of disease within hypertension, congestive heart failure, and Type II diabetes are presented in Table 8.2. Consistent with previously reported results for SF-36 scales and summary measures (Ware, Kosinski, Bayliss, et al., 1995), there were no significant differences between groups of hypertension patients differing in severity (defined by diastolic blood pressure; see Appendix B); likewise average scores did not differ for patients differing in the severity of Type II diabetes (defined in terms of duration and complications; see Appendix B). As in previous studies, average scores differed markedly for patients differing in the severity of congestive heart failure, particularly for the SF-12 and SF-36 physical summary measures and SF-36 scales previously shown to be the best measures of physical health. Relative to the best SF-36 scale, PCS-12 had an RV coefficient of .58, which compares with a RV of .68 for PCS-36. The ordering of the two CHF groups differing in severity was the same for PCS-12 and PCS-36 and the two measures led to the same statistical conclusions.

These results suggest that, when the SF-36 detects differences in functional health and well-being due to severity of CHF, PCS-12 is likely to reach the same conclusion about those differences. However, relative validity coefficients favor PCS-36 over PCS-12 by about 10% in these tests. As expected, MCS-12 was much less useful in detecting the impact of CHF, although the differences were significant and MCS-12 performed better than MCS-36.

Comorbid Conditions

Adjusted estimates of the impact of 16 comorbid conditions studied in the MOS for summary measures based on the SF-12 and SF-36 and the eight SF-36 scales are presented in Tables 8.3 and 8.4. PCS-12 achieved an overall RV coefficient of 0.77 in these tests in comparison with an RV of 0.94 for PCS-36. As expected, the MCS summaries performed much less well in discriminating among these comorbidities, which did not include any psychiatric disorders. Not surprisingly, given that many of these conditions involve pain (rheumatoid arthritis, osteoarthritis, back/sciatica, and angina) the SF-36 Bodily Pain scale did best in discriminating the impact of these comorbid conditions. As can be seen in the two right-hand columns of Table 8.3, PCS-12 reached the same statistical conclusion as PCS-36 and SF-12 and SF-36 agreed substantially in estimating the magnitude of differences associated with various comorbid conditions. As shown in Table 8.4, only asthma impacted significantly on mental health; this effect was detected by both MCS-12 and MCS-36.

Acute Symptoms

Table 8 presents correlations between symptoms in five categories and SF-12 and SF-36 scales and summary measures. The pattern of correlations with these symptoms was very similar for summary measures based on the SF-12 and the SF-36. As expected, RV coefficients were much higher for physical summary measures than for mental summaries for the first three symptom groups. In those tests RV coefficients of 0.67, 0.51, and 0.43 were observed for PCS-12, in comparison with RV coefficients of 0.78, 0.60, and 0.55 for PCS-36. For symptoms in the central nervous system cluster, higher correlations were observed for MCS-36 and MCS-12 (RV coefficients 0.82 and 0.67, respectively). Symptoms in the GI/GU cluster correlated highest with the MCS summary measure (0.92 and 0.98 for MCS-36 and MCS-12, respectively). Symptoms in the GI/GU group were not significant for either of the two physical summary measures.

These results suggest that the summary measures based on the SF-12 and SF-36 have very similar associations with acute symptoms. Although SF-12 always reached the same statistical conclusions in the tests documented here, it did so with lower empirical validity than PCS-36. MCS-12 did better than MCS-36 in terms of correlating with symptoms in the GI/GU symptom clusters.

Age Differences

As shown in Table 8.6, age differences were concentrated in the physical component of health status, as hypothesized. Relative to the best SF-36 scale (PF scale), the PCS-12 and PCS-36 summary measures achieved RV coefficients of 0.78 and 0.71 respectively. Average scores across the three age groups were very similar for PCS-12 and PCS-36 and they reached the same statistical conclusion about age differences. In contrast, smaller but significantly more favorable mental health scores were observed for older age groups in both the MCS-12 and MCS-36 summary measures. These mental health summary measures also reached the same statistical conclusion.

Self-Evaluated Health Transitions

To test the validity of SF-12 summary measures in discriminating among patients who reported different changes in physical and mental health over a one-year period, changes in SF-12 and SF-36 scores were compared for five groups of patients who reported and improvement or worsening in health status. Average changes in SF-12 and SF-36 scores for five groups differing in self-evaluations of changes in physical and mental health are summarized in Table 8.7. Additional results are reported elsewhere (Ware, Kosinski, and Keller, 1996).

As expected, changes in PCS-12 were more responsive than MCS-12 in comparisons of groups who reported changes in physical health status (RV = 0.73 and 0.11, respectively). Likewise, MCS-12 change scores did better than PCS-12 change scores in discriminating among those who reported changes in mental health status (RV = 0.93 and 0.08, respectively). In the same tests, summary measures based on the SF-36 reached the same statistical conclusions as the SF-12 and performed better (by about 5-10%) than the SF-12, as documented in the table.

Impact of Clinical Depression

As documented in Table 8.8, the SF-36 Mental Health (MH) scale was the best scale in discriminating between patients with and without clinical depression. Relative to the MH scale, MCS-12 and MCS-36 were the best of the summary measures in discriminating between those groups (RV = 0.98 and 1.03, respectively). As expected, patients with and without clinical depression differed little if at all in scores for the summary physical health measures.

A similar pattern of results was observed in analyses of MOS patients who recovered from clinical depression over a two-year period (Table 8.9). Consistent with results from cross-sectional tests, the MH scale was the best SF-36 scale in responding to clinical improvement over time. As expected, the MCS-12 and MCS-36 measures were better than the physical summary measures in detecting an improvement (RV = 0.91 and

1.38, respectively). Physical summaries based on SF-12 and SF-36 showed no differences.

Summary of Results

As documented and discussed elsewhere, (Ware, Kosinski, and Keller, 1996) 16 tests of validity have been performed for the SF-12 to date and results for all tests have been compared with results for summary measures and the eight scales based on the SF-36 (a total of 192 RV coefficients). In the 12 tests based on criterion variables defining differences in physical health, statistical conclusions based on PCS-12 agreed with PCS-36 consistently and agreed with the three best SF-36 physical scales (PF, RP, BP) 30 out of 36 times. RV coefficients for PCS-12 ranged from 0.43 to 0.78, median = 0.67 in these tests. PCS-36 performed better than PCS-12 in 11 of 12 tests.

Conclusions based on MCS-12 agreed with those based on MCS-36 in all four comparisons between groups differing in the presence and severity of mental health conditions. In these mental health tests, RV coefficients for MCS-12 ranged from 0.93 to 0.98 and were below those for MCS-36 in three of four tests.

We discuss and offer recommendations based on these results in Chapter 9. Additional information and discussion are published elsewhere (Ware, Kosinski, and Keller, 1996).

TABLE 8.1: COMPARISONS OF ADJUSTED SF-36 AND SF-12 MEAN SCORES FOR FOUR CHRONIC CONDITIONS

	SF-36 Scales[a]								SF-36 Summary Scales		SF-12 Summary Scales	
	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS	PCS	MCS
Hypertension	77.34 (0.9)	65.35 (1.6)	74.71 (1.1)	66.19 (0.8)	60.43 (0.9)	90.40 (0.7)	80.08 (1.4)	79.85 (0.6)	45.64 (0.4)	53.31 (0.3)	46.47 (0.4)	52.99 (0.4)
Congestive Heart Failure[a]	58.14 (3.2)	46.58 (4.5)	70.85 (3.5)	50.17 (2.3)	47.83 (2.6)	78.14 (3.8)	67.50 (4.1)	79.40 (1.9)	38.29 (1.3)	51.53 (1.1)	40.02 (1.2)	51.15 (1.1)
Myocardial Infarction, Recent	69.71 (3.1)	53.62 (5.5)	73.38 (2.9)	58.19 (2.9)	56.78 (2.5)	85.57 (2.7)	74.02 (5.6)	75.39 (1.6)	42.69 (1.5)	51.64 (1.2)	42.34 (1.6)	51.52 (1.0)
Type II Diabetes[a]	73.23 (2.0)	62.44 (3.8)	73.58 (2.3)	59.09 (1.9)	58.78 (1.8)	86.17 (2.0)	80.17 (2.8)	78.45 (1.5)	43.69 (1.0)	52.87 (0.8)	44.84 (0.9)	52.49 (0.8)
F for Four Conditions	17.22***	8.11***	2.77*	22.71***	7.44***	6.41***	2.84*	2.48*	13.31***	1.14	14.76***	1.14
RV[b]	.76	.36	.12	**1.00**	.33	.28	.12	.11	.59	-	.65	-
Adjusted R^2 (Four Conditions)	0.253	0.115	0.072	0.087	0.102	0.067	0.042	0.089	0.189	0.078	0.199	0.080
N	1238	1238	1238	1238	1238	1238	1238	1238	1238	1238	1238	1238

*** $p < 0.001$, two-tailed test
** $p < 0.01$, two-tailed test
* $p < 0.05$, two-tailed test

[a] Mean scores were estimated from linear regression models that controlled for demographics and MOS design variables.

[b] RV = Relative Validity (not reported for non-significant F's). For each test, the "best" of the eight SF-36 scales (with highest F-ratio) is labeled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.2: COMPARISON OF ADJUSTED SF-36 AND SF-12 MEAN SCORES FOR SEVERITY GROUPS

	SF-36 Scales[a]								SF-36 Summary Scales		SF-12 Summary Scales	
	PF	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**	**PCS**	**MCS**	**PCS**	**MCS**
Hypertension												
Severity 1	76.54	63.56	74.18	65.98	60.32	90.09	79.01	79.61	45.32	53.22	46.08	52.98
	(0.9)	(1.7)	(1.2)	(0.9)	(1.0)	(0.8)	(1.7)	(0.7)	(0.4)	(0.4)	(0.4)	(0.4)
Severity 2	81.15	75.98	77.68	67.18	60.62	92.05	86.39	81.30	47.42	53.34	48.65	52.95
	(3.0)	(5.3)	(3.5)	(3.0)	(2.0)	(1.9)	(3.5)	(1.9)	(1.4)	(0.7)	(1.3)	(0.9)
F for Severity	2.30	4.76*	0.87	0.14	0.02	0.84	3.15	0.62	1.98	0.45	3.30	0.00
RV[b]	-	**1.00**	-	-	-	-	-	-	-	-	-	-
Congestive Heart Failure												
Severity 1	66.79	57.74	75.12	56.68	53.89	83.92	78.72	82.50	41.49	53.72	42.78	53.48
	(3.9)	(5.8)	(4.3)	(3.1)	(3.3)	(4.3)	(5.2)	(1.8)	(1.8)	(1.2)	(1.5)	(1.2)
Severity 2	45.28	30.58	64.41	40.67	38.47	69.78	51.32	74.62	33.67	48.24	36.27	47.56
	(3.5)	(4.9)	(5.1)	(3.1)	(3.1)	(6.2)	(6.0)	(3.7)	(1.4)	(2.1)	(1.4)	(1.8)
F for Severity	18.65***	15.63***	2.86	13.91***	13.35***	3.66	10.72***	3.92*	12.60***	5.03*	10.91***	7.22**
RV[b]	**1.00**	.84	-	.74	.72	-	.57	.21	.68	.27	.58	.39
Type II Diabetes												
Severity 1	74.98	64.59	73.62	60.79	59.12	87.48	83.63	78.44	44.28	53.23	45.67	52.58
	(2.9)	(4.9)	(2.8)	(2.5)	(2.1)	(2.7)	(4.0)	(2.0)	(1.3)	(1.1)	(1.2)	(1.1)
Severity 2	75.11	65.40	77.15	59.01	60.38	87.48	73.51	77.39	45.34	51.61	46.01	51.27
	(2.8)	(5.5)	(3.1)	(3.3)	(3.4)	(2.8)	(4.7)	(2.6)	(1.3)	(1.4)	(1.3)	(1.3)
Severity 3	72.86	64.90	74.69	59.13	59.45	85.43	86.76	82.22	43.22	54.62	44.43	55.05
	(4.9)	(8.5)	(5.3)	(5.7)	(3.2)	(4.2)	(6.1)	(3.0)	(2.3)	(1.3)	(2.3)	(1.4)
Severity 4	63.24	50.63	66.48	52.39	54.28	78.26	72.73	76.28	39.52	51.62	41.00	51.03
	(4.2)	(7.4)	(5.5)	(2.9)	(3.7)	(4.7)	(5.7)	(2.3)	(1.9)	(1.3)	(1.7)	(1.2)
F for Severity	2.50	1.55	1.23	2.25	0.79	1.43	1.89	1.28	2.55	1.39	2.43	2.12
RV[b]	-	-	-	-	-	-	-	-	-	-	-	-
F for Severity Overall	5.24***	4.16***	1.46	3.25**	2.98**	2.22*	2.89**	1.52	3.69***	1.43	3.31**	1.78
RV[b]	**1.00**	.79	-	.62	.57	.42	.55	-	.70	-	.63	-
Adjusted R^2 (Severity)	0.272	0.132	0.077	0.100	0.109	0.800	0.060	0.091	0.205	0.083	0.215	0.088
N	1238	1238	1238	1238	1238	1238	1238	1238	1238	1238	1238	1238

*** $p < 0.001$, two-tailed test

** $p < 0.01$, two-tailed test

* $p < 0.05$, two-tailed test

[a] Mean scores were estimated from linear regression models that controlled for demographics and MOS design variables.

[b] RV = Relative Validity (not reported for non-significant F's). For each test the "best" of the eight SF-36 scales (with highest F-ratio) is labelled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.3. COMPARISONS OF DIFFERENCES IN SF-36 AND SF-12 SCORES FOR PATIENTS WITH COMORBIDITIES VERSUS PATIENTS WITHOUT COMORBIDITIES

	SF-36 Scales[a]								SF-36 Summary		SF-12 Summary	
	PF		RP		BP		GH		PCS		PCS	
Asthma	-11.55	(6.6)	-14.68	(8.0)	-6.20	(6.0)	-4.84	(5.8)	-2.01	(2.9)	-2.69	(2.6)
C.O.P.D.	-4.76	(3.3)	-9.00	(5.7)	-6.18*	(3.0)	-10.95***	(2.3)	-3.51*	(1.5)	-3.63**	(1.4)
Angina (no MI)	-6.06**	(2.1)	-15.00***	(3.5)	-6.70**	(2.2)	-8.01***	(1.9)	-4.22***	(.9)	-4.43***	(.9)
Past MI	-8.54**	(2.8)	-3.74	(4.3)	-3.40	(3.0)	-5.03*	(2.3)	-2.61*	(1.1)	-2.16	(1.2)
Other Lung	2.31	(4.6)	4.82	(6.9)	3.43	(5.3)	.32	(6.2)	2.45	(2.8)	2.56	(2.4)
Back/Sciatica	-5.49**	(1.8)	-14.18***	(3.2)	-14.60***	(2.0)	-3.38**	(1.3)	-4.41***	(.8)	-3.09***	(.8)
Hip Impairment	-14.92***	(4.0)	-6.32	(6.5)	-9.22**	(3.0)	-10.69***	(3.0)	-5.02**	(1.6)	-4.64*	(1.7)
Rheumatoid Arth.	-7.84	(6.7)	-17.45*	(7.2)	-15.59**	(5.0)	-5.55	(4.3)	-6.11**	(2.2)	-5.36*	(2.5)
Osteoarthritis	-8.90*	(3.7)	-14.14**	(5.4)	-14.22***	(3.3)	-2.25	(2.4)	-5.22***	(1.5)	-4.23***	(1.3)
Musculoskeletal Complaints	-4.98**	(1.9)	-8.71*	(3.5)	-6.44**	(1.9)	-3.59	(1.9)	-2.71***	(.8)	-1.97*	(.8)
Irritable Bowel	-5.93	(3.4)	-5.13	(5.7)	-4.33	(3.5)	-6.29	(3.7)	-2.71*	(1.4)	-2.65*	(1.3)
Ulcers	-12.61*	(6.1)	-13.53	(7.2)	-7.85	(4.4)	-9.36*	(4.0)	-5.12	(3.0)	-4.63	(2.4)
Kidney Disease	-6.22	(6.4)	-16.26	(13.2)	-12.50	(7.9)	-6.01	(6.5)	-4.46	(3.4)	-5.01	(3.2)
U.T.I.	-.64	(2.3)	-.89	(4.8)	-3.84	(2.9)	-1.51	(2.8)	-.91	(1.2)	-.62	(1.3)
Dermatitis	3.72*	(1.7)	-6.40	(4.2)	-1.96	(2.6)	-2.13	(1.7)	-.22	(.8)	-.14	(.9)
Anemia	-4.53	(4.5)	-5.57	(6.9)	-3.34	(3.7)	-8.23*	(3.3)	-3.18	(2.0)	-2.75	(1.9)
Intercept	86.64***	(2.5)	75.63***	(4.3)	89.35***	(2.9)	71.96***	(2.3)	50.77***	(1.1)	51.00***	(1.0)
F for Significance of Comorbidities	7.94***		7.30***		13.66***		9.58***		12.88***		10.55***	
RV[b]	.58		.53		**1.00**		.70		.94		.77	
Adjusted R^2	0.3529		0.2188		0.2513		0.2017		0.3437		0.3273	
	n = 1238		n = 1238		n = 1238		n = 1238		n = 1238		n = 1238	

*** p < 0.001

** p < 0.01

* p < 0.05

[a] Differences in scores were estimated from linear regression models that controlled for demographics, diagnosis, and MOS design variables.

[b] RV = Relative Validity (not reported for non-significant F's). For each test, the "best" of the eight SF-36 scales (with highest F-ratio) is labeled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.4. COMPARISONS OF DIFFERENCES IN SF-36 AND SF-12 SCORES FOR PATIENTS WITH COMORBIDITIES VERSUS PATIENTS WITHOUT COMORBIDITIES

	SF-36 Scales[a]								SF-36 Summary		SF-12 Summary	
	VT		SF		RE		MH		MCS		MCS	
Asthma	-8.69	(5.1)	-2.15	(5.5)	-41.39***	(10.7)	-5.65	(4.2)	-6.26*	(2.8)	-5.91*	(2.5)
C.O.P.D.	-4.06	(2.2)	-5.25	(3.6)	-.42	(4.0)	-4.32	(2.3)	-1.16	(1.2)	-.24	(1.0)
Angina (no MI)	-8.82***	(2.1)	-1.74	(1.9)	-4.28	(3.4)	-1.82	(1.7)	-.76	(.9)	-.54	(.8)
Past MI	-2.58	(2.2)	-5.54*	(2.7)	-4.76	(5.3)	-1.32	(2.6)	-.78	(1.5)	-.82	(1.5)
Other Lung	12.34*	(5.2)	-1.41	(3.7)	-.36	(10.9)	-5.80	(5.4)	-.98	(2.1)	-.53	(2.0)
Back/Sciatica	-5.35**	(1.7)	-6.09***	(1.6)	-3.66	(3.2)	-2.64*	(1.3)	-.83	(.8)	-1.17	(.7)
Hip Impairment	-6.66*	(3.3)	-11.08*	(4.5)	-10.99	(6.8)	-2.45	(3.0)	-2.03	(1.9)	-2.48	(2.0)
Rheumatoid Arth.	2.29	(3.8)	-3.16	(3.7)	.06	(8.8)	1.95	(4.8)	2.49	(2.2)	2.55	(2.6)
Osteoarthritis	-5.30	(3.5)	-3.19	(2.9)	2.36	(5.4)	-2.40	(2.3)	.69	(1.3)	.28	(1.0)
Musculoskeletal Complaints	-2.99	(1.9)	-.20	(1.5)	-.76	(3.4)	-3.54*	(1.5)	-.32	(.8)	-.68	(.8)
Irritable Bowel	-5.10	(3.3)	-6.01*	(3.0)	-2.49	(6.8)	-1.57	(2.7)	-1.04	(1.5)	-1.02	(1.4)
Ulcers	-6.36	(4.3)	-6.79	(4.4)	-6.69	(7.2)	-3.85	(3.7)	-1.31	(2.0)	-1.74	(1.8)
Kidney Disease	-7.30	(4.3)	-13.83	(7.3)	.14	(14.9)	-7.12	(7.4)	-2.61	(4.5)	-2.92	(4.2)
U.T.I.	-8.44**	(2.8)	2.75	(2.4)	-1.43	(6.3)	-.46	(2.4)	-.66	(1.5)	-.46	(1.4)
Dermatitis	-1.53	(1.8)	-.08	(2.1)	-1.12	(3.2)	-2.06	(2.0)	-.91	(.8)	-1.04	(.7)
Anemia	-4.09	(3.9)	-.64	(4.3)	5.13	(4.1)	-.36	(3.3)	.90	(1.6)	.42	(1.6)
Intercept	66.10***	(2.2)	95.71***	(2.1)	80.01***	(3.6)	82.54***	(2.0)	53.40***	(1.1)	52.72***	(1.0)
F for Significance of Comorbidities	6.64***		4.54***		2.12**		3.41***		1.87*		1.89*	
RV[b]	.48		.33		.16		.25		.14		.14	
Adjusted R^2	0.1925		0.1414		0.0852		0.1218		0.0974		0.1044	
	n = 1238		n = 1238		n = 1238		n = 1238		n = 1238		n = 1238	

*** $p < 0.001$

** $p < 0.01$

* $p < 0.05$

[a] Differences in scores were estimated from linear regression models that controlled for demographics, diagnosis, and MOS design variables.

[b] RV = Relative Validity (not reported for non-significant F's). For each test, the "best" of the eight SF-36 scales (with highest F-ratio) is labeled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.5. CORRELATIONS BETWEEN SYMPTOMS AND SF-36 AND SF-12 SCALES (N=1,250)

			SF-36 Scales								SF-36 Summary Scales		SF-12 Summary Scales	
Symptoms[a]	**mean**	**sd**	**PF**	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**	**PCS**	**MCS**	**PCS**	**MCS**
Ears, Nose, and Throat														
Blurred Vision	1.53	1.00	-.32	-.34	-.30	-.31	-.31	-.30	-.19	-.23	-.33	-.19	-.32	-.20
Dry Mouth	2.11	1.26	-.39	-.38	-.36	-.35	-.38	-.32	-.23	-.32	-.38	-.23	-.39	-.23
Lump in throat	1.29	0.73	-.20	-.24	-.25	-.22	-.29	-.24	-.22	-.31	-.19	-.26	-.21	-.25
F for Significance			9.8***	3.1*	0.9	4.1**	1.8	2.8*	0.6	9.5***	7.7***	4.2**	6.6***	2.4
RV[b]			**1.00**	.32	-	.42	-	.28	-	.97	.78	.43	.67	-
Under Central Nervous System Control														
Fainting or passing out	1.03	0.25	-.10	-.11	-.08	-.08	-.11	-.13	-.14	-.12	-.08	-.13	-.08	-.12
Shortness of breath (lying down)	1.37	0.89	-.43	-.37	-.35	-.36	-.40	-.33	-.20	-.26	-.41	-.19	-.40	-.19
Feeling drowsy or sedated	1.89	1.10	-.34	-.43	-.38	-.40	-.53	-.42	-.35	-.42	-.36	-.40	-.35	-.39
Feeling dizzy when standing up	1.64	0.93	-.23	-.33	-.31	-.28	-.39	-.31	-.26	-.35	-.26	-.31	-.27	-.30
Chest pain relieved by nitroglycerin	1.30	0.78	-.35	-.30	-.24	-.31	-.27	-.23	-.16	-.13	-.34	-.10	-.34	-.10
Heart pounding or palpitations	1.48	0.86	-.30	-.32	-.33	-.31	-.36	-.29	-.27	-.34	-.29	-.28	-.31	-.27
Headaches more than usual	1.65	1.01	-.16	-.25	-.34	-.25	-.37	-.36	-.30	-.44	-.18	-.40	-.19	-.38
F for Significance			23.9***	12.8***	7.1***	12.3***	33.0***	15.6***	11.3***	27.5***	19.8***	27.2***	16.7***	22.1***
RV[b]			.72	.39	.21	.37	**1.00**	.47	.34	.83	.60	.82	.51	.67
Musculoskeletal/Extremities														
Backaches or lower back pains	2.49	1.38	-.36	-.38	-.53	-.30	-.35	-.29	-.22	-.23	-.42	-.17	-.43	-.17
Pins and needles in your feet	1.74	1.16	-.34	-.35	-.39	-.34	-.32	-.26	-.19	-.22	-.38	-.16	-.37	-.16
Heavy Feeling in arms and legs	1.50	0.95	-.39	-.41	-.46	-.36	-.42	-.40	-.32	-.32	-.40	-.29	-.40	-.30
Stiffness, pain in muscles	2.96	1.36	-.46	-.48	-.65	-.38	-.42	-.33	-.21	-.21	-.56	-.13	-.54	-.16
F for Significance			48.5***	46.4***	175.3***	19.4***	22.2***	14.7***	5.9***	2.5*	96.2***	6.2***	74.7***	4.7***
RV[b]			.28	.26	**1.00**	.11	.13	.08	.03	.01	.55	.03	.43	.03
GI/GU														
Acid indigestion after meals	2.20	1.18	-.23	-.28	-.34	-.28	-.34	-.25	-.23	-.27	-.27	-.24	-.26	-.25
Trouble passing urine	1.23	0.68	-.13	-.20	-.16	-.18	-.18	-.16	-.18	-.20	-.15	-.18	-.15	-.17
Nausea (upset stomach)	1.58	0.94	-.21	-.31	-.35	-.30	-.33	-.35	-.30	-.39	-.24	-.35	-.24	-.35
F for Significance			3.3*	0.8	3.2*	1.5	1.3	5.6***	3.3*	8.5***	0.7	7.8***	0.8	8.3***
RV[b]			.39	-	.39	-	-	.66	.39	**1.00**	-	.92	-	.98
Other														
Waking early, unable to go back to sleep	2.22	1.23	-.36	-.40	-.34	-.33	-.38	-.31	-.30	-.33	-.35	-.27	-.36	-.27
Coughing producing sputum	1.83	1.20	-.22	-.28	-.26	-.29	-.27	-.23	-.16	-.20	-.27	-.16	-.27	-.15
F for All Symptoms			48.8***	49.5***	84.2***	34.6***	54.9***	32.9***	19.6***	36.3***	61.7***	28.1***	57.6***	25.1***
RV[b]			.58	.59	**1.00**	.41	.65	.39	.23	.43	.73	.33	.68	.30
Adjusted R^2			0.42	0.42	0.56	0.34	0.45	0.33	0.22	0.35	0.48	0.29	0.45	0.27

[a] Reported frequency in the past 4 weeks scored as followed: 1=never, 2=once or twice, 3=a few times, 4=fairly often, 5=very often

[b] The F-statistics and RV's summarized here and in the text are based on comparison of means for the symptom clusters. Those statistics are summarized as correlations coefficients.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.6. COMPARISONS OF CROSS SECTIONAL AGE RELATED DIFFERENCES IN SF-36 AND SF-12 SCORES: UNCOMPLICATED HYPERTENSION[a]

	Age 18-44 (n=161)	**Age 45-64 (n=397)**	**Age 65+ (n=269)**	**(F)**	**RV[b]**
Best SF-36 Scale	90.1	81.0	70.8	46.2***	**1.00**
(Physical Functioning)	(1.2)	(1.0)	(1.4)		
PCS-36	50.2	46.4	42.6	32.9***	0.71
	(0.6)	(0.5)	(0.6)		
MCS-36	51.6	54.0	55.9	15.1***	0.33
	(0.6)	(0.4)	(0.5)		
PCS-12	50.9	47.5	43.4	35.9***	0.78
	(0.6)	(0.5)	(0.6)		
MCS-12	51.2	53.5	55.2	13.6***	0.29
	(0.6)	(0.4)	(0.4)		

*** p < .001
** p < .01
* p < .05

[a] Uncomplicated Hypertension is defined as patients with hypertension and classified as "Minor Medical" in previous sickgroup comparisons.

[b] RV = Relative Validity (not reported for non-significant F's). For each test the "best" of the eight SF-36 scales (with highest F-ratio) is labelled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.7. COMPARISONS OF SF-36 AND SF-12 ONE-YEAR DIFFERENCE SCORES BY SELF-REPORTED PHYSICAL AND MENTAL TRANSITIONS

		SF-36 Scales																SF-36 Summary Scales				SF-12 Summary Scales			
		PF		RP		BP		GH		VT		SF		RE		MH		PCS		MCS		PCS		MCS	
	N	mean	se	mean	se	mean	se	mean	se	mean	se	mean	se	mean	se	mean	se	mean	se	mean	se	mean	se	mean	se
Physical																									
Lot more	79	-11.7	2.6	-13.6	4.0	-12.6	4.1	-13.8	2.3	-3.4	2.4	-13.9	3.5	-0.4	5.0	-1.8	1.9	-7.1	1.3	-0.2	1.1	-6.1	1.3	-2.0	1.2
Some more	162	-8.9	1.4	-7.9	3.4	-3.4	2.5	-10.9	1.3	-4.6	1.4	-6.5	2.0	0.8	3.8	-1.6	1.2	-4.1	0.7	-0.3	0.8	-4.5	0.7	-0.7	0.8
Same	737	-0.0	0.5	5.0	1.3	1.5	0.9	-2.0	0.6	2.4	0.6	-0.3	0.7	5.2	1.4	1.7	0.5	0.0	0.3	1.1	0.3	0.0	0.3	0.5	0.3
Some less	124	2.2	1.7	3.0	3.6	2.4	2.5	-4.0	1.5	2.3	1.7	1.1	2.0	-2.6	3.7	2.0	1.3	0.1	0.8	0.3	0.7	0.5	0.8	-0.7	0.8
Lot less	131	11.8	1.8	19.3	3.6	7.6	2.4	5.9	1.5	8.3	1.8	10.8	2.1	16.8	3.9	7.4	1.6	4.0	0.8	4.2	1.0	4.3	0.8	3.0	0.9
F		39.7***		14.4***		8.1***		28.5***		11.5***		17.8***		4.6***		8.0***		32.1***		4.9***		28.9***		4.5**	
RV[a]		**1.00**		0.36		0.20		0.72		0.29		0.45		0.11		0.20		0.81		0.12		0.73		0.11	
Mental																									
Lot more	44	-4.2	4.0	-13.1	6.6	-8.6	5.0	-15.7	3.4	-10.3	3.9	-18.7	4.1	-15.9	6.4	-12.5	3.1	-2.8	1.7	-7.5	1.7	-3.6	1.8	-8.3	1.6
Some more	120	-7.1	1.7	-7.1	3.7	-4.8	2.9	-9.5	1.7	-5.6	1.7	-11.3	2.3	-13.1	4.4	-6.0	1.2	-2.5	0.9	-4.1	0.8	-2.7	1.0	-5.0	0.9
Same	612	-0.4	0.6	4.2	1.5	0.7	1.0	-4.1	0.6	1.1	0.6	-1.0	0.8	1.6	1.4	0.2	0.4	-0.2	0.3	0.2	0.3	0.0	0.3	-0.6	0.3
Some less	200	0.7	1.2	6.2	2.7	4.4	2.0	-0.2	1.1	4.3	1.3	1.1	1.6	12.5	2.9	3.4	1.0	0.0	0.5	2.8	0.7	-0.2	0.6	2.1	0.7
Lot less	223	2.9	1.3	6.9	2.7	3.0	1.9	2.0	1.3	7.7	1.4	8.1	1.7	20.2	2.9	10.8	1.2	-0.6	0.7	6.7	0.7	-0.3	0.7	5.7	0.8
F		7.4***		5.2***		3.7**		17.5***		17.4***		23.4***		21.0***		45.4***		2.8*		48.3***		3.6**		42.2***	
RV[a]		0.16		0.11		0.08		0.38		0.38		0.51		0.46		**1.00**		0.06		1.06		0.08		0.93	

[a] RV = Relative Validity (not reported for non-significant F's). For each test, the "best" of the eight SF-36 scales (with highest F-ratio) is labelled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.8. COMPARISON OF SF-36 AND SF-12 SCORES FOR CROSS-SECTIONAL TESTS IN DETECTING DIFFERENCES IN MENTAL HEALTH BETWEEN PATIENTS WITH CLINICAL DEPRESSION AND PATIENTS WITH MINOR MEDICAL CONDITIONS

Measures	Clinical Depression (n=242)	Minor Medical (n=898)	Mean Diff. (S.E.)	F	RV[a]
Best SF-36 Scale	46.06	81.75	-35.69	950.49***	**1.00**
(Mental Health)	(1.3)	(0.5)	(1.4)		
PCS-36	47.86	46.17	1.69	5.24*	0.01
	(0.7)	(0.3)	(0.8)		
MCS-36	33.42	54.16	-20.75	981.57***	1.03
	(0.8)	(0.3)	(0.8)		
PCS-12	47.88	47.10	0.78	1.21	0.00
	(0.7)	(0.3)	(0.8)		
MCS-12	34.42	53.62	-19.20	929.64***	0.98
	(0.7)	(0.3)	(0.8)		

*** p < .001
** p < .01
* p < .05

[a] RV = Relative Validity (not reported for non-significant F's). For each test, the "best" of the eight SF-36 scales (with highest F-ratio) is labelled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

TABLE 8.9. COMPARISONS OF SF-36 AND SF-12 SCORES FOR LONGITUDINAL TESTS IN DIFFERENCES IN MENTAL HEALTH AFTER RECOVERY FROM CLINICAL DEPRESSION (N=78)

Measures	Baseline	Follow-Up[a]	Average Change	F	RV[b]
Best SF-36 Scale (Mental Health)	52.08 (2.2)	68.70 (1.9)	16.62 (2.4)	46.38***	**1.00**
PCS-36	50.93 (1.1)	49.81 (1.2)	-1.12 (1.2)	0.85	0.02
MCS-36	36.63 (1.5)	48.02 (1.0)	11.39 (1.4)	64.00***	1.38
PCS-12	50.69 (1.1)	49.96 (1.2)	-0.73 (1.2)	0.37	0.00
MCS-12	37.46 (1.4)	46.91 (1.0)	9.45 (1.4)	42.38***	0.91

*** $p < .001$
** $p < .01$
* $p < .05$

[a] Follow-up scores were obtained two years after baseline (n=78).

[b] RV = Relative Validity (not reported for non-significant F's). For each test, the "best" of the eight SF-36 scales (with highest F-ratio) is labelled RV=1.00 and is boldfaced.

PF = Physical Functioning; **RP** = Role Physical; **BP** = Bodily Pain; **GH** = General Health; **VT** = Vitality; **SF** = Social Functioning; **RE** = Role Emotional; **MH** = Mental Health; **PCS** = Physical Component Summary; **MCS** = Mental Component Summary.

❦ 9. Comments

Experience to date with the SF-12 Health Survey suggests that we have accomplished at least two of our objectives: a) reproduction of physical and mental health summary measures based on the SF-36 with an accuracy of 90% or better; and b) the practical objective of a health survey that can be printed on a single-page scannable form and that can be administered in two minutes or less. As documented and discussed elsewhere (Ware, Kosinski, and Keller, 1996), ongoing studies are evaluating scoring algorithms for SF-12 items to accurately reproduce average scores for the eight-scale SF-36 health profile in large-group studies.

The SF-12 Health Survey represents a calculated compromise between practicality, which is essential for widespread use, comprehensiveness of content (i.e. content validity), and such psychometric considerations as the statistical precision of scores. The 12 items chosen include one or more items from each of the eight health concepts in the SF-36 and provide a representative sample of the various indicators used to operationally define those concepts, including what respondents are able to do, how they feel in terms of both distress and well-being, how their everyday lives are affected, and how they evaluate their health status.

Because of the high degree of correspondence between summary health measures estimated using the SF-12 and SF-36 health surveys, it appears that general population norms and other interpretation guidelines that have been extensively documented for the SF-36 in the *Summary Measures User's Manual* (Ware, Kosinski, and Keller, 1994), will be useful in interpreting the SF-12. Evidence to date suggests that results will be interchangeable for both general and specific patient populations. Interpretation guidelines in the *Summary Measures Manual* include cross-sectional norms and norms for one-year change scores in general and specific populations, and content-related interpretation guidelines. Criterion-based guidelines for the two summary measures include predictions of in-patient and out-patient utilization of health care services, subsequent job loss due to health problems among employed adults, five-year survival probabilities at various levels of scale scores, and cut-off scores for screening for psychiatric and physical conditions.

Advantages and Disadvantages

An advantage of the SF-12 is that its items are a subset of the SF-36 Health Survey, which is widely used in a variety of applications. Because of this overlap, scoring algorithms documented here can be used to achieve comparability between data sets including SF-12 or SF-36 health surveys. This overlap will also accelerate the development of translations of the SF-12. Because SF-12 items were translated along with the SF-36 by the IQOLA Project in over 30 countries, normative data and other interpretation guidelines useful in multinational studies are rapidly becoming available for the SF-12 Health Survey. Translations are available in five non-English languages commonly spoken in the U.S. (Spanish, French, German, Italian, and Japanese). Translations in Chinese, Korean, and Vietnamese, three of the fastest growing non-English speaking populations in the U.S., are currently being evaluated (Ware, Gandek, Keller, et al., 1995).

Because of the comparability of scores for SF-12 and SF-36 physical and mental health summary measures, information useful in their interpretation is likely to accumulate more rapidly. Applications and datasets include: norms for general and specific populations, benchmarks for use in comparing the burden of chronic conditions, algorithms for use in patient screening, algorithms for use in predicting utilization of health care services, algorithms for use in severity adjustment when comparing health outcomes, and estimates of the physical and mental health benefits of various treatments.

Although the SF-12 yields efficient estimates of physical and mental health, its brevity was achieved by eliminating all but 1-2 items for each of eight health concepts. Scales this short have been shown to have significantly less precision than well-constructed multi-item scales (McHorney, Ware, Rogers, et al., 1992). However, empirical studies of these tradeoffs to date suggest that the SF-12 summary measures are a practical alternative to the SF-36 and that the SF-12 will rarely miss differences or changes in physical or mental health status captured by the SF-36 summary measures.

Like the SF-36 Health Survey, the SF-12 offers the option of analyzing physical and mental health summary measures, which reduces the number of statistical comparisons, or analyzing an 8-concept health profile. As illustrated and discussed elsewhere (Ware, Kosinski, and Keller, 1994; Ware, Kosinski, Bayliss, et al., 1995), these options each have advantages and disadvantages in the interpretation and presentation of results.

A disadvantage of the SF-12 physical and mental health summary scores is that their empirical validity typically has fallen about 10% below that observed for the SF-36 physical and mental health summary measures in studies to date (see Chapter 8 and Ware, Kosinski, and Keller, 1996). The reason is that the SF-12 items achieve less replication and define fewer scale levels. Therefore, the SF-12 yields less precise estimates of individual health scores. In group comparisons reported to date, SF-12 and SF-36 summary measures have reached the same statistical conclusions about group differences. For very large group comparisons and longitudinal monitoring efforts, differences in measurement reliability between SF-12 and SF-36 are not as important. For purposes of such applications, the trade-off between reduced questionnaire length and reduced precision is likely to prove to be a good tradeoff.

Preliminary conclusions about the performance of the SF-12 were based on analyses of questionnaire items embedded within the SF-36 (Ware, Kosinski, and Keller, 1996). Results documented in Appendix A support the assumption on which these conclusions were based, namely, that tests of scaling assumptions and conclusions about reliability and validity do not vary to a noteworthy extent when SF-12 items were embedded with other SF-36 items in comparison with SF-12 items administered alone.

Computer-administered telephone interviews and "pen pad" and "touch screen" systems of administering the SF-12 are currently being evaluated (Ware, Kosinski, DeBrota et al., 1995). An optical scan card version of the SF-12 and accompanying software program are also available to reduce the burden of data entry and analyses.

Although we have strongly argued against the "blind" adoption of either the SF-36 or the SF-12 for studies at the individual patient level, adoption of both forms for this purpose appears to be underway. There is little data on this important issue. Published data suggests that PCS and MCS *summary* measures based on the SF-36 may prove useful at the individual patient level (Ware, Kosinski, and Keller, 1994). The very high correlations observed between PCS scores based on the SF-12 and the SF-36 and between MCS scores based on the SF-12 and SF-36 suggest that SF36 results will generalize to the SF-12. However, we view this as a research agenda rather than a proven application.

Conclusions

The SF-12 Health Survey is likely to be a useful alternative to the SF-36 for purposes of measuring and monitoring health status in large group studies. The reduction in questionnaire length achieved by the SF-12 is large enough to make a difference in whether health status can be measured in some large-scale studies. The SF-12 can be self-administered using a single-page scannable form and it can be administered to the great majority of respondents in about two minutes or less.

The choice between the SF-36 and SF-12 health surveys is a choice between more or less information and precision in measuring health status. It is also a choice between a less practical and a more practical survey tool. The SF-12 Health Survey is most likely to be a successful "stand in" for the SF-36 when large samples from general or specific populations are being monitored and when the focus is on overall physical and mental health outcomes.

❦ Appendix A: Comparison of Responses to SF-12 Items Administered Alone and Embedded in the SF-36

Background

The SF-12 is comprised of a subset of 12 questionnaire items from the SF-36. Although the order of concepts and items is the same across forms, an important issue is whether the 24 items that are in the SF-36 but not in the SF-12 create a "context effect" causing responses to SF-12 items to be different when they are embedded in the SF-36. This issue is crucial because the construction of the SF-12 and the results from psychometric evaluations reported here and elsewhere (Ware, Kosinski, and Keller, 1996) were based on analyses of responses to SF-12 items embedded in the SF-36.

A similar issue arose during the initial studies of the SF-36, which were based on responses to items contained in the 149-item MOS Functional Health and Well Being (FHWB) questionnaire (Stewart and Ware, 1992). On the strength of results showing that the SF-36 performed at least as well when administered alone as when embedded in the FHWB, we hypothesized that SF-12 items would perform at least as well when administered alone as when embedded in the SF-36. Methods and results pertaining to this important issue are briefly summarized below.

Data and Methods

To assess the comparability of responses to SF-12 items across methods, we analyzed available data sets from studies that administered the SF-12 with (the "embedded" context) and without (the "unembedded" context) the other 24 SF-36 items. Dataset A includes responses for 525 employees who completed the SF-36 (including the embedded SF-12 items) in 1993 and completed the unembedded SF-12 a year later in 1994. Dataset B includes responses to the SF-36 (including embedded SF-12 items) for 1,819 adults representing the general US population in 1994: this dataset is the subject of current norming studies for the SF-36. Dataset C includes responses for 27,361 employees who completed the unembedded SF-12 in 1994. These datasets, thus, provide two opportunities to compare embedded and unembedded forms: 1) comparison of results for embedded (1993) and unembedded (1994) administrations to the same respondents in Dataset A; and 2) comparison of responses to embedded (Dataset B) and unembedded (Dataset C) administrations across two independent general population samples.

We evaluated results for the unembedded and embedded SF-12 items with respect to three criteria: (1) item-level descriptive statistics (mean scores and standard deviations); (2) correlations among the SF-12 items and their factor content in relation to physical and mental factors; and (3) empirical tests of the validity of the SF-12 Physical Component Summary (PCS-12) and Mental Component Summary (MCS-12) scales, in comparison to results for the SF-36 PCS and MCS, in cross-sectional studies of different age groups, groups differing in physical condition, and groups differing in depression.

Hypotheses

(1) We hypothesized that mean scores for SF-12 items would be similar and would be ordered the same across data sets B and C (two independent general populations) and across embedded and unembedded administrations in Dataset A (single sample that completed SF-12 with and without the other 24 items a year apart).

(2) We hypothesized that items measuring physical health would correlate highest with the physical health component (convergent validity) and lowest with the mental component (discriminant validity); in contrast, items measuring mental health were hypothesized to correlate highest with the mental component and lowest with the physical component. We expected there to be very little variation in the magnitude of these correlations across context.

(3) We hypothesized that SF-12 Physical Component Summary (PCS-12) scores would be lower for older age groups, regardless of context, and that PCS-12 would discriminate better than MCS-12 between groups known to differ in physical condition. We hypothesized that SF-12 Mental Component Summary (MCS-12) scores would not differ with age, regardless of context, and that MCS-12 would discriminate best between groups known to differ in depression.

Preliminary Results and Comments

Mean scores and standard deviations for favorably-scored SF-12 items are compared in Table A.1 for all three datasets. Because respondents were not randomly assigned to form and because those in Dataset A who completed embedded and unembedded SF-12 surveys did so a year apart and those in Datasets B and C are different samples, we did not expect means to be identical. Results in Table 1 support the hypothesized equivalence of descriptive statistics for responses to SF-12 items when they are and are not embedded in the SF-36. In Dataset A, the ordering of favorably-scored item means is the same across embedded and unembedded administrations a year apart with one exception. The highest means and the seven lowest means are the same across contexts and administrations. One minor reversal in item ordering was observed (items ranked 4 and 5). The product-moment correlation between 12 item means for embedded and unembedded contexts was very high in Dataset A (r=0.999). The slight decline in means, particularly for the first four physical health items expected with a year of aging, is also apparent in Dataset A.

The order of item-mean scores is also very similar in Datasets B and C. Eight of the 12 item means are ranked identically across embedded and unembedded contexts. The rank order of item means across the three data sets and embedded and unembedded contexts is strikingly similar. The product-moment correlation between the 12 item means for embedded and unembedded contexts is also very high for Dataset B and Dataset C (r=0.994).

Table A.2 presents correlations between the hypothesized physical and mental components (orthogonal rotation) extracted independently in each of the three datasets, including embedded and unembedded administrations. These results pertain to the second criterion. Results in Table 2 strongly support the hypothesized two-dimensional SF-12 factor structure across all three general population samples and across embedded and unembedded contexts. This structure, first observed in the SF-36, was the basis for constructing the SF-12 (Ware, Kosinski, and Keller, 1996). Incidentally, unit eigenvalue, Scree Test and other factor extraction and rotational criteria confirmed the hypothesized two-dimensional model in all three datasets (results not reported).

The factor content of the 12 items is strikingly similar across datasets and contexts. Specifically, the ordering of SF-12 items from the strongest to weakest correlations with the first (physical) factor is comparable across datasets and contexts; the same holds for the second (mental) factor (Table A.2).

As shown in Table A.2, the average difference in the correlations with the physical component between contexts was 0.04 in Dataset A (the largest difference was 0.24), and 0.06 between Datasets B and C (the largest difference was 0.19). The average difference in the correlations with the mental component between contexts was 0.02 in Dataset A (the largest difference was 0.22), and 0.01 between Datasets B and C (the largest difference was 0.21).

Results from comparisons of differences in mean SF-12 scores across age groups are reported for Datasets B (embedded) and C (unembedded) in Table A.3. These results confirm the hypothesized poorer physical health among older age groups in both data sets and across embedded and unembedded contexts. As hypothesized, mean scores for the PCS-12, which measures physical health, declined with older age groups. This pattern of scores is apparent in both datasets regardless of context. A noteworthy discrepancy between the two results is the substantially lower mean (more than 4 points, i.e. more than one-third of a standard deviation) for the elderly group for the embedded SF-12 in Dataset B. This result is largely due to there being a larger proportion of respondents in Dataset B who were 75 years and older (14.8%) than in Dataset C (2.3%). Another noteworthy difference between the two datasets was the larger difference in mental health across age groups for the unembedded SF-12 in Dataset C. Although slightly more favorable mental health scores have been observed among older age groups in previous studies of the SF-36, we did not hypothesize that this trend would be greater in an employed population than in the general population. Further analyses are necessary to determine whether this represents a population difference in the relationship between age and mental health or a difference in the validity of the SF-12 depending on whether embedded or unembedded contexts are used. We are currently pursuing these issues.

Mean scores for those reporting and not reporting a physical condition (e.g., diabetes, arthritis) are presented in Table A.4 for Dataset A. (Data on comorbid conditions was not available for Dataset C and, therefore, comparisons between B and C of the effect of physical conditions are not presented). Hypothesized lower scores, particularly for the measures shown to be most related to physical health in previous studies, are apparent for the impaired group without exception. As hypothesized, PCS-12 discriminated between those with and without physical conditions better than MCS-12. The hypothesized superiority of PCS-12 is evident in the larger mean differences and larger F-ratios (relative to MCS-12) in both the embedded and unembedded administrations of the SF-12 in Dataset A. Interestingly, PCS-12 showed a larger mean difference relative to its standard error in the unembedded administration of the SF-12 in 1994 (Note the larger F-ratio for that context).

Table A.4 also reports mean differences observed in Dataset A for those who screened positive and negative for depression (MOS Depression Screener) in Dataset A. It should be noted that the depression screener was administered only in 1994 and is therefore concurrent with the unembedded SF-12 which was also administered in 1994 in Dataset A. Thus, the presentation of results for Dataset A for embedded SF-12 (1993 data collection) is a weak test of the validity of the SF-12 when embedded.

Because the first analysis uses a psychiatric "criterion" from a later year, we did not expect the performance of the embedded SF-12 in Dataset A to be as good as that previously reported for the SF-36 or as good as the unembedded SF-12 in Dataset A. These expectations are confirmed in Table A.4. In the second analysis reported in Table A.4 for Dataset A (unembedded SF-12), results confirm our hypothesis of a substantial difference in mental health between those who screened positive and negative for depression. These results for the unembedded SF-12 MCS-12 are nearly identical to those that have been previously published for the SF-36 MCS scale. These results strongly support the validity of the SF-12 and particularly the MCS scale which is hypothesized to be the most valid of the two summary measures in measuring mental health for both the embedded and unembedded versions of the SF-12.

Results for the effect of chronic conditions adjusted for age and gender mirrored those in Table A.4 and are not reported.

Preliminary Conclusions

Based on our previous experience in studying the SF-36, which was embedded in a longer MOS form in earlier studies and later administered unembedded, we expected the SF-12 to perform as well when administered alone as when embedded in the SF-36. We interpret the analyses summarized above as providing strong support for this hypothesis.

Although much remains to be learned about the performance of the SF-12 Health Survey and its psychometric and normative properties in embedded and unembedded administrations, we see no trends in the results to date that would cause us to communicate a cautionary note with respect to whether performance in unembedded administrations will match its performance in analyses when embedded in the longer SF-36. We see support for the use of SF-12 in three kinds of criteria, including: 1) the ordering of item means was strikingly similar across embedded and unembedded contexts; 2) the factor structure of the SF-12 and the factor content of each of the SF-12 items was virtually indistinguishable in embedded and unembedded administrations, suggesting that each of these items has the same interpretation across embedded and unembedded surveys; and, finally, 3) the performance of the unembedded SF-12 in discriminating groups known to differ in physical health and mental health suggests that the psychometric properties of the unembedded SF-12 go hand in hand with results of empirical tests of validity of the embedded SF-12 in relation to external criteria, as has been the case in studies of the SF-36 to date.

Analyses of datasets for unembedded administrations of the SF-12 summarized above and reports from numerous other population surveys in progress confirm that a high degree of data quality is achieved in terms of both return rates and the preliminary completeness of returned surveys in unembedded administrations of the SF-12 (data not reported). At this point in the development and evaluation of the SF-12, we know much more about its psychometric properties, normative results, and its empirical validity in relation to external criteria than we did at the same stage in the development and evaluation of the SF-36. We also have a much better understanding of the relationship between results from different evaluation criteria than we had at that time. On the strength of this understanding and the pattern of results that we have observed, we recommend with confidence the use of the SF-12 for its primary purpose which is the monitoring of physical and mental health in large general and specific *populations*.

Table A.1. Item Means (and Standard Deviations) for the SF-12 Health Survey, Embedded and Unembedded Items

	Dataset A				Dataset B		Dataset C	
SF-12 Items	**Embedded**	**Rank**	**Unembedded**	**Rank**	**Embedded**	**Rank**	**Unembedded**	**Rank**
Moderate activities (PF02)	2.84 (0.4)	7	2.83 (0.4)	7	2.60 (0.7)	7	2.78 (0.5)	7
Climbing several flights of stairs (PF04)	2.78 (0.5)	8	2.75 (0.5)	8	2.43 (0.7)	8	2.72 (0.6)	8
Accomplished less due than you would like (RP2)	1.88 (0.3)	11	1.84 (0.4)	11	1.65 (0.5)	12	1.81 (0.4)	11
Limited in kind of work or activities (RP3)	1.93 (0.3)	9	1.90 (0.3)	9	1.72 (0.4)	10	1.84 (0.4)	10
How much did pain interfere with normal work (BP2)	4.58 (0.7)	3	4.44 (0.9)	3	4.20 (1.0)	2	4.31 (1.0)	3
In general, would you say your health is (GH1)	3.90 (0.8)	6	3.91 (0.8)	6	3.40 (1.0)	6	3.61 (0.8)	6
Have a lot of energy (VT2)	3.98 (1.2)	5	4.01 (1.2)	4	3.66 (1.3)	5	3.90 (1.2)	5
How much time health interferes w/social activities (SF2)	4.61 (0.8)	2	4.51 (0.9)	2	4.12 (1.2)	3	4.43 (0.9)	2
Accomplish less than you would like (RE2)	1.83 (0.4)	12	1.82 (0.4)	12	1.71 (0.5)	11	1.81 (0.4)	12
Didn't do work or other activities as carefully as usual (RE)	1.91 (0.3)	10	1.88 (0.3)	10	1.81 (0.4)	9	1.87 (0.4)	9
Felt calm and peaceful (MH3)	4.13 (1.2)	4	4.00 (1.2)	5	3.99 (1.3)	4	3.98 (1.2)	4
Felt downhearted and blue (MH4)	5.13 (1.0)	1	5.02 (1.0)	1	4.94 (1.1)	1	4.89 (1.1)	1
Correlation between embedded and unembedded vectors of item means	0.999				0.995			

Table A.2. Correlation Between SF-12 Items and Rotated Principal Components, Embedded and Unembedded Items

	Physical Component				Mental Component			
	Dataset A		Datasets B & C		Dataset A		Datasets B & C	
SF-12 Items	E-PC	U-PC	E-PC	U-PC	E-MC	U-MC	E-MC	U-MC
Moderate activities (PF02)	.76	.75	.85	.76	-.06	-.02	.03	-.01
Climbing several flights of stairs (PF04)	.74	.65	.80	.70	-.10	.11	.08	.03
Accomplished less due than you would like (RP2)	.55	.73	.73	.71	.26	.23	.25	.24
Limited in kind of work or activities (RP3)	.68	.77	.81	.78	.08	.09	.13	.13
How much did pain interfere with normal work (BP2)	.65	.71	.70	.72	.30	.26	.32	.22
In general, would you say your health is (GH1)	.53	.52	.69	.51	.27	.36	.24	.21
Have a lot of energy (VT2)	.36	.33	.51	.37	.64	.64	.55	.60
How much time health interferes w/social activities (SF2)	.42	.47	.38	.41	.63	.58	.42	.63
Accomplish less than you would like (RE2)	.10	.14	.16	.12	.76	.77	.72	.78
Didn't do work or other activities as carefully as usual (RE)	-.10	.14	.19	.12	.65	.67	.69	.73
Felt calm and peaceful (MH3)	.09	.07	.08	.05	.75	.71	.75	.73
Felt downhearted and blue (MH4)	.10	.04	.08	.03	.77	.73	.77	.75
Average Difference in Item to Component Correlation (Embedded/Unembedded)	0.04		0.06		0.02		0.01	
Largest Difference in Item to Component Correlation (Embedded/Unembedded)	0.24		0.19		0.22		0.21	

U-PC = Unembedded PCS-12
U-MC = Unembedded MCS-12
E-PC = Embedded PCS-12
E-MC = Embedded MCS-12

Table A.3. Mean (SD) PCS-12 and MCS-12 Scores Across Age Groups, Embedded and Unembedded SF-12 Items

	Age Groups				
	18-44	**45-54**	**55-64**	**65+**	F
Dataset B: Embedded (n=1,819)					
PCS-12	52.49 (7.0)	50.21 (9.3)	47.32 (10.2)	40.69 (11.9)	394.4*
MCS-12	48.21 (9.9)	49.07 (10.2)	51.05 (8.2)	51.35 (9.9)	30.5*
Dataset C: Unembedded (n=27,361)					
PCS-12	52.45 (7.1)	51.01 (8.1)	49.36 (8.9)	45.53 (10.7)	1385.5*
MCS-12	48.42 (9.8)	50.08 (9.4)	52.94 (7.9)	53.67 (8.5)	1055.9*

* p < .001

Table A.4. Mean Differences in PCS-12 and MCS-12 Scale Scores Between Groups Known to Differ in Chronic Physical and Mental Conditions, Embedded and Unembedded SF-12 Items

	1+ Physical Condition vs. None			Depression vs. No Depression		
	Mean Difference (SE)	F	RV	Mean Difference (SE)	F	RV
Dataset A						
Embedded						
PCS12	-4.06 (0.7)	30.43**	1.00	-3.60 (1.3)	8.49*	0.21
MCS12	-3.16 (0.9)	12.91**	0.42	-9.90 (1.5)	39.98**	1.00
N	292			384		
Dataset A						
Unembedded						
PCS12	-5.69 (0.8)	50.47**	1.00	-1.85 (1.4)	1.73	0.01
MCS12	-2.22 (0.8)	6.91*	0.14	-17.77 (1.4)	161.08**	1.00
N	292			384		

** p < .001
* p < .05

❦ Appendix B: Definitions of Criterion Variables

TABLE B.1. DEFINITIONS OF CRITERIA USED IN EMPIRICAL VALIDATION OF SF-12 SCALES

CRITERION	DEFINITION
MOS Tracer Conditions (Table 8.1)	
Hypertension	Physician report of current hypertension (or independently derived probability of hypertension if physician report missing or questionable).
Congestive Heart Failure	Physician report of current congestive heart failure (or independently derived probability of CHF if physician report missing or questionable).
MI (Recent)	Physician report of MI within the past year (or independently derived probability of MI if physician report of MI missing or questionable).
Diabetes, Type II	Physician report of diabetes with age at onset 30 years or older (or independently derived probability of diabetes and age at onset if actual information missing or questionable).
Severity of MOS Tracer Conditions (Table 8.2)	
Hypertension	Severity defined by diastolic blood pressure above 100 mm Hg (2 levels)
Congestive Heart Failure	Severity defined by the presence of dyspnea on one-block exertion or while lying flat (2 levels)
MI (Recent)	Severity defined by the presence of premature ventricular contractions and/or angina (2 levels)
Diabetes, Type II	Severity defined by the presence of complications and duration of diabetes (4 levels: 1-free of complications and duration less than 10 years; 2-free of complications and duration 10 or more years; 3-complications of eye of foot only; 4-complications of diabetic heart and/or kidney disease)
MOS Comorbid Conditions (Tables 8.3 and 8.4)*	
Asthma	Had any asthma attacks in past six months
COPD	Now have lung disease ever diagnosed by physician as chronic obstructive pulmonary disease (like chronic bronchitis or emphysema) in past six months.
Angina - ever**	Ever told by physician have angina.

(continued)

TABLE B.1. DEFINITIONS OF CRITERIA USED IN EMPIRICAL VALIDATION OF SF-12 SCALES (continued)

CRITERION	DEFINITION
Angina, recent - no MI	Symptoms of angina in past six months in the absence of an MI within one year.
MI, past	Ever had a heart attack diagnosed by physician, more than one year ago.
Other lung disease	Any other lung disease such as tuberculosis or pneumonia in past six months.
Back pain/sciatica	Attacks of back pain or sciatica last six months.
Hip impairments	Ever told by physician have hip impairments
Rheumatoid arthritis	Now have active condition physician ever diagnosed as arthritis and physician labeled it rheumatoid arthritis and morning stiffness.
Osteoarthritis	Now have active condition physician ever diagnosed as arthritis and physician labeled it osteoarthritis or degenerative arthritis and patient is ≥ 55 years old.
Musculoskeletal complaints	Active condition physician ever diagnosed as arthritis but criteria for osteoarthritis or rheumatoid arthritis not met.
Other rheumatic disease**	Now have active rheumatic disease other than arthritis physician ever diagnosed (e.g., systemic lupus erythematosus, scleroderma, or gout).
Colitis**	Now have active disease physician ever diagnosed as Crohn's disease or ulcerative colitis.
Diverticulitis**	Now have active disease physician ever diagnosed as diverticulitis.
Fistulas**	Now have active disease physician ever diagnosed as intestinal fistulas.
Gallbladder disease**	Now have active disease physician ever diagnosed as chronic gallbladder disease.
Irritable bowel disease	Ever told by physician have irritable bowel syndrome or functional bowel disease,.
Liver disease**	Now have active disease physician ever diagnosed as chronic hepatitis or cirrhosis.
Diabetes, Type I**	Physician report of diabetes with age at onset younger than 30 years.
Ulcer	Now have active disease physician ever diagnosed as an ulcer (peptic, gastric, stomach, or duodenal).
Kidney disease**	Disease physician ever diagnosed as serious kidney disease in last six months.
Benign Prostatic Hypertrophy**	Male, age ≥ 50 years, history of nocturia in past six months, no serious kidney disease ever diagnosed, and no report of prostatic cancer.

(continued)

TABLE B.1. DEFINITIONS OF CRITERIA USED IN EMPIRICAL VALIDATION OF SF-12 SCALES (continued)

CRITERION	DEFINITION
UTI	Kidney or bladder infection diagnosed by physician in past six months.
Varicosities**	Now have condition physician ever diagnosed as varicose veins/deep varicosities.
Cancer**	Ever had cancer.
Dermatitis	Repeated episodes of dermatitis or skin rash in past six months.
Anemia	Told by doctor have anemia (past six months.)
Symptom Clusters (Table 8.5)	
Ear, nose & throat	Patient reported frequency of blurred vision, dry mouth or lump in throat in the past four weeks.
Central Nervous system	Patient report of fainting, drowsiness or dizziness, shortness of breath, chest pain heart palpitations, or frequent headaches in the past four weeks.
Musculoskeletal	Patient report of stiffness or soreness in the joints, backache, heavy feeling in arm or legs, or numbness in the feet in the past four weeks.
GI/GU	Patient report of acid indigestion, heartburn, nausea, or trouble passing urine in the past four weeks.
Clinical Depression Groups (Tables 8.8 and 8.9)	
Cross-sectional	NIMH (DIS) criteria met for major depression and/or dysthymia at baseline assessment.
Longitudinal	Major depression and/or dysthymia present at one-year follow-up but *not present* at two-year follow-up.
Age Groups (Tables 8.6)	
Age 18-44	Uncomplicated hypertensives (patients with hypertension and no other major medical conditions), age 18 - 44
Age 45-64	Uncomplicated hypertensives (patients with hypertension and no other major medical conditions), age 45 - 64
Age 65 or older	Uncomplicated hypertensives (patients with hypertension and no other major medical conditions), age >= 65

(continued)

TABLE B.1. DEFINITIONS OF CRITERIA USED IN EMPIRICAL VALIDATION OF SF-12 SCALES (continued)

CRITERION	DEFINITION
Self-Reported Transition Groups (Table 8.7)	
Physical	Patient report at two-year follow-up of change in physical health over two years: (a lot more limited now, a little more limited now, about the same, somewhat less limited now or a lot less limited now.
Mental	Patient report at two-year follow-up of change in mental health over two years: (a lot more limited now, a little more limited now, about the same, somewhat less limited now or a lot less limited now.

* Information regarding the comorbid medical conditions was obtained from the patient during a structured medical history interview conducted by a trained clinician. If information regarding a condition (or conditions) was missing, an independently derived probability of each diagnosis was substituted.

** Because of very low prevalence, the following conditions are incorporated into an index of eleven comorbid conditions: angina-ever, other rheumatic disease, colitis, diverticulitis, intestinal fistulas, gallbladder disease, liver disease, benign prostatic hypertrophy, varicosities, cancer, and type I diabetes.

❦ Appendix C: SAS Scoring Program

```
FILENAME IN 'A:\SF12RAW.DAT';
*****************************************************************************************
*;
 PROGRAM: SF12SUMM.SCR
 PURPOSE: SAS SCORING PROGRAM FOR THE SF-12 SUMMARY SCALES
   SF-12 SUMMARY SCALE SCORING EXERCISE (FIRST EDITION).
   COPYRIGHT 1995, 1994 MEDICAL OUTCOMES TRUST, ALL RIGHTS RESERVED.

 SF-12 IS A REGISTERED TRADEMARK OF MEDICAL OUTCOMES TRUST.
 SAS IS A REGISTERED TRADEMARK OF SAS INSTITUTE, INC., CARY NC.
*****************************************************************************************
;

*****************************************************************************************
;
    INPUT DATA
*****************************************************************************************
;

DATA SF12DATA;
INFILE IN;
INPUT ID $ 1-3
  @ 5 (GH1 PF02 PF04 RP2 RP3 RE2 RE3 BP2
   MH3 VT2 MH4 SF2) (1.0);
RUN;

****************************************************************************************;
    STEP 1: DATA CLEANING/REVERSE SCORING
****************************************************************************************;

****************************************************************************************;
 USING THE SAS DATASET CREATED IN PART 1, CHANGE OUT-OF-RANGE
 VALUES TO MISSING FOR EACH ITEM.
****************************************************************************************;

DATA SF12SCAL;
 SET SF12DATA;

ARRAY TWOPT RP2 RP3 RE2 RE3;
 DO OVER TWOPT;
 IF TWOPT LT 1 OR TWOPT GT 2 THEN TWOPT = .;
END;

ARRAY THREEPT PF02 PF04;
 DO OVER THREEPT;
 IF THREEPT LT 1 OR THREEPT GT 3 THEN THREEPT = .;
END;
```

```
ARRAY FIVEPT GH1 BP2 SF2;
 DO OVER FIVEPT;
 IF FIVEPT LT 1 OR FIVEPT GT 5 THEN FIVEPT = .;
END;

ARRAY SIXPT VT2 MH3 MH4;
 DO OVER SIXPT;
 IF SIXPT LT 1 OR SIXPT GT 6 THEN SIXPT = .;
END;

************************************************************************************;
 USING THE SAS DATASET CREATED IN PART 1, REVERSE SCORE
 FOUR ITEMS SO THAT HIGHER SCORE INDICATES BETTER HEALTH
************************************************************************************;

RBP2=6-BP2;
RGH1=6-GH1;
RVT2=7-VT2;
RMH3=7-MH3;

*************************************************************************************;
    STEP 2: CREATE INDICATOR VARIABLES FOR
      ITEM RESPONSE CHOICES
*************************************************************************************;

PF02_1 = .;
 if PF02 = . then PF02_1 = .; else
 if PF02 = 1 then PF02_1 = 1; else PF02_1 = 0;

PF02_2 = .;
 if PF02 = . then PF02_2 = .; else
 if PF02 = 2 then PF02_2 = 1; else PF02_2 = 0;

PF04_1 = .;
 if PF04 = . then PF04_1 = .; else
 if PF04 = 1 then PF04_1 = 1; else PF04_1 = 0;

PF04_2 = .;
 if PF04 = . then PF04_2 = .; else
 if PF04 = 2 then PF04_2 = 1; else PF04_2 = 0;

RP2_1 = .;
 if RP2 = . then RP2_1 = .; else
 if RP2 = 1 then RP2_1 = 1; else RP2_1 = 0;

RP3_1 = .;
 if RP3 = . then RP3_1 = .; else
 if RP3 = 1 then RP3_1 = 1; else RP3_1 = 0;

BP2_1 = .;
 if RBP2 = . then BP2_1 = .; else
 if RBP2 = 1 then BP2_1 = 1; else BP2_1 = 0;

BP2_2 = .;
 if RBP2 = . then BP2_2 = .; else
 if RBP2 = 2 then BP2_2 = 1; else BP2_2 = 0;
```

```
BP2_3 = .;
 if RBP2 = . then BP2_3 = .; else
 if RBP2 = 3 then BP2_3 = 1; else BP2_3 = 0;

BP2_4 = .;
 if RBP2 = . then BP2_4 = .; else
 if RBP2 = 4 then BP2_4 = 1; else BP2_4 = 0;

GH1_1 = .;
 if RGH1 = . then GH1_1 = .; else
 if RGH1 = 1 then GH1_1 = 1; else GH1_1 = 0;

GH1_2 = .;
 if RGH1 = . then GH1_2 = .; else
 if RGH1 = 2 then GH1_2 = 1; else GH1_2 = 0;

GH1_3 = .;
 if RGH1 = . then GH1_3 = .; else
 if RGH1 = 3 then GH1_3 = 1; else GH1_3 = 0;

GH1_4 = .;
 if RGH1 = . then GH1_4 = .; else
 if RGH1 = 4 then GH1_4 = 1; else GH1_4 = 0;

VT2_1 = .;
 if RVT2 = . then VT2_1 = .; else
 if RVT2 = 1 then VT2_1 = 1; else VT2_1 = 0;

VT2_2 = .;
 if RVT2 = . then VT2_2 = .; else
 if RVT2 = 2 then VT2_2 = 1; else VT2_2 = 0;

VT2_3 = .;
 if RVT2 = . then VT2_3 = .; else
 if RVT2 = 3 then VT2_3 = 1; else VT2_3 = 0;

VT2_4 = .;
 if RVT2 = . then VT2_4 = .; else
 if RVT2 = 4 then VT2_4 = 1; else VT2_4 = 0;

VT2_5 = .;
 if RVT2 = . then VT2_5 = .; else
 if RVT2 = 5 then VT2_5 = 1; else VT2_5 = 0;

SF2_1 = .;
 if SF2 = . then SF2_1 = .; else
 if SF2 = 1 then SF2_1 = 1; else SF2_1 = 0;

SF2_2 = .;
 if SF2 = . then SF2_2 = .; else
 if SF2 = 2 then SF2_2 = 1; else SF2_2 = 0;

SF2_3 = .;
 if SF2 = . then SF2_3 = .; else
 if SF2 = 3 then SF2_3 = 1; else SF2_3 = 0;

SF2_4 = .;
```

```
 if SF2 = . then SF2_4 = .; else
 if SF2 = 4 then SF2_4 = 1; else SF2_4 = 0;

RE2_1 = .;
 if RE2 = . then RE2_1 = .; else
 if RE2 = 1 then RE2_1 = 1; else RE2_1 = 0;

RE3_1 = .;
 if RE3 = . then RE3_1 = .; else
 if RE3 = 1 then RE3_1 = 1; else RE3_1 = 0;

MH3_1 = .;
 if RMH3 = . then MH3_1 = .; else
 if RMH3 = 1 then MH3_1 = 1; else MH3_1 = 0;

MH3_2 = .;
 if RMH3 = . then MH3_2 = .; else
 if RMH3 = 2 then MH3_2 = 1; else MH3_2 = 0;

MH3_3 = .;
 if RMH3 = . then MH3_3 = .; else
 if RMH3 = 3 then MH3_3 = 1; else MH3_3 = 0;

MH3_4 = .;
 if RMH3 = . then MH3_4 = .; else
 if RMH3 = 4 then MH3_4 = 1; else MH3_4 = 0;

MH3_5 = .;
 if RMH3 = . then MH3_5 = .; else
 if RMH3 = 5 then MH3_5 = 1; else MH3_5 = 0;

MH4_1 = .;
 if MH4 = . then MH4_1 = .; else
 if MH4 = 1 then MH4_1 = 1; else MH4_1 = 0;

MH4_2 = .;
 if MH4 = . then MH4_2 = .; else
 if MH4 = 2 then MH4_2 = 1; else MH4_2 = 0;

MH4_3 = .;
 if MH4 = . then MH4_3 = .; else
 if MH4 = 3 then MH4_3 = 1; else MH4_3 = 0;

MH4_4 = .;
 if MH4 = . then MH4_4 = .; else
 if MH4 = 4 then MH4_4 = 1; else MH4_4 = 0;

MH4_5 = .;
 if MH4 = . then MH4_5 = .; else
 if MH4 = 5 then MH4_5 = 1; else MH4_5 = 0;


  *************************************************************************************;
      STEP 3: WEIGHTING AND AGGREGATION OF
        INDICATOR VARIABLES USING
        PHYSICAL AND MENTAL REGRESSION WEIGHTS
  *************************************************************************************;
```

```
RAWPCS12 = (-7.23216*PF02_1) + (-3.45555*PF02_2) +
 (-6.24397*PF04_1) + (-2.73557*PF04_2) + (-4.61617*RP2_1) +
 (-5.51747*RP3_1) + (-11.25544*BP2_1) + (-8.38063*BP2_2) +
 (-6.50522*BP2_3) + (-3.80130*BP2_4) + (-8.37399*GH1_1) +
 (-5.56461*GH1_2) + (-3.02396*GH1_3) + (-1.31872*GH1_4) +
 (-2.44706*VT2_1) + (-2.02168*VT2_2) + (-1.6185*VT2_3) +
 (-1.14387*VT2_4) + (-0.42251*VT2_5) + (-0.33682*SF2_1) +
 (-0.94342*SF2_2) + (-0.18043*SF2_3) + (0.11038*SF2_4) +
 (3.04365*RE2_1) + (2.32091*RE3_1) + (3.46638*MH3_1) +
 (2.90426*MH3_2) + (2.37241*MH3_3) + (1.36689*MH3_4) +
 (0.66514*MH3_5) + (4.61446*MH4_1) + (3.41593*MH4_2) +
 (2.34247*MH4_3) + (1.28044*MH4_4) + (0.41188*MH4_5);


RAWMCS12 = (3.93115*PF02_1) + (1.8684*PF02_2) +
 (2.68282*PF04_1) + (1.43103*PF04_2) + (1.4406*RP2_1) +
 (1.66968*RP3_1) + (1.48619*BP2_1) + (1.76691*BP2_2) +
 (1.49384*BP2_3) + (0.90384*BP2_4) + (-1.71175*GH1_1) +
 (-0.16891*GH1_2) + (0.03482*GH1_3) + (-0.06064*GH1_4) +
 (-6.02409*VT2_1) + (-4.88962*VT2_2) + (-3.29805*VT2_3) +
 (-1.65178*VT2_4) + (-0.92057*VT2_5) + (-6.29724*SF2_1) +
 (-8.26066*SF2_2) + (-5.63286*SF2_3) + (-3.13896*SF2_4) +
 (-6.82672*RE2_1) + (-5.69921*RE3_1) + (-10.19085*MH3_1) +
 (-7.92717*MH3_2) + (-6.31121*MH3_3) + (-4.09842*MH3_4) +
 (-1.94949*MH3_5) + (-16.15395*MH4_1) + (-10.77911*MH4_2) +
 (-8.09914*MH4_3) + (-4.59055*MH4_4) + (-1.95934*MH4_5);


****************************************************************************************;
   STEP 4: NORM-BASED STANDARDIZATION OF SCALE SCORES
****************************************************************************************;

PCS12 = RAWPCS12 + 56.57706;

MCS12 = RAWMCS12 + 60.75781;
```

❦ Appendix D: Standard and Acute Forms; Interviewer Script

SF-12 HEALTH SURVEY (STANDARD)

INSTRUCTIONS: This questionnaire 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**

SF-12 HEALTH SURVEY (ACUTE)

INSTRUCTIONS: This questionnaire 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 <u>your health now limit you</u> 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 <u>past week</u>, have you had any of the following problems with your work or other regular daily activities <u>as a result of your physical health</u>?

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

During the past week, 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 week, 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 week. 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 week -

		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 week, 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**

SCRIPT FOR PERSONAL INTERVIEW SF-12 ADMINISTRATION

The script included in this appendix is recommended for interviewer administrations of the SF-12 items. It can be administered both by telephone and in-person. Standard SF-12 instructions should precede this script. Interviewers also should follow standard procedures for repeating questions and response choices as required by the respondent.

The first question is about your health now and your current daily activities. Please try to answer the question as accurately as you can.

Q1 In general, would you say your health is...

1. *excellent*
2. *very good*
3. *good*
4. *fair*
5. *poor*

Now I'm going to read a list of activities that you might do during a typical day. As I read each item, please tell me if your health now limits you a lot, limits you a little, or does not limit you at all in these activities.

Q2 ...moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf. Does your health now limit you a lot, limit you a little, or not limit you at all?

If R says s/he does not do activity, probe:

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*

3. No, not limited at all

Q3 ...climbing several flights of stairs. Does your health now limit you a lot, limit you a little, or not limit you at all?

If R says s/he does not do activity, probe:

Is that because of your health?

1. Yes, limited a lot

2. Yes, limited a little

3. No, not limited at all

The following two questions ask you about your physical health and your daily activities.

Q4 During the past 4 weeks, have you accomplished less than you would like as a result of your physical health?

1. Yes

2. No

Q5. During the past 4 weeks, were you limited in the kind of work or other regular daily activities you do as a result of your physical health?

1. Yes

2. No

The following two questions ask about your emotions and your daily activities:

Q6 During the past 4 weeks, have you accomplished less than you would like as a result of any emotional problems, such as feeling depressed or anxious?

1. Yes

2. No

Q7 During the past 4 weeks, did you not do work or other regular activities as carefully as usual as a result of any emotional problems, such as feeling depressed or anxious?

1. *Yes*

2. *No*

Q8 During the past 4 weeks, how much did pain interfere with your normal work, including both work outside the home and housework? Did it interfere...

1. *not at all*

2. *a little bit*

3. *moderately*

4. *quite a bit*

5. *or extremely*

Q9 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 or relatives? Has it interfered...

1. *all of the time*

2. *most of the time*

3. *some of the time*

4. *a little of the time*

5. *or none of the time*

The next questions are about how you feel and how things have been with you during the past 4 weeks.

As I read each statement, please give me the one answer that comes closest to the way you have been feeling; is it all of the time, most of the time, a good bit of the time, some of the time, a little of the time, or none of the time?

Q10 How much of the time during the past 4 weeks ... have you felt calm and peaceful? Read categories only if necessary

1. *all of the time*
2. *most of the time*
3. *a good bit of the time*
4. *some of the time*
5. *a little of the time*
6. *none of the time*

Q11 How much of the time during the past 4 weeks ... did you have a lot of energy? Read categories only if necessary

1. *all of the time*
2. *most of the time*
3. *a good bit of the time*
4. *some of the time*
5. *a little of the time*
6. *none of the time*

Q12 How much of the time during the past 4 weeks ... have you felt downhearted and blue? Read categories only if necessary

1. *all of the time*
2. *most of the time*
3. *a good bit of the time*
4. *some of the time*
5. *a little of the time*
6. *none of the time*

❦ Appendix E: SF-12 User Mailing List Registration Form

Users of the *SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales* on our mailing list are sent updates without charge as they become available. If you would like to be on this mailing list, please fill out and mail or FAX this form.

Contact person ....

Title ....

Organization ....

Address ....

....

....

Telephone ....

Facsimile ....

We welcome your comments, including suggestions for improvement:

....

....

....

....

....

....

....

Please return form to:

SF-12 Manual Mailing List, The Health Institute, NEMC - Box 345, 750 Washington Street, Boston, MA 02111, or FAX to: 617-636-8077.

❦ References

Brazier J, Jones N, and Kind P. Testing the validity of the EuroQOL and comparing it with the SF-36 Health Survey Questionnaire. *Quality of Life Research* 1993;2:169-180.

Kerlinger FN. *Foundations of Behavioral Research.* New York: Holt, Rinehart, and Winston, 1973.

McHorney CA, Kosinski M, and 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. *Medical Care* 1994;32(6):551-567.

McHorney CA, Ware JE, and Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care* 1993;31(3):247-263.

McHorney CA, Ware JE, Rogers WH, Raczek A, and Lu JFR. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: results from the Medical Outcomes Study. *Medical Care* 1992;30(Suppl 5):MS253-MS265.

Medical Outcomes Trust. *How to Score the SF-36 Health Status Survey.* Boston, MA: The Health Institute, New England Medical Center Hospitals, 1991.

Medical Outcomes Trust. *Scoring Exercise for the MOS SF-36 Health Survey, 2nd edition.* Boston, MA: Medical Outcomes Trust, 1994.

Nunnally JC and Bernstein IH. *Psychometric Theory, 3rd Edition.* New York: McGraw-Hill, 1994.

Stewart AL and Ware JE, editors. *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach.* Durham, NC: Duke University Press, 1992.

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;262:925-930.

Ware JE. The status of health assessment 1994. *Annual Review of Public Health* 1995;16:327-354.

Ware JE. *How to Score the Revised MOS Short-Form Health Scale (SF-36).* Boston, MA: The Health Institute, New England Medical Center Hospitals, 1988.

Ware JE, Gandek B, Keller SD, and the IQOLA Group. Evaluating instruments used cross-nationally: methods from the IQOLA Project. In: Spilker B, editor. *Quality of Life and Pharmacoeconomics in Clinical Trials, Second Edition.* New York: Raven Press, 1995.

Ware JE, Keller SD, Gandek B, Brazier JE, Sullivan M, and the IQOLA Project Group. Evaluating translations of health status questionnaires: methods from the IQOLA Project. *International Journal of Technology Assessment in Health Care* 1995;11(3):525-551.

Ware JE, Kosinski M, and Keller SD. *SF-36 Physical and Mental Summary Scales: A User's Manual.* Boston, MA: The Health Institute, 1994.

Ware JE, Kosinski M, and Keller SD. *SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales.* Boston, MA: The Health Institute, New England Medical Center, First Edition, March 1995.

Ware JE, Kosinski M, and Keller SD. A 12-Item Short-Form Health Survey (SF-12): construction of scales and preliminary tests of reliability and validity. *Medical Care* 1996;32(3):220-233.

Ware JE, Kosinski M, Bayliss MS, McHorney CA, Rogers WH, and Raczek A. 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* 1995;33(Suppl 4):AS264-AS279.

Ware JE, Kosinski M, DeBrota DJ. Comparison of patient responses to SF-36 Health Surveys that are self-administered, interviewer administered by telephone, and computer administered by telephone. Abstract presented at the American Federation for Clinical Trials Eastern Regional Meeting, Boston, MA, October 28, 1995.

Ware JE, Snow KK, Kosinski M, and Gandek B. *SF-36 Health Survey Manual and Interpretation Guide.* Boston, MA: New England Medical Center, The Health Institute, 1993.