Copyright © 1994

Copyright © 1994 – 2001, John E. Ware, Jr., Ph.D.

All rights reserved. No part of this manual covered by the copyrights hereon may be reproduced or transmitted in any form or by any means—electronic, mechanical, including photocopy, recording, or any information storage or retrieval system—without permission of the copyright holder. Permission to reproduce the SF-36™ Health Survey and scoring algorithms is hereby granted to the holder of this manual for his/her own personal use.

ISBN: 1-891810-09-X

The SF-36™ Health Survey is reproduced with permission of the Medical Outcomes Trust (MOT), a non-profit organization that ensures the availability of the SF-36™ Health Survey, while preserving standardization of its content, scoring, and labeling. Permission to use the SF-36™ Health Survey is routinely granted by the MOT, upon receipt of a request, to individuals and organizations for their own non-commercial use on a royalty-free basis.

Requests for permission to reproduce parts of this manual should be sent to John E. Ware, Jr., Ph.D., QualityMetric Incorporated, 640 George Washington Highway, Lincoln, RI 02865, info@qmetric.com.

Permission to reproduce sample SF-36 items and to make reference to specific normative data from this manual for purposes of documenting or interpreting a specific published study is hereby granted to the holder of this manual with the understanding that proper citation will be made to this manual as the source.

SF-36 is a registered trademark of the Medical Outcomes Trust.

Suggested citation:

Ware JE and Kosinski M. SF-36 PHYSICAL & MENTAL HEALTH SUMMARY SCALES: A MANUAL FOR USERS OF VERSION 1,
Second Edition, Lincoln, RI: QualityMetric, 2001

Janssen Biotech, Inc.
Exhibit 2100
Accord v. Janssen
IPR2026-00257
Ex. 2100-001

Table of Contents

Copyright © 1994 ..... **i**

Table of Contents ..... **ii**

Acknowledgements ..... **iv**

Preface ..... **vi**

SECOND EDITION VI
FIRST EDITION VII

Chapter 1: How To Use This Manual **1**

Chapter 2: Introduction **3**

Chapter 3: Construction Of Summary Measures **5**

SF-36 MEASUREMENT TOOL 5
METHODOLOGICAL ISSUES 12
SF-12 HEALTH SURVEY 16

Chapter 4: Norm-Based Scoring And Interpretation **18**

ADVANTAGES OF NORM-BASED SCORING 18

Chapter 5: Norm-Based Scoring Of SF-36 Scales **22**

IMPORTANCE OF STANDARDIZATION 22
NORM-BASED SCORING OF SCALE SCORES, STANDARD FORM (4-WEEK RECALL) 23
NORM-BASED SCORING OF SCALE SCORES, ACUTE FORM (1-WEEK RECALL) 25

Chapter 6: Norm-Based Scoring For Physical & Mental Summary Measures **28**

STANDARDIZATION OF SCALES (Z-SCORES), ACUTE FORM 32
MISSING DATA ESTIMATION 33
FEATURES OF PCS AND MCS SCORES 33

Chapter 7: Reliability And Statistical Power **36**

ESTIMATION OF RELIABILITY 37
SUMMARY OF RELIABILITY ESTIMATES 37
STATISTICAL POWER 39

Chapter 8: Empirical Validation **45**

INITIAL VALIDATION STUDIES 46
CORRELATIONS WITH SPECIFIC SYMPTOMS 53
CORRELATIONS WITH OTHER MOS SCALES 54

Chapter 9: Interpretation: Content And Criterion-Based **57**

CONTENT-BASED INTERPRETATION 57
CRITERION-BASED INTERPRETATION 64
LIFE SATISFACTION AND MCS 69
BURDEN OF CHRONIC CONDITIONS 70
EFFECTS OF AGING 74
SUMMARY OF EFFECT SIZES 75

Evaluation of Changes In PCS And MCS Scores 83
Final Comment 86

Chapter 10: Norms: 1998 General U.S. Population 88

Background 88
How The SF-36 Was Normed 90

Chapter 11: Applications: Outcomes Research 120

Heart Valve Replacement 121
DUODENAL ULCER TREATMENT 127
Method For Relating PCS And MCS Scale Scores To "Criterion" Scores 128

Chapter 12: Applications: Clinical Practice 131

Charts For Longitudinal Data 132
Profiles For Individual Patients 134
Patient Profiles: More Examples 136
MCS And Clinical Depression 138
PCS Scores And Physical Disease 139

Chapter 13: Individual License Information 141

Types Of Licenses Offered 142
Single-User, Non-Commercial License 142
Disclaimer Of Warranty 143

References 144

PCS/MCS References 154

Appendix A: Definitions Of Criterion Variables 158

Appendix B: Table Of Results 162

Appendix C: Scoring Algorithms: 1990 Norm-Based Scoring of PCS and MCS 171

Steps In Scoring 172
Features Of PCS And MCS Scores 173
Scoring Checks 174

Appendix D: 1990 Norms for PCS and MCS 176

General Population Database 178
MOS Database 179
Effects Of Data Collection Method 180
U.S. General Population Norms 181
Norms For Chronic Conditions, General U.S. Population 182
Norms For Chronic Conditions: General U.S. Population 183

Acknowledgements

Development and validation of the SF-36 Health Survey was supported by a grant (#89-6515) from the Henry J. Kaiser Family Foundation, Menlo Park, California, to The Health Institute at New England Medical Center Hospitals (J.E. Ware, Jr., Ph.D., Principal Investigator).

Development of the physical (PCS) and mental (MCS) component summary health measures for the SF-36 was supported 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, and by The Health Institute at New England Medical Center Hospitals (NEMCH) from its own research funds.

The national norming of the SF-36 Health Survey in the US population in 1998 was supported, in part, by the Musculoskeletal Education and Research Institute of the American Academy of Orthopedic Surgeons and by QualityMetric, Inc., Lincoln, RI, from its own research funds. The 1990 national norming of the SF-36 in the US population, also documented here, was sponsored by the Functional Outcomes Program of the Henry J. Kaiser Family Foundation, The Health Institute, NEMCH, Boston, MA.

Several publications documenting the SF-36 were very useful in preparing the second edition of this manual. Most useful was the first edition entitled, *SF-36 Physical and Mental Health Summary Scales: A User's Manual* (Ware, Kosinski and Keller, 1994). We gratefully acknowledge the contributions of Susan D. Keller, PhD, our co-author, for her many contributions to that manual. Other useful documents include: *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); and *SF-36 Health Survey Manual and Interpretation Guide* (Ware, Snow, Kosinski, et al., 1993).

We also gratefully acknowledge our IQOLA Project collaborators in 10 countries for the very informative studies of the PCS and MCS, as summarized here, including: the Director of the IQOLA Project (Barbara Gandek, MS), and other collaborators from Denmark (Jakob Bjorner, Md, PhD), France (Alan Leplege, MD, PhD), Germany (Monika Bullinger, PhD), Italy (Giovanni Apolone, MD), The Netherlands (Neil K Aaronson, PhD), Norway (Stein Kaasa, Md,PhD), Spain (Jordi Alonso, MD, PhD), Sweden (Marianne Sullivan, PhD), and the United Kingdom (John Brazier, PhD). In addition to sharing their results, which are also published elsewhere (Gandek and Ware, 1998), our IQOLA Project collaborators provided valuable comments regarding the scoring and interpretation of the PCS and MCS summary health measures from the SF-36.

We gratefully acknowledge: William H. Rogers, PhD for his invaluable assistance with our analyses of mortality data from the Medical Outcomes Study (MOS) and for providing statistical consultation throughout the development of the SF-36 and other tools in the "SF" family of measures; Colleen A. McHorney, PhD, for her contribution to initial studies of the factor structure of the SF-36 in the U.S.; Martha Bayliss, MSc for her efforts during the first analyses of the PCS and MCS in the MOS; and both Marie Perrone and Rosamaria Amoros-Cherry for their assistance in preparing the second edition of this manual.

Applications: Clinical Practice

Chapter 12 presents examples of how SF-36 summary measures can be used in monitoring individual patients over time, in screening patients, and illustrates different display formats.

Scoring Algorithms

Appendix D documents the algorithms for aggregating the eight SF-36 scales to score the PCS and MCS summary scales. Using 1990 norm-based scoring methods.

Norm-Based Interpretation

Appendix E presents 50 tables of norms for a representative sample of the 1990 general U.S. population (for seven age groups, men and women), as well as for MOS patients, including norms for one-year change scores. This chapter also describes sampling methods and sample characteristics.

References And Bibliography

Complete citations for more than 125 referenced publications, including more than 87 publications about the SF-36, are included.

Chapter 2: Introduction

The cycle of measurement development involves a model of health status components and the search for the most valid operational definitions of those components. The cycle must grapple with both uncertainty about the underlying conceptual framework of health and the validity of assessment methods. Out of necessity, the process is circular and tests of hypotheses about the structure of health are vulnerable to the limits of the measures used in hypothesis testing. Advances in understanding of the structure of health can lead to breakthroughs in measurement strategies.

Physical And Mental Health

Discovery of the physical and mental components of health has substantial implications for the construct validation of health measures and a new strategy for creating more useful summary measures of the information they contain. The observation that the eight SF-36 scales define distinct physical and mental health clusters in factor analytic studies of both patients participating in the Medical Outcomes Study (MOS) (McHorney, Ware, Raczek, et al., 1993; Ware, Gandek, & the IQOLA Project Group, 1994) and in the general U.S. population (Ware, Snow, Kosinski, et al., 1993) constitutes considerable support for the construct validity of the SF-36. The generalizability of these findings is underscored by results showing that the factor content of each SF-36 scale -- the extent to which each scale measures a physical and/or mental health component -- is also very similar across populations (Ware, Snow, Kosinski, et al., 1993; Ware, Kosinski, Bayliss, et al., 1995). These results have proven very useful in establishing interpretation guidelines for each of the SF-36 scales, as documented in the original *SF-36 Health Survey Manual and Interpretation Guide* (Ware, Snow, Kosinski, et al., 1993).

The discovery that from 80 to 85 percent of the reliable variance in the eight SF-36 scales is accounted for by physical and mental components of health opens the door for a breakthrough in scoring and interpretation. Specifically, this result suggests that psychometrically-based summary measures have the potential to reduce the number of statistical comparisons required in analyzing SF-36 data from eight to two without substantial loss of information. Preliminary tests of their potential have yielded promising results (Ware, Kosinski, Bayliss, et al., 1995).

Psychometric-Based Summary Measures

In support of aggregating highly related health measures using psychometric methods, initial results from the MOS suggested that scales with the same factor content are much more likely to lead to the same conclusions about health outcomes than scales with different factor content (McHorney, Ware, & Raczek, 1993). Thus, measures having the same factor content are good candidates for aggregation, because less information is likely to be lost. Factor analytic methods are likely to prove useful in deriving the weights used in aggregation. Although factor analytic methods have been used in construct validation (Ware, 1976; Goldberg & Hillier, 1979; Ware, Brook, Davies-Avery, et al., 1980; Veit & Ware, 1983; Derogatis, 1986; Wiklund, Lindvall, Swedberg, et al., 1987; Mason, Anderson, & Meenan, 1988; Hall, Epstein, & McNeil, 1989; Hays & Stewart 1990; Schag, Heinrich, Aadland, et al., 1990), the scoring of higher-order factors to achieve summary measures of health status has been pursued rarely (Davies & Ware, 1981; Veit & Ware, 1983). As a result, little is known about the tradeoffs between the simplicity of fewer statistical comparisons with aggregate summary measures versus the greater sensitivity of a scale to effects concentrated in a particular dimension of health (Ware, 1984). In theory, to the extent that results are the same across conceptually related scales, a summary measure that aggregates them should be *more* useful than any one of them in detecting a difference in health status at a point in time or a change in health over time. To the extent that differences are concentrated in a particular subscale, a summary measure is *less* likely to capture that difference or change (Ware, Kosinski, Bayliss, et al., 1995).

Standardized Scoring

Following several years of evaluating the factor structure of the SF-36 among those with various chronic conditions and among the "well" in general populations in the U.S. and in other countries, it is clear that physical and mental factors account for the great majority of the variance in SF-36 scales across populations. It follows that the standardized scoring of measures of physical and mental factors will add a useful option for the scoring of SF-36 data with advantages in interpreting and presenting results.

This user's manual summarizes the empirical work leading to the development of the SF-36 Physical Component Summary (PCS) and Mental Component Summary (MCS) scales and provides the documentation necessary to score and interpret those measures.

Chapter 3: Construction Of Summary Measures

This chapter documents the factor analytic methods used to evaluate the construct validity of the SF-36 in relation to a two-factor -- physical and mental -- model of health across populations. The construction of the PCS and MCS summary measures of physical and mental health based on those methods and some of the more important methodological issues are also explained. The PCS and MCS summary measures are referred to as "component" measures (Ware, Kosinski, Bayliss, et al., 1995) because they were derived and scored using a factor analytic method called principal *components* analysis (Harman, 1976).

Factor analyses of correlations among the eight SF-36 scales have consistently identified two factors, as documented below. On the strength of the pattern of their correlations with the eight scales, they have been interpreted as "physical" and "mental" dimensions of health status. These physical and mental components accounted for 81.5% of the reliable variance in SF-36 scales in the general U.S. population (Ware, Kosinski, Bayliss, et al., 1995), and 82.4% in the MOS (McHorney, Ware, Raczek, 1993). Similar physical and mental components have been observed for other comprehensive surveys, including the Health Insurance Experiment Medical History Questionnaire (Ware, Brook, Davies-Avery, et al., 1980), the MOS Functioning and Well-Being Profile (Hays & Stewart, 1990), and the Sickness Impact Profile (Bergner, Bobbitt, Carter, et al., 1981). In further support of the generalizability of this two-dimensional model of health, we would expect factor analytic studies of other comprehensive surveys to yield recognizably similar physical and mental components.

The psychometric approach to summarizing health measures illustrated here is in contrast to a utility index, in which measures are aggregated without regard to their inter-relationships. A utility index achieves a single summary score at the expense of sensitivity and specificity to physical versus mental components of health status. A strength of the PCS and MCS summary scales described here is their value in distinguishing a physical from a mental health outcome.

SF-36 Measurement Model

Figure 3.1 illustrates the measurement model underlying the construction of SF-36 multi-item scales and summary measures. This model has three levels: (1) items, (2) scales that aggregate items, and (3) summary measures that aggregate scales. All but one of the 36 items (self-reported health transition) are used to score the eight SF-36 scales.

Psychometric-Based Summary Measures

In support of aggregating highly related health measures using psychometric methods, initial results from the MOS suggested that scales with the same factor content are much more likely to lead to the same conclusions about health outcomes than scales with different factor content (McHorney, Ware, & Raczek, 1993). Thus, measures having the same factor content are good candidates for aggregation, because less information is likely to be lost. Factor analytic methods are likely to prove useful in deriving the weights used in aggregation. Although factor analytic methods have been used in construct validation (Ware, 1976; Goldberg & Hillier, 1979; Ware, Brook, Davies-Avery, et al., 1980; Veit & Ware, 1983; Derogatis, 1986; Wiklund, Lindvall, Swedberg, et al., 1987; Mason, Anderson, & Meenan, 1988; Hall, Epstein, & McNeil, 1989; Hays & Stewart 1990; Schag, Heinrich, Aadland, et al., 1990), the scoring of higher-order factors to achieve summary measures of health status has been pursued rarely (Davies & Ware, 1981; Veit & Ware, 1983). As a result, little is known about the tradeoffs between the simplicity of fewer statistical comparisons with aggregate summary measures versus the greater sensitivity of a scale to effects concentrated in a particular dimension of health (Ware, 1984). In theory, to the extent that results are the same across conceptually related scales, a summary measure that aggregates them should be *more* useful than any one of them in detecting a difference in health status at a point in time or a change in health over time. To the extent that differences are concentrated in a particular subscale, a summary measure is *less* likely to capture that difference or change (Ware, Kosinski, Bayliss, et al., 1995).

Standardized Scoring

Following several years of evaluating the factor structure of the SF-36 among those with various chronic conditions and among the "well" in general populations in the U.S. and in other countries, it is clear that physical and mental factors account for the great majority of the variance in SF-36 scales across populations. It follows that the standardized scoring of measures of physical and mental factors will add a useful option for the scoring of SF-36 data with advantages in interpreting and presenting results.

This user's manual summarizes the empirical work leading to the development of the SF-36 Physical Component Summary (PCS) and Mental Component Summary (MCS) scales and provides the documentation necessary to score and interpret those measures.

Chapter 3: Construction Of Summary Measures

This chapter documents the factor analytic methods used to evaluate the construct validity of the SF-36 in relation to a two-factor -- physical and mental -- model of health across populations. The construction of the PCS and MCS summary measures of physical and mental health based on those methods and some of the more important methodological issues are also explained. The PCS and MCS summary measures are referred to as "component" measures (Ware, Kosinski, Bayliss, et al., 1995) because they were derived and scored using a factor analytic method called principal *components* analysis (Harman, 1976).

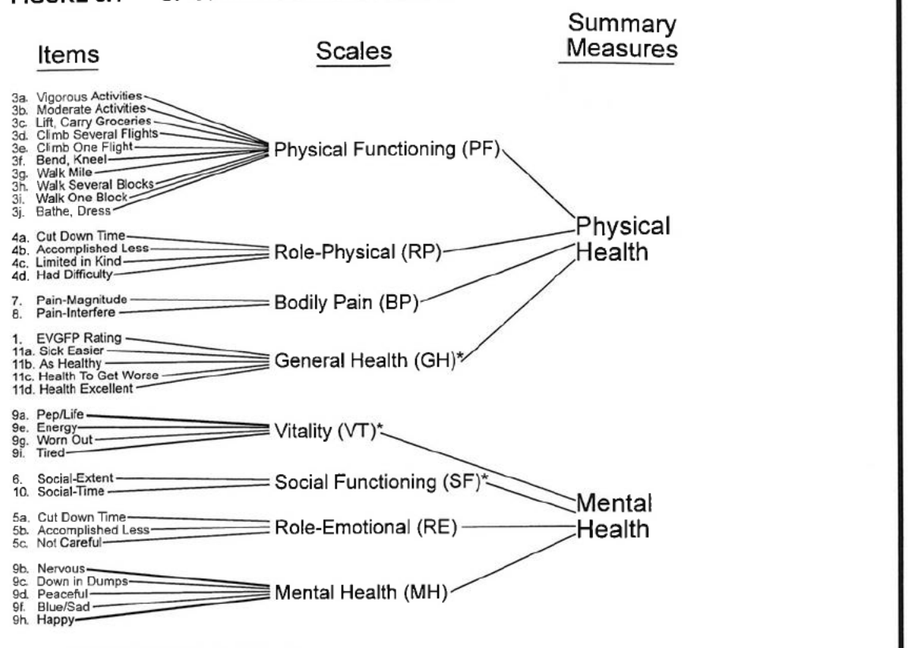
Factor analyses of correlations among the eight SF-36 scales have consistently identified two factors, as documented below. On the strength of the pattern of their correlations with the eight scales, they have been interpreted as "physical" and "mental" dimensions of health status. These physical and mental components accounted for 81.5% of the reliable variance in SF-36 scales in the general U.S. population (Ware, Kosinski, Bayliss, et al., 1995), and 82.4% in the MOS (McHorney, Ware, Raczek, 1993). Similar physical and mental components have been observed for other comprehensive surveys, including the Health Insurance Experiment Medical History Questionnaire (Ware, Brook, Davies-Avery, et al., 1980), the MOS Functioning and Well-Being Profile (Hays & Stewart, 1990), and the Sickness Impact Profile (Bergner, Bobbitt, Carter, et al., 1981). In further support of the generalizability of this two-dimensional model of health, we would expect factor analytic studies of other comprehensive surveys to yield recognizably similar physical and mental components.

The psychometric approach to summarizing health measures illustrated here is in contrast to a utility index, in which measures are aggregated without regard to their inter-relationships. A utility index achieves a single summary score at the expense of sensitivity and specificity to physical versus mental components of health status. A strength of the PCS and MCS summary scales described here is their value in distinguishing a physical from a mental health outcome.

SF-36 Measurement Model

Figure 3.1 illustrates the measurement model underlying the construction of SF-36 multi-item scales and summary measures. This model has three levels: (1) items, (2) scales that aggregate items, and (3) summary measures that aggregate scales. All but one of the 36 items (self-reported health transition) are used to score the eight SF-36 scales.

FIGURE 3.1 SF-36 MEASUREMENT MODEL



* Significant correlation with other summary measure.

As shown in the figure, each item is used in scoring only one scale. Tests of assumptions underlying the algorithms used in scoring the eight scales have been strongly supported in the U.S. (Ware, Snow, Kosinski, et al., 1993; McHorney, Ware, Lu, et al., 1994) and in other countries (Ware, Keller, Gandek, et al., 1998). The eight scales also form two distinct higher-ordered clusters. Three scales (Physical Functioning, Role-Physical, and Bodily Pain) correlate most highly with the physical component and contribute most to scoring of the PCS measure of that component. The mental component correlates most highly with the Mental Health, Role-Emotional, and Social Functioning scales, which contribute most to the scoring of the MCS measure of that component. Three of the scales have noteworthy correlations with both components: the Vitality scale correlates substantially with both; General Health correlates with both but higher with the physical component; and Social Functioning correlates much higher with the mental component. Reasons for these patterns of factor loadings are discussed elsewhere (Ware, Snow, Kosinski, et al., 1993).

Criteria And Hypotheses

Criteria commonly used to evaluate factor analyses using the principal components method were routinely applied (Harman, 1976) using SAS software (SAS Institute, 1989). First, we examined whether eigenvalues for the first two components were both greater than unity, as required for rotation. Other criteria, previously used, including the scree test, five percent rule, and common factor test, were also satisfied (Ware, Davies-Avery, & Brook, 1980). Second, we evaluated the pattern of correlations between the two rotated components and eight SF-36 scales to determine the basis for their interpretations as physical and mental components.

We hypothesized that the PF scale would have the highest and a very strong loading ($r > .80$) on the "physical" component followed by the RP and BP scales. The MH scale was hypothesized to have the strongest loading ($r > .80$) on the "mental" component, followed by the RE and SF scales. Because they are general measures, the GH and VT scales were hypothesized to correlate with both components. Lastly, we evaluated the adequacy of the two components in explaining the variation in each SF-36 scale. We hypothesized that a large proportion of the reliable variance in each scale (three-fourths) would be explained by the two components.

To evaluate the generalizability of the two-dimensional SF-36 model of health across different patient groups, we conducted principal component analyses in 23 subgroups of MOS patients differing in demographic characteristics and medical conditions. These are the same subgroups analyzed in tests of scaling assumptions for the eight SF-36 scales, as described elsewhere (McHorney, Ware, Lu, et al., 1994). These tests are extended here to include 12 subgroups differing in demographic characteristics from the general U.S. population, as well as general population samples from ten countries participating in the International Quality of Life Assessment (IQOLA) Project. In all analyses, we used the same method to extract two principal components from the correlations among the SF-36 scales and rotated them to orthogonal simple structure to facilitate interpretation.

Results

Tables 3.1 through 3.3 present results for the principal component analyses of SF-36 scales for the total 1998 general U.S. population standard (4-week recall) and acute (1-week recall) forms; the total patient sample from the MOS (McHorney, Ware, & Raczek, 1993); the total 1990 general U.S. population; and general populations in Denmark, France, Germany, Italy, Netherlands, Norway, Spain, Sweden, and the United Kingdom (Ware, Kosinski, Gandek, et al., 1998). Results for the 12 subgroups from the general U.S. population are summarized in Table 3.4. Results for the 23 patient subgroups from the MOS are summarized in Table 3.5. Results for age and gender subgroups in 9 countries are summarized in Table 3.6.

TABLE 3.1 CORRELATIONS BETWEEN SF-36 SCALES AND ROTATED PRINCIPAL COMPONENTS IN THE 1998 GENERAL U.S. POPULATION, STANDARD (N=1,982) AND ACUTE RECALL (N=1,646) FORMS

	Standard Form (4-Week Recall)			Acute Form (1-week Recall)		
Scales	**PCS**	**MCS**	**h^2/r_{tt}**	**PCS**	**MCS**	**h^2/r_{tt}**
Physical Function	.87	.13	.82	.87	.16	.83
Role Physical	.83	.28	.88	.82	.30	.88
Bodily Pain	.77	.31	.74	.79	.29	.84
General Health	.61	.50	.76	.64	.47	.75
Vitality	.43	.70	.76	.43	.69	.75
Social Function	.55	.65	.84	.55	.66	.86
Role Emotion	.25	.70	.70	.30	.73	.75
Mental Health	.10	.91	.98	.11	.90	.99
Reliable Variance[1]	81.3%			82.9%		

h^2/r_{tt} Variance in each SF-36 scale explained by the two principal components (h^2) divided by the reliability of each SF-36 scale (r_{tt}).

[1] Percent of the total reliable variance in SF-36 scales explained by the two principal components.

As hypothesized, eigenvalues for the first two components were both greater than unity in all analyses, strongly supporting the two-dimensional model of health.

TABLE 3.2 CORRELATIONS BETWEEN SF-36 SCALES AND ROTATED PRINCIPAL COMPONENTS IN THE MEDICAL OUTCOMES STUDY (MOS) AND THE GENERAL U.S. POPULATION

	MOS (N = 3,445)			U.S. POPULATION (N = 2,474)		
SF-36 SCALES	**PCS**	**MCS**	**h^2/r_{tt}**	**PCS**	**MCS**	**h^2/r_{tt}**
Physical Functioning (PF)	.88	.04	.84	.85	.12	.78
Role Physical (RP)	.78	.30	.83	.81	.27	.82
Bodily Pain (BP)	.77	.24	.79	.76	.28	.72
General Health (GH)	.68	.32	.72	.69	.37	.78
Vitality (VT)	.59	.57	.77	.47	.64	.75
Social Functioning (SF)	.44	.71	.82	.42	.67	.92
Role Emotional (RE)	.19	.81	.83	.17	.78	.78
Mental Health (MH)	.12	.90	.91	.17	.87	.92
Reliable Variance[1]			82.4%			81.5%

h^2/r_{tt} Variance in each SF-36 scale explained by the two principal components (h^2) divided by the reliability of each SF-36 scale (r_{tt}).

[1] Percent of the total reliable variance in SF-36 scales explained by the two principal components.

The pattern of correlations between SF-36 scales and the two rotated components in Tables 3.1 and 3.2 strongly support the physical and mental health interpretation of the two components. In each analysis, the PF scale loaded strongest on the "physical" component and weakest on the "mental" component. Also, the RP and BP scales both had stronger loadings on the "physical" component than the "mental" component. The MH scale had the highest loading and the RE and SF scales had stronger loadings on the "mental" component than the "physical" component. The GH and VT scales had noteworthy loadings (r > .30) on both physical and mental components, as expected for general measures.

TABLE 3.3 CORRELATIONS BETWEEN SF-36 SCALES AND ROTATED PRINCIPAL COMPONENTS IN 9 COUNTRIES

	Denmark		France		Germany		Italy		Netherlands		Norway		Spain		Sweden		United Kingdom	
Scales	**PC**	**MC**	**PC**	**MC**	**PC**	**MC**	**PC**	**MC**	**PC**	**MC**	**PC**	**MC**	**PC**	**MC**	**PC**	**MC**	**PC**	**MC**
PF	0.86	0.10	0.85	0.12	0.85	0.16	0.86	0.13	0.86	0.16	0.83	0.15	0.81	0.19	0.84	0.15	0.84	0.20
RP	0.78	0.30	0.78	0.29	0.79	0.28	0.64	0.42	0.69	0.44	0.82	0.26	0.76	0.24	0.78	0.30	0.80	0.27
BP	0.69	0.32	0.78	0.32	0.82	0.24	0.73	0.37	0.80	0.20	0.80	0.25	0.74	0.26	0.76	0.28	0.80	0.19
GH	0.70	0.41	0.60	0.49	0.72	0.38	0.75	0.34	0.75	0.31	0.74	0.39	0.73	0.34	0.66	0.50	0.73	0.39
VT	0.39	0.75	0.41	0.73	0.52	0.64	0.44	0.70	0.52	0.66	0.40	0.71	0.51	0.61	0.43	0.74	0.54	0.64
SF	0.33	0.72	0.37	0.78	0.33	0.75	0.27	0.81	0.49	0.68	0.33	0.78	0.50	0.62	0.32	0.78	0.55	0.62
RE	0.27	0.70	0.28	0.71	0.20	0.71	0.23	0.76	0.13	0.84	0.21	0.78	0.18	0.78	0.25	0.71	0.20	0.78
MH	0.11	0.89	0.09	0.91	0.16	0.87	0.25	0.81	0.25	0.86	0.14	0.90	0.22	0.84	0.15	0.90	0.18	0.87
Reliable Variance[1]	81.3		84.7		82.9		84.0		82.1		83.8		76.3		82.6		82.0	

Abbreviations: PC = Physical Component; MC = Mental Component; PF = Physical Functioning; RP = Role Physical; BP = Bodily Pain; GH = General Health; VT = Vitality; SF = Social Functioning; RE = Role Emotional; MH = Mental Health.

Source: Ware JE, Kosinski M, Gandek BG, et al.; 1998.

Table 3.4 summarizes the range of factor loadings observed in 12 analyses of subgroups of the general U.S. population differing in age, gender, race, and education. The two components explained from 77% to 87% (median = 81%) of the reliable variance in the eight scales, and the range of correlations observed for each scale and each component is strikingly similar to that observed in Table 3.1 for the general U.S. population.

Table 3.5 summarizes results for analyses of 23 subgroups of MOS patients differing in sociodemographic characteristics, diagnosis, and other clinical variables defined elsewhere (McHorney, Ware, Lu, et al., 1994). From 75% to 83% (median = 80%) of the reliable variance in the eight scales was explained. Again, the amount of reliable variance explained and the pattern of correlations between scales and the two components replicates other results to date.

Table 3.6 presents therange of correlations between SF-36 scales and rotated principal components for five subgroups differing in age (three groups) and gender (two groups) within each of the nine countries. The pattern of correlations observed are consistent with the patterns observed for the total populations in each country (see Table 3.3). Only in Italy are departures observed, particularly among the 65 and older group.

The results presented in Tables 3.1-3.6 also demonstrate the adequacy of the two components in explaining the reliable variance in each SF-36 scale. In all but four of the 40 factor analyses, the total reliable variance in SF-36 scales explained by the two components exceeded 81%. As hypothesized, the two components explained at least three-fourths of the reliable variance in each SF-36 scale, with few exceptions.

TABLE 3.4 RANGE OF CORRELATIONS BETWEEN SF-36 SCALES AND ROTATED PRINCIPAL COMPONENTS, 12 SUBGROUPS OF THE GENERAL U.S. POPULATION[1] (N = 2,474)

SF-36 SCALES	PCS	MCS
Physical Functioning (PF)	.77 - .88	.05 - .30
Role Physical (RP)	.67 - .82	.16 - .43
Bodily Pain (BP)	.70 - .84	.17 - .46
General Health (GH)	.53 - .76	.29 - .69
Vitality (VT)	.31 - .73	.44 - .82
Social Functioning (SF)	.34 - .62	.46 - .73
Role Emotional (RE)	.06 - .48	.57 - .83
Mental Health (MH)	.11 - .27	.84 - .90

1 Results above come from analyses of 12 subgroups in the general U.S. population: 3 age, 2 gender, 3 race, and 4 education groups.

TABLE 3.5 RANGE OF CORRELATIONS BETWEEN SF-36 SCALES AND ROTATED PRINCIPAL COMPONENTS, 23 MOS SUBGROUPS[1] (N = 3,445)

SF-36 SCALES	PCS	MCS
Physical Functioning (PF)	.83 - .90	.01 - .16
Role Physical (RP)	.69 - .81	.20 - .41
Bodily Pain (BP)	.65 - .83	.12 - .42
General Health (GH)	.56 - .76	.17 - .51
Vitality (VT)	.43 - .79	.27 - .66
Social Functioning (SF)	.30 - .72	.37 - .82
Role Emotional (RE)	.11 - .38	.66 - .88
Mental Health (MH)	.01 - .34	.75 - .92

1 Results above come from principal components analyses of 23 subgroups in the MOS: 3 age, 2 gender, 3 race, 4 education, 2 poverty status, 6 diagnoses, and 3 clinical groups.

TABLE 3.6 RANGE OF CORRELATIONS BETWEEN SF-36 SCALES AND ROTATED PRINCIPAL COMPONENTS ACROSS AGE AND GENDER GROUPS IN 9 COUNTRIES

	Denmark		France		Germany		Italy		Netherlands	
Scale	Physical	Mental	Physical	Mental	Physical	Mental	Physical	Mental	Physical	Mental
PF	.79-.86	.04-.16	.82-.86	.06-.19	.81-.86	.10-.24	.81-.86	.04-.20	.81-.86	.09-.20
RP	.72-.80	.25-.39	.68-.78	.25-.41	.73-.81	.26-.31	.60-.83	.29-.42	.60-.75	.30-.47
BP	.68-.79	.23-.37	.73-.81	.27-.34	.76-.86	.17-.32	.43-.77	.31-.63	.76-.85	.18-.26
GH	.58-.78	.33-.56	.56-.65	.44-.48	.62-.75	.33-.46	.53-.77	.33-.57	.58-.77	.22-.49
VT	.29-.56	.64-.79	.32-.51	.69-.76	.34-.69	.46-.76	.30-.48	.66-.82	.37-.70	.46-.74
SF	.22-.39	.69-.75	.33-.47	.71-.80	.27-.38	.70-.78	.26-.29	.76-.81	.38-.58	.56-.75
RE	.13-.44	.55-.79	.15-.41	.63-.76	.14-.24	.68-.73	.19-.61	.40-.79	.09-.24	.73-.91
MH	.06-.16	.86-.91	.07-.17	.90-.92	.12-.25	.82-.89	.18-.27	.80-.86	.11-.19	.70-.89

	Norway		Spain		Sweden		United Kingdom	
Scale	Physical	Mental	Physical	Mental	Physical	Mental	Physical	Mental
PF	.77-.85	.12-.22	.69-.81	.07-.23	.81-.85	.08-.29	.76-.86	.15-.26
RP	.76-.83	.21-.35	.62-.80	.14-.40	.75-.78	.26-.33	.76-.81	.21-.33
BP	.79-.83	.19-.27	.67-.77	.19-.35	.74-.80	.26-.30	.78-.81	.14-.27
GH	.66-.79	.36-.41	.49-.83	.19-.42	.61-.69	.47-.54	.57-.78	.35-46
VT	.37-.52	.66-.72	.19-.53	.59-.76	.42-.62	.64-.74	.37-.59	.59-.70
SF	.29-.46	.65-.80	.43-.53	.55-.66	.28-.35	.77-.79	.43-.56	.61-.74
RE	.10-.28	.75-.84	.13-.19	.66-.90	.11-.45	.64-.77	.10-.27	.76-.81
MH	.12-.20	.86-.91	.03-.36	.70-.86	.13-.19	.88-.91	.11-.29	.82-.89

Methodological Issues

Factor analysis has proven to be very useful in testing hypotheses about the structure of health and in evaluating the construct validity of the SF-36 and other health surveys (Ware, 1976; Goldberg & Hillier, 1979; Ware, Brook, & Davies-Avery, 1980; Veit & Ware, 1983; Derogatis, 1986; Hall, Epstein, & McNeil, 1989; Wiklund, Lindvall, Swedberg, et al., 1987; Mason, Anderson, & Meenan, 1988; Schag, Heinrich, Aadland, et al., 1990; Hays & Stewart, 1990; McHorney, Ware, & Raczek, 1993). We have given considerable attention to the implications of different methods of factor extraction and rotation. In many cases, conclusions do not vary across methods. When they do, the choice among methods depends on the purpose(s) of the factor analysis (Nunnally & Bernstein, 1994).

The choice of method for the SF-36 studies follows from the first author's earlier work and considerations of work published by others (Ware, Miller, & Snyder, 1973; Snyder & Ware, 1974; Ware & Snyder, 1975; Ware, Davies-Avery, & Brook, 1980). Consistent with guidelines suggested by Harris & Harris (1971), conclusions about the factor structure of the SF-36 were not of great consequence in choosing a method. That structure is robust across methods and populations. In fact, a good test of a structural model is its robustness across factor analytic methods (Harris & Harris, 1971). For example, comparisons across methods for the same matrices were often employed during the development of the Health Perceptions Questionnaire (Ware, Miller, & Snyder, 1973; Ware, 1976; Ware & Karmos, 1976), from which items were selected for the SF-36 GH Scale. Those studies also demonstrated the advantages of homogeneous, short, multi-item scales over single-item measures as the unit of analysis in factor analytic studies. These advantages are also well-documented in empirical studies of personality variables (Comrey, 1973).

The two-dimensional factor structure of the eight SF-36 scales has also been shown to satisfy criteria for "simple structure" (Nunnally & Bernstein, 1994) across patient and general population samples in the U.S. and across other countries (as summarized above). To facilitate re-analyses by others, the matrices of correlations for the SF-36 in general and patient populations are reproduced at the end of this section (see Tables 3.5 and 3.6).

Principal Components

As noted above, the interpretation of the first two factors as dimensions of physical and mental health has been straightforward and robust across methods. Thus, the choice of factor analytic method was not governed by considerations in interpreting the factors. The choice of the principal components method was based on other considerations, including the ease of estimation of factor scores for the two summary measures, estimation of the factor content of the eight SF-36 scales in relation to physical and mental health status, the explanatory power of the factors, and their validity in discriminating between physical and mental dimensions of health status. These considerations are discussed briefly below.

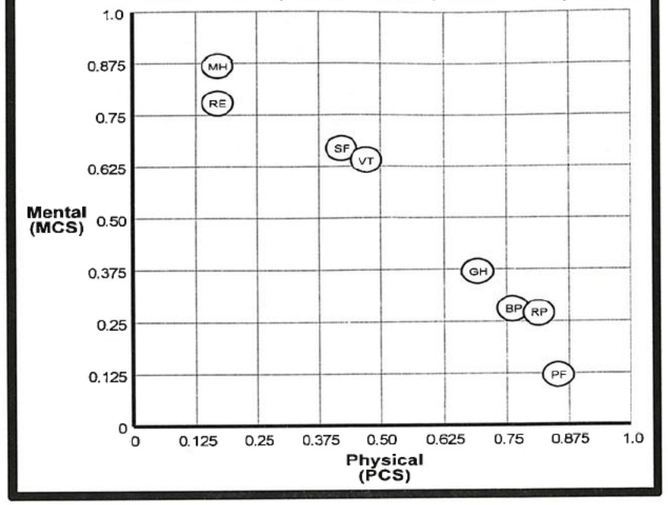
The advantages of components analysis over principal factor analysis are noteworthy for the purposes of achieving: (1) a simple additive model of factor content facilitating the interpretation of each scale; (2) summary measures that explain as much of the variance in the eight scales as possible; (3) summary scales that are easy to estimate statistically; and (4) summary scores that are interpretable as physical and mental dimensions of health. The goal in constructing the PCS and MCS scores was to explain as much of the variance in the eight SF-36 variables as possible with only two summary measures (Ware, Kosinski, Bayliss, et al., 1995). Components analysis attempts to explain as much of the variance as possible, in contrast to principal factors, which attempt to reproduce the original correlation matrix (Harman, 1976). Second, as explained in Chapter 4, the computation of scores for each principal component is a straightforward estimation using scores for the observed variables (i.e., the eight SF-36 scale scores) in contrast to approximations involved in estimating scores for principal factors. These differences and the advantages of components analysis are discussed in numerous texts on factor analysis and psychometric methods (e.g., Harman, 1976; Nunnally & Bernstein, 1994).

There are good theoretical arguments for both orthogonal and oblique factor rotations (Harman, 1976; Nunnally & Bernstein, 1994). As argued in numerous texts, orthogonal components are ideal for our purposes. Our initial objective in factor analyzing the correlations among SF-36 scales was to test the construct validity of the SF-36 and to establish guidelines for interpreting each scale, on the basis of its physical and mental health factor content (McHorney, Ware, & Raczek, 1993; Ware, Snow, Kosinski, et al., 1993; Ware, Kosinski, Bayliss, et al., 1995). For this purpose, orthogonal components, which are not correlated, have clear advantages. For example, "factor loadings," which are product moment correlations between scales and factors, can be squared and summed across factors to estimate the amount of variance in each scale accounted for by each factor and the amount of variance in each scale that is explained by all factors (i.e., the communality). As a result, factor content and implications for the interpretation of each scale are more straightforward (Ware, Snow, Kosinski, et al., 1993).

Orthogonal Components

To provide a visual image of the contribution of the eight SF-36 scales to the physical and mental components, their factor loadings are plotted in Figure 3.2 for the general U.S. population. This plot reveals a progression from the upper left corner, with the MH scale correlating most highly with the mental component (MCS) and least with the physical component (PCS), to the lower right corner, where the highest correlation with the PCS is observed for the PF scale, which also correlates the least with the MCS. In between is a progression of loadings from the lower right (PF) to the upper left (MH). This ordering of scales from PF in the lower right to MH in the upper left corresponds to their ordering in the SF-36 profile.

FIGURE 3.2 PLOT OF SF-36 SCALE FACTOR LOADINGS ON ORTHOGONAL PHYSICAL AND MENTAL COMPONENTS, GENERAL U.S. POPULATION (N=2,474)



It is also apparent that the eight SF-36 scales form two distinct clusters with four scales (MH, RE, SF, VT) correlating highest with the MCS and lowest with the PCS, and a second cluster (PF, RP, BP, GH) correlating highest with the PCS and lowest with the MCS. As shown in Chapter 4, the factor score coefficients used to score the PCS and MCS correspond to these two clusters. Specifically, the highest positive coefficients (0.42 to 0.25) are used to weight the four best physical scales in scoring the PCS, and the highest positive coefficients (0.49 to 0.24) are used to weight the four best mental scales in scoring the MCS. Because the factor score coefficients take into account the correlations *among* the eight scales, they differ from the factor loadings; some are negative. Negative factor score coefficients were also observed in oblique principal factor solutions and are not unique to the principal components method (data not reported). Oblique solutions, which can allow substantial correlations between health dimensions (factors), can facilitate the identification of factors but they complicate understanding of the factor content of scales because loadings are not additive in an oblique solution. Correlations among factors and factor loadings must both be taken into account in interpreting an oblique solution, complicating the interpretation of each scale.

If the PF and MH scales had proven to be substantially correlated, or if the PCS and MCS were substantially correlated on cross-validation, there would be good reason to favor an oblique solution. However, it is clear that physical and mental health are only weakly positively correlated. Correlations between the PF and MH scales, the best physical and mental health measures among the eight SF-36 scales, are low, with medians ranging from only 0.22 to 0.30 in 39 patient and general population studies in the U.S., Germany, Sweden, and the U.K. Cross validation of the orthogonal two-dimensional model (using U.S. factor score coefficients) in these samples has demonstrated very low empirical correlations between

the PCS and MCS scores (medians of -0.01 to 0.07 across 39 estimates in studies to date). These correlations would be much larger upon cross-validation if the orthogonal solution and scoring were a gross distortion.

Additional convincing empirical evidence favoring the scoring of orthogonal principal components in summarizing SF-36 information about physical and mental health is their superiority in discriminating between physical and mental health outcomes in empirical tests. Comparisons of alternative scoring strategies revealed that much of the advantage in interpretation gained with the SF-36 PCS and MCS is lost when the physical and mental components are scored with an oblique solution (data not shown).

TABLE 3.7 PRODUCT MOMENT CORRELATIONS AND RELIABILITY COEFFICIENTS (IN PARENTHESES), SF-36 SCALES (4-WEEK RECALL) FOR THE 1998 GENERAL U.S. POPULATION

	PF	RP	BP	GH	VT	SF	RE	MH
PF	(.94)							
RP	.66	(.88)					N=1,982	
BP	.57	.65	(.85)					
GH	.57	.54	.57	(.84)				
VT	.44	.52	.54	.63	(.88)			
SF	.54	.64	.59	.56	.60	(.86)		
RE	.35	.46	.38	.39	.42	.58	(.82)	
MH	.26	.34	.38	.51	.63	.59	.51	(.85)

TABLE 3.8 PRODUCT MOMENT CORRELATIONS AND RELIABILITY COEFFICIENTS (IN PARENTHESES), SF-36 SCALES (1-WEEK RECALL) FOR THE 1998 GENERAL U.S. POPULATION

	PF	RP	BP	GH	VT	SF	RE	MH
PF	(.94)							
RP	.69	(.88)					N=1,646	
BP	.62	.66	(.86)					
GH	.58	.57	.60	(.84)				
VT	.46	.54	.53	.61	(.88)			
SF	.55	.66	.59	.58	.60	(.86)		
RE	.42	.52	.41	.43	.48	.64	(.83)	
MH	.29	.35	.39	.50	.61	.59	.55	(.83)

TABLE 3.9 PRODUCT MOMENT CORRELATIONS AND RELIABILITY COEFFICIENTS (IN PARENTHESES), SF-36 SCALES FOR THE 1990 GENERAL U.S. POPULATION

	PF	RP	BP	GH	VT	SF	RE	MH
PF	(.93)							
RP	.65	(.89)					N=2.474	
BP	.52	.61	(.90)					
GH	.55	.55	.56	(.81)				
VT	.44	.50	.52	.58	(.86)			
SF	.45	.52	.49	.47	.51	(.68)		
RE	.30	.42	.32	.35	.44	.53	(.82)	
MH	.28	.35	.39	.46	.63	.56	.54	(.84)

TABLE 3.10 PRODUCT MOMENT CORRELATIONS AND RELIABILITY COEFFICIENTS (IN PARENTHESES), SF-36 SCALES FOR THE MOS (N = 3,445)

	PF	RP	BP	GH	VT	SF	RE	MH
PF	(.93)							
RP	.60	(.84)						
BP	.56	.62	(.82)					
GH	.53	.48	.50	(.78)				
VT	.49	.57	.50	.57	(.87)			
SF	.45	.52	.53	.47	.57	(.85)		
RE	.29	.48	.34	.37	.51	.55	(.83)	
MH	.20	.34	.38	.41	.58	.66	.60	(.90)

SF-12 Health Survey

The construction of the PCS and MCS and evidence to date supporting their usefulness in cross-sectional and longitudinal tests provides 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 most purposes, as suggested by results reported here and elsewhere (Ware, Kosinski, Bayliss, et al., 1995), it is likely that the two summary scores can be estimated with fewer items well enough for some purposes.

Following discovery of the PCS and MCS summary scales, we began evaluating how well these scales could be reproduced using much shorter questionnaires. Those studies identified ten items from six of the eight SF-36 scales that reproduced at least 90% of the variance in both the PCS and MCS, as defined using the SF-36 scales. Addition of two more items created a 12-item short-form yielding satisfactory estimates of the PCS and MCS, as well as scores for the two additional concepts necessary to represent the profile of eight SF-36 concepts. We refer to this new short form as the SF-12. Four of the eight concepts (physical functioning, role-physical, role-emotional, and mental health) are estimated using two items each. The remaining scales (bodily pain, general health, vitality, and social functioning) are measured using one item each. The eight-concept profile based on SF-12 items (calibrated to reproduce the original eight 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 scales and very short multi-item scales has been demonstrated in previous studies (McHorney, Ware, Rogers, et al., 1992).

Results from preliminary tests of empirical validity, like those described in Chapter 6 for the SF-36 PCS and MCS summary scales, suggest that 12-item versions of the PCS and MCS will correlate with the SF-36 versions in the 0.93 to 0.97 range, upon cross-validation. However, the SF-12-based PCS and MCS define fewer levels and should be expected to yield less reliable assignments of individuals to those levels. For large group studies these differences in measurement reliability are not as important, because confidence intervals for group averages are more determined by sample

size. Given that the SF-12 can be self-administered in about two to three minutes, the trade-off may be worthwhile for many purposes.

We have prepared documentation of items and scoring algorithms for the SF-12 (Ware, Kosinski, & Keller, 1995), and results of the preliminary evaluation of empirical validity for the SF-12 have been published (Ware, Kosinski & Keller, 1995). The SF-12 should be most useful for purposes of large group and population monitoring of health status and changes in health over time. Because SF-12 items are a subset of the original SF-36, complete comparability can be maintained across data sets using SF-12 scoring algorithms and data for either short-form survey.

4. Check correlation's between the General Health scale and the other seven scales to verify that all are positive; with rare exceptions they should also be substantial in magnitude (0.30 or higher).

5. For those familiar with principal factor or components analysis, inspect correlations between the eight scales and the first unrotated factor or component extracted from the correlations among those scales. Regardless of extraction method, these correlations should be positive and substantial in magnitude (0.30 or higher).

Chapter 6: Norm-Based Scoring For Physical & Mental Summary Measures

This chapter provides scoring instructions for the SF-36 physical (PCS) and mental (MCS) component summary measures. Scoring of the SF-36 PCS and MCS summary measures involves three steps. First, the eight SF-36 scales are standardized using means and standard deviations from the 1998 general U.S. population. Second, they are aggregated using weights (factor score coefficients) from the 1990 general U.S. population. These are the same weights as those used to score PCS and MCS from the SF-36 Version in the first edition of this manual (Ware 1994). Finally, aggregate PCS and MCS scores are standardized using a linear T-score transformation to have a mean of 50 and a standard deviation of 10, in the 1998 general U.S. population. Norm-based scoring of the standard (4-week recall) and acute (1 week recall) forms is presented separately.

General U.S. population statistics used in the standardization and in the aggregation of SF-36 scale scores are presented in Table 6.1 for the standard form and in Table 6.2 for the acute form. Note that for both standard and acute forms the same factor score coefficients are used to score PCS and MCS. To make sure that the original coefficients used in this scoring "recipe" have the same affect in 1998, as in 1990, 1998 variances are utilized. Detailed information including formulas for scale aggregation and transformation of scores are presented below. Formal checks using a test dataset available upon request (info@qmetric.com) can be performed to confirm the successful reproduction of PCS and MCS scales, as discussed later in the chapter. We strongly recommend these tests.

Importance Of Standardization

As with the 1998 NBS scoring of SF-36 scales, which are aggregated to score the summary measures, standardization of the scoring of the PCS and MCS scales is vital to their interpretation. Any changes in scoring of the SF-36 scales or the algorithms for the summary measures may compromise their reliability and validity. Changes in scoring may also invalidate normative comparisons, based on the 1998 norms documented here.

The methods documented here achieve 1998 scores that are a nearly perfect transformation of the 1990 scores. For scales and summary measures, NBS based on 1998 norms only shifts the score distribution to better reflect the health of the US population in 1998. Otherwise, 1998 scores have the same interpretations as 1990 scores.

Norm-Based Scoring Of PCS And MCS, Standard Form (4-Week Recall)

Standard form PCS and MCS scales are scored using norm-based methods. The means and standard deviations used in scoring come from the 1998 general U.S. population and the factor score coefficients come from the 1990 general U.S. population (Ware et al. 1994). A linear T-score transformation method is used so that both the PCS and MCS have a mean of 50 and a standard deviation of 10 in the 1998 general U.S. population.

The advantage of the standardization and norm-based scoring of the PCS and MCS 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 1998 general U.S. population. Because the standard deviation is 10 for both PCS and MCS measures, each one point difference in scores also has a direct interpretation. A one point difference is one-tenth of a standard deviation.

TABLE 6.1 1998 GENERAL U.S. POPULATION MEANS, STANDARD DEVIATIONS AND 1990 FACTOR SCORE COEFFICIENTS USED TO DERIVE PCS AND MCS SCALE SCORES, STANDARD FORM

			Factor Score Coefficients	
SF-36 SCALE	MEAN*	STANDARD DEVIATION	PCS	MCS
PF	82.96845	23.83795	0.42402	-0.22999
RP	77.93107	35.34865	0.35119	-0.12329
BP	70.22865	23.35310	0.31754	-0.09731
GH	70.10060	21.35900	0.24954	-0.01571
VT	56.99917	21.12677	0.02877	0.23534
SF	83.56494	23.02758	-0.00753	0.26876
RE	83.10276	31.64149	-0.19206	0.43407
MH	75.21913	17.60698	-0.22069	0.48581

*__Note__: The means and standard deviations for each SF-36 scale are based on the 0-100 scoring.

It should be noted that mean scores for most of the scales vary substantially form those from the 1990 norms. For those scales that have been changed in any way from 1990 to 1998, several explanations are being explored (e.g., sampling bias in the 1990 study, downward temporal trends in physical health). Regardless of the explanations, these differences underscore the importance of using up-to-date norms and also the importance of using the same normative data when equating Versions 1 and 2 of the SF-36. 1998 norms for both make these comparisons possible.

Steps In Scoring

Following the scoring of the eight scales according to the standard SF-36 scoring algorithms (0-100 scale) explained in the SF-36 Manual and Interpretation Guide (Ware et al., 1993, PCS and MCS are scored in three steps as explained below:

Standardization Of Scales (Z-Scores), Standard Form

The first step in computing PCS and MCS consists of standardizing each of the 8 SF-36 scales using a z-score transformation. This is the same as Step 1 used in the norm-based scoring of the 8 SF-36 scales explained in Chapter 5. A z-score for each scale is computed by subtracting the mean 0-100 1998 general US population score (see Table 6.1) for each SF-36 scale and dividing the difference by the corresponding scale standard deviation. Note that the SF-36 scales scored on the 0-100 scale are used in step 1. Norm-based SF-36 scale scores are not used in this step. Formulas are listed below.

STEP 1. Formulas For Z-Score Standardization Of SF-36 Scales, (Standard Form)

PF_Z = (PF – 82.96845) / 23.83795
RP_Z = (RP – 77.93107) / 35.34865
BP_Z = (BP – 70.22865) / 23.35310
GH_Z = (GH – 70.10060) / 21.35900
VT_Z = (VT – 56.99917) / 21.12677
SF_Z = (SF – 83.56494) / 23.02758
RE_Z = (RE – 83.10276) / 31.64149
MH_Z = (MH – 75.21913) / 17.60698

Means and standard deviations are from Table 6.1.

Aggregation Of Scale Scores (Standard Form)

After a z-score has been computed for each SF-36 scale, the second step involves computation of aggregate scores for the physical and mental components using the physical and mental factor score coefficients from the 1990 general U.S. population as given in Table 6.1.

Computation of an aggregate physical component score consists of multiplying each SF-36 scale z-score by its respective physical factor score coefficient and summing the eight products, as shown below. Similarly, an aggregate mental component score is obtained by multiplying each SF-36 scale z-score by its respective mental factor score coefficient and summing the eight products.

STEP 2. Formulas For Aggregating Scales In Estimating Aggregate Physical And Mental Component Scores (Standard Form)

AGG_PHYS = (PF_Z*.42402) + (RP_Z*.35119) +
(BP_Z*.31754) + (GH_Z*.24954) + (VT_Z * .02877) +
(SF_Z*-.00753) + (RE_Z*-.19206) + (MH_Z*-.22069)

AGG_MENT= (PF_Z*-.22999) + (RP_Z*-.12329)
+ (BP_Z*-.09731) + (GH_Z*-.01571) + (VT_Z*.23534) +
(SF_Z*.26876) + (RE_Z*.43407) + (MH_Z*.48581)

Transformation Of Summary Scores

The third step involves transforming each component score to the norm-based (50, 10) scoring. This is accomplished by multiplying each aggregate component scale score by 10 and adding the resulting product to 50. Formulas are shown in step 3.

STEP 3. FORMULAS FOR T-SCORE TRANSFORMATION OF COMPONENT SCORES (STANDARD FORM)

Transformed Physical (PCS) = 50 + (AGG_PHYS * 10)
Transformed Mental (MCS) = 50 + (AGG_MENT * 10)

Norm-Based Scoring Of PCS And MCS, Acute Form

Acute form PCS and MCS scales are scored using norm-based methods. The means and standard deviations used in scoring come from the 1998 general U.S. population and the factor score coefficients come from the 1990 general U.S. population documented in the first edition of this manual (Ware et al. 1994). A linear T-score transformation method is used so that both the PCS and MCS have a mean of 50 and a standard deviation of 10 in the 1998 general U.S. population.

The advantage of the standardization and norm-based scoring of the PCS and MCS 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 1998 general U.S. population. Because the standard deviation is 10 for both PCS and MCS measures, each one point difference in scores also has a direct interpretation. A one point difference is one-tenth of a standard deviation.

TABLE 6.2 1998 GENERAL U.S. POPULATION MEANS, STANDARD DEVIATIONS AND 1990 FACTOR SCORE COEFFICIENTS USED TO DERIVE PCS AND MCS SCALE SCORES, ACUTE FORM

			FACTOR SCORE COEFFICIENTS	
SF-36 SCALE	MEAN*	STANDARD DEVIATION	PCS	MCS
PF	82.34934	24.55363	0.42402	-0.22999
RP	78.37950	35.25259	0.35119	-0.12329
BP	73.37809	24.09701	0.31754	-0.09731
GH	70.69938	21.60431	0.24954	-0.01571
VT	57.64847	21.81907	0.02877	0.23534
SF	85.61860	22.83605	-0.00753	0.26876
RE	82.44273	32.52898	-0.19206	0.43407
MH	76.23511	17.25352	-0.22069	0.48581

*__Note__: The means and standard deviations for each SF-36 are based on the 0-100 scoring.

Steps In Scoring

Following the scoring of the eight scales according to the standard SF-36 scoring algorithms (0-100 scale) explained earlier in SF-36 Manual and Interpretation Guide (Ware et al., 1993), PCS and MCS are scored in three steps as explained below:

Standardization Of Scales (Z-Scores), Acute Form

The first step in computing PCS and MCS consists of standardizing each of the 8 SF-36 scales using a z-score transformation. This is the same as Step 1 used in the norm-based scoring of the 8 SF-36 scales explained in Chapter 5. A z-score for each scale is computed by subtracting the mean 0-100 1998 general US population score (see Table 6.2) for each SF-36 scale and dividing the difference by the corresponding scale standard deviation. Note that the SF-36 scales scored on the 0-100 scale are used in step 1. Norm-based SF-36 scale scores are not used in this step. Formulas are listed below.

STEP 1. FORMULAS FOR Z-SCORE STANDARDIZATION OF SF-36 SCALES (ACUTE VERSION)

PF_Z = (PF –82.34934) / 24.55363
RP_Z = (RP – 78.37950) / 35.25259
BP_Z = (BP – 73.37809) / 24.09701
GH_Z = (GH – 70.69938) / 21.60431
VT_Z = (VT – 57.64847) / 21.81907
SF_Z = (SF – 85.61860) / 22.83605
RE_Z = (RE – 82.44273) / 32.52898
MH_Z = (MH – 76.23511) / 17.25352

Means and standard deviations are from Table 6.2

Aggregation Of Scale Scores (Acute Form)

After a z-score has been computed for each SF-36 scale, the second step involves computation of aggregate scores for the physical and mental components using the physical and mental factor score coefficients from the 1990 general U.S. population as given in Table 6.2.

Computation of an aggregate physical component score consists of multiplying each SF-36 scale z-score by its respective physical factor score coefficient and summing the eight products, as shown in step 2. Similarly, an aggregate mental component score is obtained by multiplying each SF-36 scale z-score by its respective mental factor score coefficient and summing the eight products.

STEP 2. FORMULAS FOR AGGREGATING STANDARDIZED SCALES IN ESTIMATING AGGREGATE PHYSICAL AND MENTAL COMPONENT SCORES (ACUTE FORM)

AGG_PHYS = (PF_Z * .42402) + (RP_Z * .35119) + (BP_Z * .31754) + (GH_Z * .24954) + (VT_Z * .02877) + (SF_Z * -.00753) + (RE_Z * -.19206) + (MH_Z * -.22069)

AGG_MENT = (PF_Z * -.22999) + (RP_Z * -.12329) + (BP_Z * -.09731) + (GH_Z * -.01571) + (VT_Z * .23534) + (SF_Z * .26876) + (RE_Z * .43407) + (MH_Z * .48581)

Transformation Of Summary Scores

The third step involves transforming each component score to the norm-based (50, 10) scoring. This is accomplished by multiplying each aggregate component scale score by 10 and adding the resulting product to 50. Formulas are shown in step 3.

STEP 3. FORMULAS FOR T-SCORE TRANSFORMATION OF COMPONENT SCORES

Transformed Physical (PCS) = 50 + (AGG_PHYS * 10)
Transformed Mental (MCS) = 50 + (AGG_MENT * 10)

Missing Data Estimation

Results from ongoing evaluations of options to score PCS and MCS when a respondent is missing any one of the eight SF-36 scales has shown considerable promise. Evaluation of missing data rates across general and clinical populations has shown that 50% of those who have missing PCS and MCS scores are missing PCS and MCS because of missing data on one SF-36 scale. QualityMetric's Missing Data Estimator (MDE) will enable you to score PCS and MCS with data for 7 of the 8 SF-36 scales. PCS and MCS scores from MDE have proven to be reliable and valid (Ware et al., forthcoming). Further efforts are underway to evaluate options for scoring PCS and MCS when more than 1 SF-36 scale score is missing. More information will be available at these websites www.sf-36.com and www.qmetric.com.

Until QualityMetric's MDE is made available, it is recommended that component scale scores be set to missing if the respondent is missing any one of the eight SF-36 scales. To minimize the number of component scores missing, we recommend estimating each of the eight scale scores if half or more of the items are complete, as documented earlier.

Features Of PCS And MCS Scores

The PCS and MCS were constructed and scored to achieve a number of advantages, in addition to reducing the SF-36 from an eight-scale profile to two summary measures without substantial loss of information. Features of the PCS and MCS scores, including their reliability, confidence intervals (CI), skewness (percent ceiling and floor), and number of levels observed in the general U.S. population, are summarized in Table 6.3. These results confirm some of the theoretical advantages of the two summary measures as compared to the eight SF-36 scales, including a very large increase in the number of levels defined, smaller confidence intervals relative to each of the eight scales, as well as the elimination of both floor and ceiling effects.

TABLE 6.3 COMPARISON OF FEATURES OF SF-36 SCALES AND SUMMARY MEASURES, 1998 GENERAL U.S. POPULATION

		SUMMARY MEASURES	
	SF-36 SCALES[b]	PCS	MCS
Reliability	.84 - .95	.95	.93
95% CI (±)	12 – 17	4.5	5.2
% Floor	0.0 – 2.1	0	0
% Ceiling	2.1 – 60.2	0	0
# of Levels[a]	8 - 20	486	494

[a] Scores rounded to first decimal place
[b] Statistics are presented as the range of results found across the eight SF-36 scales (standard and acute forms) in the 1998 general U.S. population.

Scoring Checks

Because errors can lead to inaccurate scale scores, we strongly recommend formal scoring checks of SF-36 scales prior to computing the SF-36 component summary scales. These formal scoring checks are explained in full detail in the SF-36 Manual and Interpretation Guide (Ware et al., 1993).

The following scoring checks are also strongly recommended for the SF-36 component summary scales. Any discrepancies should be investigated for scoring errors:

Check correlation's between the eight SF-36 scales and the PCS and MCS scales. The PF, RP, and BP scales should correlate highest with the PCS and lowest with the MCS. The MH, RE, and SF scales should correlate highest with the MCS and lowest with the PCS. The GH and VT scales should correlate moderately with both physical and mental component scales.

Check the correlation between the physical and mental component summary scales. The correlation should be very low, generally lower than 0.30.

Scoring Exercise

The algorithms for scoring the SF-36 scales and summary measures for use with SAS software (along with a test data set) are available in electronic format to purchasers of this manual – by sending an email request to info@qmetric.com. QualityMetric also offers a Scoring & Data Quality Analysis service (information available at www.qmetric.com).

Table 6.4 presents descriptive statistics for the eight SF-36 scales (0-100 and norm-based) and the physical and mental component summary measures from the test dataset for the standard form. Table 6.5 presents descriptive statistics for the eight SF-36 scales (0-100 and norm-based) and the physical and mental component summary measures from the test dataset for the acute form.

After scoring the test dataset, you should observe the same means, standard deviations, and minimum and maximum observed values as those presented in Table 6.4 for the standard form and in Table 6.5 for the acute form.

TABLE 6.4 TEST DATASET DESCRIPTIVE STATISTICS*: SF-36 SCALES AND SUMMARY MEASURES (N = 100), STANDARD FORM

		0-100 SCORES				NORM-BASED SCORES			
	NO. OF CASES	MEAN	STANDARD DEVIATION	MIN. VALUE	MAX. VALUE	MEAN	STANDARD DEVIATION	MIN. VALUE	MAX. VALUE
Physical Functioning	100	82.55	23.74	5	100	49.82	9.96	17.29	57.14
Role Physical	100	78.25	35.82	0	100	50.09	10.13	27.95	56.24
Bodily Pain	100	70.34	23.11	0	100	50.05	9.90	19.93	62.75
General Health	100	65.32	22.72	0	100	47.76	10.63	17.18	63.99
Vitality	100	58.22	22.36	0	94	50.58	10.59	23.02	70.35
Social Functioning	100	83.00	22.79	0	100	49.75	9.90	13.71	57.14
Role Emotional	100	80.33	35.17	0	100	49.12	11.11	23.74	55.34
Mental Health	100	73.74	22.15	0	100	49.16	12.58	7.28	64.07
Physical Component Summary (PCS)	100	N/A	-	-	-	49.76	9.78	20.92	63.12
Mental Component Summary (MCS)	100	N/A	-	-	-	49.34	11.70	8.67	63.46

*All descriptive statistics are rounded to the second decimal point.

TABLE 6.5 TEST DATASET DESCRIPTIVE STATISTICS*: SF-36 SCALES AND SUMMARY MEASURES (N = 100), ACUTE FORM

		0-100 SCORES				NORM-BASED SCORES			
	NO. OF CASES	MEAN	STANDARD DEVIATION	MIN. VALUE	MAX. VALUE	MEAN	STANDARD DEVIATION	MIN. VALUE	MAX. VALUE
Physical Functioning	100	77.75	27.44	0	100	48.13	11.17	16.46	57.19
Role Physical	100	71.25	39.79	0	100	47.98	11.29	27.77	56.13
Bodily Pain	100	69.85	24.42	0	100	48.54	10.13	19.55	61.05
General Health	100	68.75	22.42	5	100	49.10	10.38	19.59	63.56
Vitality	100	59.98	22.51	0	100	51.07	10.32	23.58	69.41
Social Functioning	100	84.88	23.52	0	100	49.67	10.30	12.51	56.30
Role Emotional	100	83.67	31.96	0	100	50.38	9.82	24.66	55.40
Mental Health	100	78.35	15.85	8	100	51.23	9.19	10.45	63.77
Physical Component Summary (PCS)	100	47.49	10.87	19.00	63.13	47.49	10.87	19.00	63.13
Mental Component Summary (MCS)	100	51.76	8.61	13.43	62.06	51.76	8.61	13.44	62.07

*All descriptive statistics are rounded to the second decimal point.

SF-36 algorithms have been made available to computer software vendors and other organizations providing scoring and analysis services for the SF-36. Look for the symbol below.

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-36 Health Survey.



Chapter 7: Reliability And Statistical Power

Reliability estimates for the PCS and MCS scales were calculated using data from the Medical Outcomes Study (MOS), the 1990 and 1998 general U.S. populations, and data from general population surveys in Denmark, France, Italy, Netherlands, Norway, Spain, Sweden, and the U.K. As documented below, reliability estimates based on both internal consistency and test-retest methods have been high, ranging from 0.89 to 0.94 for the PCS and from 0.74 to 0.91 for the MCS studies of patients in the U.S. and in general population studies across countries. While in previous studies of the eight SF-36 scales (McHorney, Ware, Lu, et al., 1994), reliability estimates were slightly lower for more disadvantaged study participants, PCS and MCS reliability estimates were similar for disadvantaged and more advantaged groups.

Background

Indices of reliability give an indication of the extent to which the scores produced by a particular measurement procedure are consistent and reproducible. A measurement procedure is reliable to the extent that items within the same scale give the same results or to the extent that an individual scores the same across repeated administrations of the scale (Nunnally & Bernstein, 1994). A reliability coefficient is an estimate of how much of the variation in a score is real or truth, as opposed to chance or random errors. For example, a reliability coefficient of 0.80 indicates that 80% of the total measured variance is true score. Suggested levels of reliability are 0.70 or greater for scales used in group-level analyses and 0.90 or greater for scales used in decisions at the individual level (Nunnally & Bernstein, 1994).

There are several methods that can be used to estimate reliability. Reliability can be estimated by correlating scores from one administration with scores from another (test-retest reliability), or by correlating scores and testing the equivalence of individual answers across alternative forms of an instrument (alternative forms reliability), or by examining the equivalence of responses within the same test from a single administration (internal consistency reliability) (Nunnally & Bernstein, 1994). Most studies to date have used the internal consistency method and Cronbach's coefficient alpha to estimate the reliability of SF-36 scales, although all three methods listed above have been used in one or more studies (Ware, Snow, Kosinski, et al., 1993; McHorney & Ware, 1995).

Appendix C: Scoring Algorithms: 1990 Norm-Based Scoring of PCS and MCS

This appendix reproduces Chapter 4 from the first edition of this manual (Ware et al., 1994) to document SF-36 (v1) PCS and MCS scoring based on 1990 norms.

Scoring of the Physical (PCS) and Mental (MCS) Component Summary measures involves three steps. First, the eight SF-36 scales are standardized using means and standard deviations from the general U.S. population. Second, they are aggregated using weights (factor score coefficients) from the general U.S. population. Finally aggregate PCS and MCS scores are standardized using a linear T-score transformation to have a mean of 50 and a standard deviation of 10, in the general U.S. population.

General U.S. population statistics used in the standardization and in the aggregation of SF-36 scale scores are presented in Table C.1. Detailed information including formulas for scale aggregation and transformation of scores are presented below. Formal checks using a test dataset recorded on a computer diskette (see back cover) can be performed to confirm the successful reproduction of PCS and MCS scales, as discussed later in the chapter. We strongly recommend these tests.

TABLE C.1 GENERAL U.S. POPULATION MEANS, STANDARD DEVIATIONS AND FACTOR SCORE COEFFICIENTS USED TO DERIVE PCS AND MCS SCALE SCORES

			FACTOR SCORE COEFFICIENTS	
SF-36 SCALES	MEAN	SD	PCS	MCS
Physical Functioning (PF)	84.52404	22.89490	0.42402	-0.22999
Role Physical (RP)	81.19907	33.79729	0.35119	-0.12329
Bodily Pain (BP)	75.49196	23.55879	0.31754	-0.09731
General Health (GH)	72.21316	20.16964	0.24954	-0.01571
Vitality (VT)	61.05453	20.86942	0.02877	0.23534
Social Functioning (SF)	83.59753	22.37642	-0.00753	0.26876
Role Emotional (RE)	81.29467	33.02717	-0.19206	0.43407
Mental Health (MH)	74.84212	18.01189	-0.22069	0.48581

Importance Of Standardization

As with the scoring of SF-36 items and scales, which are aggregated to score the summary measures, standardization of the scoring of the PCS and MCS scales is vital to their interpretation. Any changes in scoring of the SF-36 items, scales, or the algorithms for the summary scales may compromise their reliability and validity. Changes in scoring have also been shown to invalidate normative comparisons, and changes are likely to complicate or prevent meaningful comparisons of results across studies. (Ware, Snow, Kosinski, et al., 1993)

Norm-Based Scoring

The PCS and MCS scales are scored using norm-based methods. The means, standard deviations, and factor score coefficients used in scoring come from the general U.S. population. A linear T-score transformation method is used so that both the PCS and MCS have a mean of 50 and a standard deviation of 10 in the general U.S. population. This transformation is in contrast to the 0-100 scoring used to date for the

eight SF-36 scales. The eight SF-36 scales have means ranging from 61 to 84 and standard deviations ranging from 18 to 34 in the general U.S. population (Ware, Snow, Kosinski, et al., 1993).

The advantage of the standardization and norm-based scoring of the PCS and MCS 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 and MCS measures, each one point difference in scores also has a direct interpretation. A one point difference is one-tenth of a standard deviation. Advantages of norm-based scoring are illustrated in Appendix D.

Steps In Scoring

Following the scoring of the eight scales according to the standard SF-36 scoring algorithms (Medical Outcomes Trust, 1991, 1994; Ware, Snow, Kosinski, et al., 1993), PCS and MCS are scored in three steps as explained below.

Standardization Of Scales

First, each SF-36 *scale* is standardized using a z-score transformation and SF-36 scale means and standard deviations from the general U.S. population as given in Table 1. A z-score for each scale is computed by subtracting the general U.S. population mean from each SF-36 scale score and dividing the difference by the corresponding scale standard deviation from the general U.S. population. Formulas are listed below.

Step 1. Formulas for z-score standardizations of SF-36 scales.

PF_Z = (PF – 84.52404) / 22.89490
RP_Z = (RP – 81.19907) / 33.79729
BP_Z = (BP – 75.49196) / 23.55879
GH_Z = (GH – 72.21316) / 20.16964
VT_Z = (VT – 61.05453) / 20.86942
SF_Z = (SF – 83.59753) / 22.37642
RE_Z = (RE – 81.29467) / 33.02717
MH_Z = (MH – 74.84212) / 18.01189

Aggregation Of Scale Scores

After a z-score has been computed for each SF-36 scale, the second step involves computation of aggregate scores for the physical and mental components using the physical and mental factor score coefficients from the general U.S. population as given in Table C.1.

Computation of an aggregate physical component score consists of multiplying each SF-36 scale z-score by its respective physical factor score coefficient and summing the eight products, as shown below. Similarly, an aggregate mental component score is obtained by multiplying each SF-36 scale z-score by its respective mental factor score coefficient and summing the eight products.

Step 2. Formulas for aggregating standardized scales in estimating aggregate physical and mental component scores:

AGG_PHYS = (PF_Z * .42402) + (RP_Z * .35119) + BP_Z * .31754+ (GH_Z * .24954) + (VT_Z * .02877) + (SF_Z * -.00753) + (RE_Z * -.19206) + (MH_Z * -.22069)

AGG_MENT = (PF_Z * .22999) + (RP_Z * .12329) + BP_Z * .09731+ (GH_Z * .01571) + (VT_Z * .23534) + (SF_Z * -.26876) + (RE_Z * -.43407) + (MH_Z * -.48581)

Pending results from ongoing evaluations of other options, it is recommended that component scale scores be set to missing if the respondent is missing any one of the eight SF-36 scales. To minimize the number of component scores missing, we recommend estimating each of the eight scale scores if half or more of the items are complete, as documented elsewhere (Medical Outcomes Trust, 1991, 1994; Ware, Snow, Kosinski, et al., 1993)

Transformation Of Summary Scores

The third step involves transforming each component score to the norm-based (50, 10) scoring. This is accomplished by multiplying each aggregate component scale score by 10 and adding the resulting product to 50. Formulas are listed below.

Step 3. Formulas for T-score transformation of component scores.

Transformed Physical (PCS) = 50 + (AGG_PHYS * 10)
Transformed Mental (MCS) = 50 + (AGG_MENT * 10)

Features Of PCS And MCS Scores

The PCS and MCS were constructed and scored to achieve a number of advantages, in addition to reducing the SF-36 from an eight-scale profile to two summary measures without substantial loss of information. Features of the PCS and MCS scores, including their reliability, confidence intervals (CI), skewness (percent ceiling and floor), and number of levels observed in the general U.S. population, are summarized in Table C.2. These results confirm some of the theoretical advantages of the two summary measures as compared to the eight SF-36 scales, including a very large increase in the number of levels defined, smaller confidence intervals relative to each of the eight scales, as well as the elimination of both floor and ceiling effects. Reliability estimates and confidence intervals are discussed further in Chapter 7. Tradeoffs between the two summary measures versus the eight-scale profile are evaluated in Chapter 8.

TABLE C.2 COMPARISON OF FEATURES OF SF-36 SCALES AND SUMMARY MEASURES, GENERAL U.S. POPULATION

		SUMMARY MEASURES	
	SF-36 SCALES[b]	PCS	MCS
Reliability	.78 - .93	.93	.88
95% CI (±)	13 – 33	5.7	6.3
% Floor	1 – 24	0	0
% Ceiling	1 – 56	0	0
# of Levels[a]	4 – 26	567	493

[a] Scores rounded to first decimal place

[b] Statistics are presented as the range of results found across the eight SF-36 scales in the general U.S. population.

Scoring Checks

Because errors can lead to inaccurate scale scores, we strongly recommend formal scoring checks of SF-36 scales prior to computing the SF-36 component summary scales. These formal scoring checks are explained in full detail in the SF-36 Scoring Exercise available through the Medical Outcomes Trust (Medical Outcomes Trust, 1994).

The following scoring checks are also strongly recommended for the SF-36 component summary scales. Any discrepancies should be investigated for scoring errors:

1. Check correlations between the eight SF-36 scales and the PCS and MCS scales. The PF, RP, and BP scales should correlate highest with the PCS and lowest with the MCS. The MH, RE, and SF scales should correlate highest with the MCS and lowest with the PCS. The GH and VT scales should correlate moderately with both physical and mental component scales.

2. Check the correlation between the physical and mental component summary scales. The correlation should be very low.

Tables and text in Chapter 3 summarize results from these tests across numerous studies.

Scoring Exercise

A test dataset and SAS code for scoring the SF-36 PCS and MCS scales has been provided on a computer diskette (see inside back cover). The SAS code begins with algorithms for scoring the eight SF-36 scales and finishes with the computation of the component summary scales. The purpose of this scoring exercise is to help SF-36 users evaluate results from each step in the process of calculating SF-36 component summary scale scores. The test dataset for this scoring exercise contains 100 administrations of the SF-36 Health Survey. The test dataset is called "RAWDATA." The SAS code for scoring the scales is called "SF36SUMM.SCR."

Table C.3 presents descriptive statistics for the eight SF-36 scales and the physical and mental component scales from the test dataset. After scoring the test data set, you should observe the same means, standard deviations, and minimum and maximum observed values as those presented in Table C.3.

Note that the missing data rate for each scale is artificially large as part of the scoring exercise. Consequently, the missing data rate for the two component summary scales is uncharacteristically high.

TABLE C.3 TEST DATASET DESCRIPTIVE STATISTICS SF-36 SCALES AND SUMMARY MEASURES (N = 100)

SF-36 SCALES	# OF CASES	MEAN	STANDARD DEVIATION	MINIMUM OBSERVED VALUE	MAXIMUM OBSERVED VALUE
Physical Functioning (PF)	99	75.8	25.1	5	100
Role Physical (RP)	93	58.0	40.5	0	100
Bodily Pain (BP)	98	69.5	25.1	0	100
General Health (GH)	96	59.6	22.9	5	100
Vitality (VT)	100	56.2	19.3	15	95
Social Functioning (SF)	100	83.5	24.7	12.5	100
Role Emotional (RE)	92	72.5	39.1	0	100
Mental Health (MH)	100	74.4	19.5	8	100
PCS	88	45.2	10.9	17.2	66.7
MCS	88	49.7	11.2	12.7	64.2