



User's Manual for the
**SF-36v2 Health Survey,
Third Edition**



USER'S MANUAL FOR THE SF-36v2 HEALTH SURVEY

THIRD EDITION

Edited by
Mark E. Maruish, PhD

Copyright © 1993 by John E. Ware, Jr., PhD
Copyright © 2011 by QualityMetric Incorporated.

SF-36, SF-36v2[®], SF-12[®], and SF-12v2[®] are trademarks of the Medical Outcomes Trust and are used under license. ACT[™], DYNHA[®], QualityMetric[™], QualityMetric Health Outcomes[™], SF-8[™], and SF-10[™] are trademarks of QualityMetric Incorporated.

The SF-36v2[®] Health Survey is copyrighted by QualityMetric Incorporated.

All rights reserved. No part of this manual covered by the copyrights herein 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.

ISBN: 1-891810-28-6

Permission to reproduce and to use the SF-36v2[®] Health Survey and the associated trademark(s) documented in this manual is routinely granted royalty-free to individuals and organizations that collect their own data for purposes of scholarly research. Permissions for both scholarly and commercial use of the SF-36v2[®] Health Survey can be obtained by completing a Survey Information Request Form. All other uses, commercial and noncommercial, may require payment of an annual license fee.

Completion of the Survey Information Request Form will result in the quotation of any user fees and, upon user request and approval by QualityMetric[™], the issuance of a license and invoice. Any organization or individual wishing to reproduce the survey documented herein and/or any associated intellectual property (e.g., the trademarks, scoring algorithms, interpretation guidelines, and/or normative data documented in this manual) for any purpose must register or obtain a license from QualityMetric. For information about registering or obtaining a license, visit <http://www.qualitymetric.com>.

Requests for permission to reproduce portions of this manual should be sent to QualityMetric Incorporated, 24 Albion Road, Building 400, Lincoln, RI 02865, or to info@qualitymetric.com.

Suggested citation:

Maruish, M. E. (Ed.). *User's manual for the SF-36v2 Health Survey* (3rd ed.). Lincoln, RI: QualityMetric Incorporated.

Contents

Copyright	ii
Table of Contents	iii
Figures and Tables	xi
Acknowledgments	xxiii
Preface	xxv

PART I: INTRODUCTION

Chapter 1 Introduction

Context for Health Status Assessment	3
Improvement of Health Status Surveys	4
The Evolution of Short Form Health Status Surveys.....	4
The Health Insurance Experiment (HIE).....	4
The Medical Outcomes Study (MOS).....	5
The International Quality of Life Assessment (IQOLA) Project	5
The Medicare Health Outcomes Study (HOS).....	5
1998 National Survey of Functional Health Status (NSFHS)	6
QualityMetric 2009 Norming Study.....	6
Improvements in Standards for Measurement Evaluation.....	7
New Standards for Health Status Measurement: The Short Form Health Surveys	7
SF-36 Health Survey	7
SF-12 Health Survey	8
SF-36v2 Health Survey	8
SF-12v2 Health Survey	9
SF-8 Health Survey	10
QualityMetric’s Item Banks and Computerized Adaptive Testing (CAT) Tool	10
A New Conceptual Framework for Health Status Assessment.....	12
Use of This Manual	14

Chapter 2 Concepts, Measures, and Applications

Concepts and Measures	15
Health Domain Scales	15
Physical and Mental Component Summary (PCS and MCS) Measures	16
Profile of Scores	19
SF-6D Health Utility Index	19
Applications	20
Evaluating and Monitoring Individual Patients in Clinical Practice	20

Monitoring Populations	21
Estimating the Burden of Disease	22
Evaluating Treatment Effects in Clinical Trials	22
Disease Management	23
Risk Prediction and Cost-Effectiveness	24
Patient-Provider Relations	24
Direct-to-Consumer Information	25
Survey Validation	26
A Final Comment on Applications	26

Chapter 3 The Short Form Family of Health Survey Instruments

The Short Form Instruments	29
The SF-36v2 Health Survey	29
The SF-12v2 Health Survey	30
The SF-8 Health Survey	30
Computerized Adaptive Testing (CAT) and the DYNHA Computer Adaptive Health Assessments	30
The SF-10 Health Survey for Children	31
Deciding Which Short Form Survey to Use	31
Features of the Short Form Surveys	31
Matching a Form to an Application: General Considerations	34
Matching a Form to an Application: Specific Form-to-Form Considerations	36

PART II: DATA COLLECTION AND SCORING

Chapter 4 Survey Administration

Determining Respondent Eligibility	41
Age	41
Reading Ability	41
Non-English-Speaking Respondents	42
Level of Respondent Cooperation and Understanding	42
Guidelines for Administration	42
When to Administer the Survey	43
Introducing the SF-36v2 to the Respondent	43
Addressing Problems and Questions	43
Concluding Survey Administration	45
Modes of Administration	45
Paper and Pencil	45
Interviewer Script	45
Online	46
Fax	46
Smartphone	46
Tablet or Kiosk	46
Considerations for the Use of Interview, Mail, or Online Format	46
Effects of Data Collection Method	48
Additional Considerations	50

Chapter 5 Scoring Procedures

Importance of Standardization	55
Scoring the SF-36v2	56
Step 1: Entering Data	57
Step 2: Recoding Item Response Values	58

Step 3: Determining Health Domain Scale Total Raw Scores	58
Step 4: Transforming Health Domain Scale Total Raw Scores to 0–100 Scores	58
Step 5: Transforming Health Domain Scale 0–100 Scores to <i>T</i> Scores	59
Step 6: Scoring the Physical and Mental Component Summary Measures	59
Step 7: Scoring the Response Consistency Index	60
Scoring Software and Services	60
Smart Measurement System	60
Health Outcomes Scoring Software 5.0	60

PART III: INTERPRETATION

Chapter 6 Data Quality Evaluation

Considerations for Analyzing Data From Groups of Respondents or Multiple Administrations to the Same Respondent	65
Combining and Analyzing Data From Standard and Acute Forms	65
Combining and Analyzing Data From Different Data Collection Methods	65
Combining and Analyzing Data From Different Translated Forms	66
Quantitative Evaluation of Data Quality	66
Completeness of Data	67
Responses Within Range	67
Consistent Responses	67
Percentage of Estimable Scale Scores	68
Item Internal Consistency	69
Item Discriminant Validity	70
Scale Reliability	70
Confirmation of the Two-Component Structure	70
Qualitative Evaluation of Data Quality	71
Results Inconsistent With Respondent Presentation	71
Unusually Quick or Long Completion Time	71
Patterned Responses	71
Data Quality Evaluation of Individual Health Domain Scales	71

Chapter 7 General Strategies for Interpreting the SF-36v2 Profile

General Considerations for Norm-Based Interpretation	73
Interpretation of the Component Summary Measures	75
Interpretation of the Health Domain Scales	77
Additional Considerations for Interpreting SF-36v2 Findings	78
The Standard Error of Measurement (<i>SEM</i>) and Confidence Intervals (<i>CI</i> s)	78
Supplemental Norms for Age, Gender, and Gender-by-Age Groups	79
Supplemental Benchmarks for Disease-Specific Populations	80
Use of Information From Other Instruments	80

Chapter 8 Content-Based Interpretation

Interpretation of Scales and Measures Across All Score Ranges	83
Content-Based Interpretation of the Standard Form Component Summary Measures	84
Physical Component Summary (PCS)	84
Mental Component Summary (MCS)	85
Content-Based Interpretation of the Standard Form Health Domain Scales	93
Physical Functioning (PF)	93
Role-Physical (RP)	93
Bodily Pain (BP)	96

General Health (GH)	96
Vitality (VT)	96
Social Functioning (SF)	96
Role-Emotional (RE)	96
Mental Health (MH)	97
Content-Based Interpretation of the Acute Form Component Summary Measures	97
Physical Component Summary (PCS)	97
Mental Component Summary (MCS)	107
Content-Based Interpretation of the Acute Form Health Domain Scales	108
Physical Functioning (PF)	109
Role-Physical (RP)	118
Bodily Pain (BP)	118
General Health (GH)	118
Vitality (VT)	118
Social Functioning (SF)	118
Role-Emotional (RE)	119
Mental Health (MH)	119
Interpolation of Score-Related Percentages	119

Chapter 9 Criterion-Based Interpretation

Interpretation of Scales and Measures Across All Score Ranges	131
Criterion-Based Interpretation of the Standard Form Component Summary Measures	132
Physical Component Summary (PCS)	132
Mental Component Summary (MCS)	133
Criterion-Based Interpretation of the Standard Form Health Domain Scales	137
Physical Functioning (PF)	137
Role-Physical (RP)	138
Bodily Pain (BP)	141
General Health (GH)	143
Vitality (VT)	143
Social Functioning (SF)	146
Role-Emotional (RE)	146
Mental Health (MH)	147
Criterion-Based Interpretation of the Acute Form Component Summary Measures	149
Physical Component Summary (PCS)	151
Mental Component Summary (MCS)	154
Criterion-Based Interpretation of the Acute Form Health Domain Scales	160
Physical Functioning (PF)	160
Role-Physical (RP)	161
Bodily Pain (BP)	161
General Health (GH)	162
Vitality (VT)	163
Social Functioning (SF)	164
Role-Emotional (RE)	164
Mental Health (MH)	164
Interpolation of Score-Related Percentages	166

Chapter 10 Determining Important Differences in Scores

General Considerations for Determining Minimally Important Differences (MIDs)	169
Criterion- or Anchor-Based Approaches to MID	170
Distribution-Based Approaches to MID	172

MID Criteria in Relation to Individual SF-36v2 Measures and Scales	173
Physical Component Summary (PCS)	173
Mental Component Summary (MCS)	173
Physical Functioning (PF)	173
Role-Physical (RP)	174
Bodily Pain (BP)	174
General Health (GH)	174
Vitality (VT)	174
Social Functioning (SF)	174
Role-Emotional (RE)	175
Mental Health (MH)	175
Responder Definition: Criteria for Minimally Important Difference in Individual Scores	175
Summary	177

Chapter 11 Interpretation of Group Data

Case 1	179
Case 2	181
Case 3	182

Chapter 12 Interpretation of Individual Respondent Data

Considerations for Interpreting Individual Respondent Data	185
Response Consistency	185
Item Analysis	186
Situational Considerations	186
Case Studies	186
Case 1	187
Case 2	189

PART IV: DEVELOPMENT AND PSYCHOMETRIC EVALUATION

Chapter 13 Development of the SF-36v2

Published Standards for Psychometric Measures	195
<i>Standards for Educational and Psychological Testing</i>	195
Medical Outcomes Trust Instrument Review Criteria	196
U.S. Food and Drug Administration Guidelines	196
Background	196
Conceptual Framework	197
Selection and Origin of Items	198
Health Domain Scales	198
Physical Functioning (PF)	198
Role-Physical (RP)	199
Bodily Pain (BP)	200
General Health (GH)	200
Vitality (VT)	201
Social Functioning (SF)	201
Role-Emotional (RE)	201
Mental Health (MH)	202
Self-Evaluated Transition (SET)	202
Differences Between the SF-36 and the SF-36v2	202
Layout	203
Type Size and Bolding	203
Wording Changes	203

Five-Choice Response Scales	203
Scoring	211
Advantages of <i>T</i> Scores	212
1998 Norms	214
Comparability of Results	215
Psychometric Characteristics and Comparability	215
Physical and Mental Component Summary Measures	218
Methodological Issues	220
Principal Components	220
Orthogonal Components	222
Scoring the Component Summary Measures: Use of Positive and Negative Component Weights	223
Development of the PCS and MCS Measures	224
Comparability of the SF-36 and SF-36v2 PCS and MCS Measures	224
The SF-6D	225
Advances Accompanying SF-36v2 Development	225
The SF-36v2 Profile	226
Calibration of Scales on a Common Metric	226
Missing Score Estimation	226
Translations	228
Chapter 14 2009 Normative Data	
How the SF-36v2 Was Renormed	231
Sampling	232
Data Collection Forms	232
Data Collection	234
Survey Readministration	236
Sample Characteristics	236
2009 U.S. General Population Norms	236
Development of the Health Domain Scale Scoring Algorithms	236
Development of the PCS and MCS Scoring Algorithms	236
Comparison of Item Format Presentations	237
Finalization of the 2009 Normative Samples	238
Component Summary Measure, Health Domain Scale, and SF-6D 2009 U.S. General Population Norms	239
2009 Supplemental Norms and Benchmarks	240
Supplemental Norms for Age, Gender, and Gender-by-Age Groups	240
Supplemental Benchmarks for Disease-Specific Populations	241
Comparability of 2009 and 1998 Normative Data: Preliminary Findings	243
Comparison of 2009 and 1998 Mean Item Raw Scores	243
Comparison of Mean Health Domain Scale <i>T</i> Scores	243
Comparison of Mean Health Domain <i>T</i> Scores in the 2009 U.S. General Population, Scored Using 2009 and 1998 Scoring Algorithms	245
Summary, Conclusions, and Recommendations	245
Chapter 15 Reliability	
Interpreting Reliability Coefficients	255
Internal Consistency Reliability	256
PCS and MCS Internal Consistency Estimates	256
Health Domain Scale Alpha Coefficients	257
Health Domain Scale Item-Scale Correlations	257
Test-Retest Reliability	260

Standard Error of Measurement	260
New Approaches to Evaluating the Reliability of Survey Instruments	261
Chapter 16 Validity	
Types of Validity: An Overview	263
Construct Validity	264
Factor Analyses	264
Convergent and Discriminant Validity	266
Known-Groups Comparisons	267
Criterion Validity	272
Concurrent Validity	272
Predictive Validity	277
Content Validity	277
Chapter 17 Statistical Power Analysis	
Statistical Power and <i>T</i> Scores.....	281
Experimental Studies.....	281
Nonexperimental Studies.....	282
Statistical Power and Scale Measurement Properties.....	283
Appendixes	
Appendix A Sample SF-36v2 Individual Respondent Reports	285
A.1 Sample SF-36v2 Member Report	285
A.2 Sample SF-36v2 Member Report with PIQ-6	286
Appendix B Sample SF-36v2 Group-Level Reports	287
B.1 Sample SF-36v2 SF Comparison for Total Sample Report	288
B.2 Sample SF-36v2 Scores by Age Group Report	289
B.3 Sample SF-36v2 Scores by Gender Report	290
B.4 Sample SF-36v2 Data Quality Evaluation Report	291
B.5 Sample SF-36v2 Summary Report of Scale and Summary Measure Scores.....	292
B.6 Sample SF-36v2 Missing Score Estimation Report.....	293
Appendix C SF-36v2 Score Estimation Using Item Response Theory (IRT)	295
References	299
Glossary	317
Index	325

Figures and Tables

Figures

Figure 1.1	Logic of Computerized Adaptive Testing	11
Figure 1.2	Patient-Reported Outcomes (PRO) Conceptual Framework	13
Figure 1.3	Components of Disease Impact Items	13
Figure 2.1	SF-36v2 Measurement Model	18
Figure 2.2	Sample SF-36v2 Profile of Scores	20
Figure 4.1	Recommended Steps for Administering SF-36v2	43
Figure 5.1	Process for Scoring SF-36v2 Health Domain Scales and Component Summary Measures	56
Figure 7.1	Sample SF-36v2 Profile of Scores	74
Figure 8.1	Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024)	86
Figure 8.2	Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024) (continued)	87
Figure 8.3	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024)	88
Figure 8.4	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024).....	89
Figure 8.5	Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024)	90
Figure 8.6	Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024)	91
Figure 8.7	Percentage of Adults Reporting Limitations in Social Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024)	92
Figure 8.8	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024)	94
Figure 8.9	Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 4,024)	95
Figure 8.10	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,034)	98

Figure 8.11	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,034) (continued)	99
Figure 8.12	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,027)	100
Figure 8.13	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,027)	101
Figure 8.14	Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,036)	102
Figure 8.15	Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,028)	103
Figure 8.16	Percentage of Adults Reporting Limitations in Social Activities at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,029)	104
Figure 8.17	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,026).....	105
Figure 8.18	Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,028)	106
Figure 8.19	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	109
Figure 8.20	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056) (continued)	110
Figure 8.21	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	111
Figure 8.22	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056).....	112
Figure 8.23	Percentage of Adults Reporting General Health Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	113
Figure 8.24	Percentage of Adults Reporting Limitations in Vitality at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	114
Figure 8.25	Percentage of Adults Reporting Limitations in Social Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056).....	115
Figure 8.26	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	116
Figure 8.27	Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	117
Figure 8.28	Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,059)	120

Figure 8.29	Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,059) (continued)	121
Figure 8.30	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	122
Figure 8.31	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	123
Figure 8.32	Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,061)	124
Figure 8.33	Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	125
Figure 8.34	Percentage of Adults Reporting Limitations in Social Activities at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	126
Figure 8.35	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	127
Figure 8.36	Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,060)	128
Figure 10.1	Model for the Analysis of the Meaningfulness of Differences in SF-36v2 Scores	170
Figure 11.1	Unstandardized Changes in SF-36v2 Health Domain Scale and Component Summary Measure Scores Between Anemic HCV Patients Randomized to Epoetin Alfa and Placebo	180
Figure 11.2	Standardized Changes in SF-36v2 Health Domain Scale and Component Summary Measure Scores Between Anemic HCV Patients Randomized to Epoetin Alfa and Placebo	180
Figure 11.3	Mean SF-36v2 Health Domain Scale and Component Summary Measure Scores Before and After Anterior Cervical Fusion (<i>N</i> = 20)	181
Figure 11.4	Effects of Tai Chi Exercise on the Physical and Mental Health of College Students (<i>N</i> = 31).....	182
Figure 12.1	SF-36v2 Profile of Scores for Case 1	188
Figure 12.2	SF-36v2 Profile of Scores for Case 2	190
Figure 13.1	SF-36 Profile of 0–100 Scores: Adults With Asthma Compared With U.S. General Population Norms	213
Figure 13.2	SF-36 Profile of <i>T</i> Scores: Adults With Asthma Compared With U.S. General Population Norms ...	213
Figure 13.3	Interpreting SF-36 Treatment Outcomes Among Adults With Asthma	214
Figure 13.4	Plot of SF-36v2 Scale Factor Loadings on Orthogonal Physical and Mental Components for the 1998 U.S. General Population (<i>N</i> = 7,069)	223
Figure 13.5	Scoring of SF-8, SF-12v2, SF-36v2, and Dynamic Health Assessments on a Common Metric	227
Figure 14.1	Sample SF-36v2 Single-Item Screen Presentation Used for the QualityMetric 2009 Norming Study Project	234
Figure 14.2	Sample SF-36v2 Item-Grid Screen Presentation Used for the QualityMetric 2009 Norming Study Project	235
Figure C.1	Item Characteristic Curves for Physical Functioning Item 3d	295
Figure C.2	Item Characteristic Curves for Physical Functioning Items 3d, 3f, and 3i	297
Figure C.2	Relation Between Physical Functioning IRT and Sum Scores	298

Tables

Table 2.1	Abbreviated Item Content for SF-36v2 Health Domain Scales	17
Table 2.2	Comparison of Features of SF-36v2 Health Survey Health Domain Scales and Component Summary Measures Based on 2009 U.S. General Population Data	19

Table 3.1	Comparison of the Number of Items and Levels of Measurement for Each Component Summary Measure and Health Domain Scale for SF-8, SF-12v2, and SF-36v2	33
Table 3.2	Short Form Data Quality Indicators, by Survey	33
Table 3.3	Summary of Fixed-Form Short Form Health Survey Similarities and Differences	36
Table 4.1	SF-36v2 Administration Dos and Don'ts	46
Table 4.2	Effects of Method of Data Collection on Short Form Results: Findings From Selected Studies	51
Table 5.1	Bodily Pain Item 8 Response Choices and Scoring Information	58
Table 6.1	SF-36v2 Quantitative Data Quality Indicators	66
Table 6.2	SF-36v2 Standard (4-Week Recall) Form Response Consistency Index (RCI) Frequencies, 2009 U.S. General Population ($N = 4,024$)	68
Table 6.3	SF-36v2 Acute (1-Week Recall) Form Response Consistency Index Frequencies, 2009 U.S. General Population ($N = 2,056$)	68
Table 6.4	Number of Completed Items for Each SF-36v2 Health Domain Scale Required for Each Score Estimation Method	69
Table 6.5	SF-36v2 Health Domain and Component Summary Measure Sample Data Set	69
Table 7.1	Composition and Interpretation of Lowest and Highest T Scores for SF-36v2 Component Summary Measures and Health Domain Scales	76
Table 7.2	Values for Constructing Confidence Intervals Around Individual Respondent SF-36v2 Standard (4-Week Recall) Form T Scores Based on the 2009 U.S. General Population Data ($N = 4,024-4,036$)	79
Table 7.3	Values for Constructing Confidence Intervals Around Individual Respondent SF-36v2 Acute (1-Week Recall) Form T Scores Based on the 2009 U.S. General Population Data ($N =$ $2,056-2,061$)	79
Table 8.1	Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$)	86
Table 8.2	Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$) (continued).....	87
Table 8.3	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$)	88
Table 8.4	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$).....	89
Table 8.5	Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$)	90
Table 8.6	Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$)	91
Table 8.7	Percentage of Adults Reporting Limitations in Social Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$)	92
Table 8.8	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$)	94
Table 8.9	Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population ($N = 4,024$)	95
Table 8.10	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population ($N = 4,034$)	98

Table 8.11	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,034) (continued)	99
Table 8.12	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,027)	100
Table 8.13	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,027)	101
Table 8.14	Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,036)	102
Table 8.15	Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,028).....	103
Table 8.16	Percentage of Adults Reporting Limitations in Social Activities at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,029)	104
Table 8.17	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,026).....	105
Table 8.18	Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 4,028)	106
Table 8.19	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	109
Table 8.20	Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056) (continued)	110
Table 8.21	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	111
Table 8.22	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056).....	112
Table 8.23	Percentage of Adults Reporting General Health Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	113
Table 8.24	Percentage of Adults Reporting Limitations in Vitality at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	114
Table 8.25	Percentage of Adults Reporting Limitations in Social Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056).....	115
Table 8.26	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	116
Table 8.27	Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	117
Table 8.28	Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,059)	120

Table 8.29	Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,059) (continued)	121
Table 8.30	Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	122
Table 8.31	Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,056)	123
Table 8.32	Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,061)	124
Table 8.33	Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	125
Table 8.34	Percentage of Adults Reporting Limitations in Social Activities at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	125
Table 8.35	Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,057)	126
Table 8.36	Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (<i>N</i> = 2,060)	128
Table 8.37	Percentage of Adults Reporting Feeling Energetic Little or None of the Time at 9 Levels of SF-36v2 Mental Component Summary Measure Scores, 2009 U.S. General Population.....	129
Table 9.1	Percentage of Adults Reporting General Health and Quality of Life Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	134
Table 9.2	Percentage of Adults Reporting Problems in Work Performance and Other Activities at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population.....	135
Table 9.3	Percentage of Adults Reporting Health Problems and Treatment at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	136
Table 9.4	Percentage of Adults Reporting Pain Interference Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	136
Table 9.5	Percentage of Adults Reporting Future Health and Work-Related Problems at 7 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	136
Table 9.6	Percentage of Adults Reporting Behavioral Health Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	138
Table 9.7	Percentage of Adults Reporting Negative Effects of Personal, Emotional, and Physical Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	138
Table 9.8	Percentage of Adults Reporting Job Performance Problems and the Effects of Stress at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population.....	139
Table 9.9	Percentage of Adults Reporting Problems in Health, Quality of Life, Pain-Related Interference, and Energy Level at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	140

Table 9.10	Percentage of Adults Reporting Problems in Cognitive Functioning at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	141
Table 9.11	Percentage of Adults Reporting Sleep Disturbance Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	141
Table 9.12	Percentage of Adults Reporting Sleep Somnolence Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	142
Table 9.13	Percentage of Adults Reporting Problems With Sleep Quantity and Adequacy and Headaches or Shortness of Breath at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	142
Table 9.14	Percentage of Adults Reporting Future Mental Health Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	143
Table 9.15	Percentage of Adults Reporting Problems Related to Quality of Life and the Performance of Work and Other Activities at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population.....	143
Table 9.16	Percentage of Adults Reporting Work Performance Problems, Significant Illness or Injury, and Limitations Due to Pain and Physical Conditions at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population	144
Table 9.17	Percentage of Adults Reporting Problems With Bodily Pain and Its Effects at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population	144
Table 9.18	Percentage of Adults Reporting Significant Chronic Conditions, Treatment, and Disability at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population	144
Table 9.19	Percentage of Adults Reporting Quality of Life and General Health Problems and Disability at 11 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population	146
Table 9.20	Percentage of Adults Reporting Significant Chronic Conditions and Treatment at 11 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population	146
Table 9.21	Percentage of Adults Reporting Quality of Life and Level of Energy Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population....	147
Table 9.22	Percentage of Adults Reporting Quality of Life Problems and Limitations in Social Activities at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population	147
Table 9.23	Percentage of Adults Reporting Emotional Problems and Problems Related to Quality of Life, Happiness, and Stress at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population	148
Table 9.24	Percentage of Adults Reporting Cognitive Functioning and Health Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population	148
Table 9.25	Percentage of Adults Reporting Problems With Depression and Anxiety at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population	149
Table 9.26	Percentage of Adults Reporting Quality of Life, Happiness, Emotional Problems, and Stress at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population	150
Table 9.27	Percentage of Adults Reporting Problems Related to Pain and Treatment at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population	151

Table 9.28	Percentage of Adults Reporting Cognitive Functioning Problems at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population....	151
Table 9.29	Percentage of Adults Reporting General Health and Quality of Life Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	152
Table 9.30	Percentage of Adults Reporting Problems in Work Performance and Other Activities at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population.....	154
Table 9.31	Percentage of Adults Reporting Health Problems and Treatment at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	154
Table 9.32	Percentage of Adults Reporting Sleep Disturbance Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	155
Table 9.33	Percentage of Adults Reporting Sleep Somnolence Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	155
Table 9.34	Percentage of Adults Reporting Problems With Sleep Quantity and Adequacy, Snoring, and Headaches or Shortness of Breath at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	156
Table 9.35	Percentage of Adults Reporting Future Health and Work-Related Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population	156
Table 9.36	Percentage of Adults Reporting Depression and Anxiety at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	157
Table 9.37	Percentage of Adults Reporting Negative Effects of Personal, Emotional, and Physical Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	157
Table 9.38	Percentage of Adults Reporting Job Performance Problems and the Effects of Stress at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population.....	158
Table 9.39	Percentage of Adults Reporting Problems in Health, Quality of Life, and Energy Level at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population.....	159
Table 9.40	Percentage of Adults Reporting Sleep Disturbance Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	159
Table 9.41	Percentage of Adults Reporting Sleep Somnolence Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	160
Table 9.42	Percentage of Adults Reporting Problems With Sleep Quantity and Adequacy and Headaches or Shortness of Breath at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population.....	160
Table 9.43	Percentage of Adults Reporting Future Mental Health and Work-Related Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population	161
Table 9.44	Percentage of Adults Reporting Problems Related to Quality of Life and the Performance of Work and Other Activities at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population.....	161
Table 9.45	Percentage of Adults Reporting Significant Illness or Injury and Limitations Due to Physical Conditions at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population.....	162

Table 9.46	Percentage of Adults Reporting Significant Chronic Conditions, Treatment, and Disability at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population	162
Table 9.47	Percentage of Adults Reporting Quality of Life and General Health Problems and Disability at 11 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population	163
Table 9.48	Percentage of Adults Reporting Significant Chronic Conditions, Missed Workdays, and Treatment at 10 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population	163
Table 9.49	Percentage of Adults Reporting Quality of Life and Level of Energy Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population	164
Table 9.50	Percentage of Adults Reporting Quality of Life Problems and Limitations in Social Activities at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population.....	164
Table 9.51	Percentage of Adults Reporting Emotional Problems and Problems Related to Quality of Life, Happiness, and Stress at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population	165
Table 9.52	Percentage of Adults Reporting Poor Job Performance and Health Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population	165
Table 9.53	Percentage of Adults Reporting Problems With Depression and Anxiety at 11 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population.....	166
Table 9.54	Percentage of Adults Reporting Quality of Life, Happiness, and Stress at 11 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population	166
Table 10.1	Physical Component Summary Measure Score Differences as Predictors of Mortality at 2-Year Follow-Up, Medicare Health Outcomes Survey (<i>N</i> = 519,035)	171
Table 10.2	Criteria for Significant Change Scores According to Different Proposals	176
Table 12.1	Values for Determining Health Domain Scale and Component Summary Measure Confidence Intervals and Minimally Important Differences for Case 1 and Case 2	187
Table 13.1	Summary of Health Phenomena Captured by SF-36 and SF-36v2 Health Domain Scales.....	198
Table 13.2	Conceptual Origins of Short Form Survey Content.....	199
Table 13.3	Mean Current Health Scores for Respondents Choosing Each Level of SF-36v2 Item 1	201
Table 13.4	Comparison of Items in SF-36v2, SF-36, SF-36 Developmental Version, and Original MOS PAQ	204
Table 13.5	Summary of Item Wording Changes From SF-36 to SF-36v2	211
Table 13.6	Comparison of Number of Health Domain Items and Scale Levels for SF-36v2 and SF-36	212
Table 13.7	SF-36v2 Standard (4-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (<i>N</i> = 5,038)	216
Table 13.8	SF-36 Standard (4-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (<i>N</i> = 2,031)	217
Table 13.9	SF-36v2 Acute (1-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (<i>N</i> = 6,137).....	218
Table 13.10	SF-36 Acute (1-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (<i>N</i> = 1,700).....	219
Table 13.11	Comparison of Descriptive Statistics and Reliability Estimates for the Standard (4-Week Recall) Forms of SF-36v2 and SF-36, 1998 U.S. General Population Sample Using 0–100 Scoring	221
Table 13.12	Comparison of Descriptive Statistics and Reliability Estimates for the Acute (1-Week Recall) Forms of SF-36v2 and SF-36, 1998 U.S. General Population Using 0–100 Scoring	221
Table 13.13	Product-Moment Correlations and Reliability Coefficients for SF-36v2 Standard (4-Week Recall) Form Scales in the 1998 General U.S. Population (<i>N</i> = 7,069)	222

Table 13.14	Product-Moment Correlations and Reliability Coefficients for SF-36v2 Acute (1-Week Recall) Form Scales in the 1998 U.S. General Population ($N = 7,837$)	222
Table 13.15	Correlations Between Health Domain Scales and Rotated Physical and Mental Health Components Across SF-36v2 and SF-36 Standard and Acute Forms, 1998 U.S. General Population	225
Table 14.1	Composition of QualityMetric 2009 Norming Study Forms Used to Collect SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Normative and Validation Data	233
Table 14.2	Survey Wave 1 Sample Sizes and Completion Rates, by Server	235
Table 14.3	Sample Group Summary, by Study Wave	236
Table 14.4	Overall Sample Sizes and Completion Rates, by Study Wave	236
Table 14.5	Demographic Characteristics of SF-36v2 Standard (4-Week Recall) Form, 2009 U.S. General Population ($N = 4,040$)	237
Table 14.6	Demographic Characteristics of SF-36v2 Acute (1-Week Recall) Form, 2009 U.S. General Population ($N = 2,061$)	238
Table 14.7	Comparison of SF-36v2 Standard (4-Week Recall) Form Mean Health Domain Scale T Scores Based on Single-Item and Item-Grid Presentation Formats, 2009 U.S. General Population	239
Table 14.8	SF-36v2 Standard (4-Week Recall) Form Normative Data, 2009 U.S. General Population	240
Table 14.9	SF-36v2 Acute (1-Week Recall) Form Normative Data, 2009 U.S. General Population	240
Table 14.10	SF-36v2 Standard (4-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 2009 U.S. General Population ($N = 4,040$)	241
Table 14.11	SF-36v2 Acute (1-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 2009 U.S. General Population ($N = 2,061$)	242
Table 14.12	Percentage Scoring at the Floor and Ceiling of Each SF-36v2 Standard (4-Week Recall) Form Health Domain Scale by Self-Reported Disease Group, 2009 U.S. General Population	244
Table 14.13	Percentage Scoring at the Floor and Ceiling of Each SF-36v2 Acute (1-Week Recall) Form Health Domain Scale by Self-Reported Disease Group, 2009 U.S. General Population	246
Table 14.14	Characteristics of 2009 SF-36v2 Standard (4-Week Recall) Form Disease-Specific Benchmark Samples	248
Table 14.15	Characteristics of 2009 SF-36v2 Acute (1-Week Recall) Form Disease-Specific Benchmark Samples	250
Table 14.16	Mean Item Raw Scores for SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 and 1998 U.S. General Populations	252
Table 14.17	Differences in SF-36v2 Standard (4-Week Recall) Form Mean Health Domain Scale T Scores, 2009 and 1998 U.S. General Populations	253
Table 14.18	Differences in SF-36v2 Acute (1-Week Recall) Form Mean Health Domain Scale T Scores, 2009 and 1998 U.S. General Populations	253
Table 14.19	Comparison of SF-36v2 Standard (4-Week Recall) Form Mean Health Domain Scale T Scores Using 2009 and 1998 Scoring Algorithms, 2009 U.S. General Population ($N = 4,040$)	253
Table 14.20	Comparison of SF-36v2 Acute (1-Week Recall) Form Mean Health Domain Scale T Scores Using 2009 and 1998 Scoring Algorithms, 2009 U.S. General Population ($N = 2,061$)	253
Table 15.1	Internal Consistency Reliability Estimates for SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 U.S. General Population	257
Table 15.2	Internal Consistency Reliability Estimates for SF-36v2 Standard (4-Week Recall) Form Component Summary Measures and Health Domain Scales, by Respondent Subgroup in the 2009 U.S. General Population	258
Table 15.3	Internal Consistency Reliability Estimates for SF-36v2 Acute (1-Week Recall) Form Component Summary Measures and Health Domain Scales, by Respondent Subgroup in the 2009 U.S. General Population	259
Table 15.4	Test-Retest Reliability Estimates for SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 U.S. General Population	260
Table 15.5	Standard Errors of Measurement ($SEMs$) for SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 U.S. General Population	260

Table 16.1	Scale Validity and Correlations With Rotated Principal Components for SF-36v2 Standard (4-Week Recall) Form, 2009 ($N = 4,016$) and 1998 ($N = 6,742$) U.S. General Populations	265
Table 16.2	Scale Validity and Correlations With Rotated Principal Components for SF-36v2 Acute (1-Week Recall) Form, 2009 ($N = 1,876$) and 1998 ($N = 7,683$) U.S. General Populations	265
Table 16.3	Correlations Between SF-36v2 Standard (4-Week Recall) Form Component Summary Measures and Health Domain Scales, 2009 U.S. General Population ($N = 4,021$ – $4,036$)	267
Table 16.4	Correlations Between SF-36v2 Acute (1-Week Recall) Form Component Summary Measures and Health Domain Scales, 2009 U.S. General Population ($N = 2,056$ – $2,061$)	267
Table 16.5	Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical Condition Groups, 2009 U.S. General Population	270
Table 16.6	Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental and Physical Condition Groups, 2009 U.S. General Population	270
Table 16.7	Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental Condition Groups, 2009 U.S. General Population	271
Table 16.8	Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical and Mental Condition Groups, 2009 U.S. General Population	271
Table 16.9	Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical Condition Groups, 2009 U.S. General Population	273
Table 16.10	Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental and Physical Condition Groups, 2009 U.S. General Population	273
Table 16.11	Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental Condition Groups, 2009 U.S. General Population	274
Table 16.12	Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical and Mental Condition Groups, 2009 U.S. General Population	274
Table 16.13	Correlations of SF-36v2 Standard (4-Week Recall) Form Variables With Other Survey, Validation, Health Care, and Background Variables, 2009 U.S. General Population	275
Table 16.14	Correlations of SF-36v2 Acute (1-Week Recall) Form Variables With Other Survey, Validation, Health Care, and Background Variables, 2009 U.S. General Population	276
Table 16.15	Percentage of Respondents Reporting Subsequent (3–4 Months) Adverse Events by Baseline SF-36v2 Standard (4-Week Recall) Form Component Summary Measure T Scores	278
Table 16.16	Percentage of Respondents Reporting Subsequent (3–4 Months) Adverse Events by Baseline SF-36v2 Acute (1-Week Recall) Form Component Summary Measure T Scores	278
Table 16.17	Summary of Content of Widely Used General Health Surveys.....	279
Table 17.1	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD -Unit Differences Between Postintervention Scores of Two Experimental Groups With Preintervention Scores as Covariates (Change Score ANCOVA, Retest Correlation = .60).....	282
Table 17.2	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD -Unit Differences Between Postintervention Scores of Two Experimental Groups With Preintervention Scores as Covariates (Change Score ANCOVA, Retest Correlation = .40)	282
Table 17.3	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD -Unit Differences Between Two Experimental Groups, Postintervention Scores Only (ANOVA, Retest Correlation = .60)	282
Table 17.4	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD -Unit Differences Between Two Self-Selected Groups, Repeated Measures Design (Retest Correlation = .60)	283

Table 17.5	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form <i>SD</i> -Unit Differences Between Two Self-Selected Groups, Repeated Measures Design (Retest Correlation = .40)	283
Table 17.6	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form <i>SD</i> -Unit Differences Over Time Within One Group (Retest Correlation = .60)	283
Table 17.7	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form <i>SD</i> -Unit Differences Over Time Within One Group (Retest Correlation = .40).....	284
Table 17.8	Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form <i>SD</i> -Unit Differences Between a Group Mean and a Fixed Norm	284
Table C.1	Partial Credit Model Item Response Theory Parameters for SF-36v2 Physical Functioning Items	296

Acknowledgments

The development of this manual was supported by QualityMetric Incorporated from its own research funds. Several earlier publications documenting the SF-36® Health Survey and SF-36v2® Health Survey were very useful in preparing this manual, including: *User's Manual for the SF-36v2 Health Survey, Second Edition* (Ware et al., 2007); *SF-36 Health Survey Manual and Interpretation Guide* (Ware, Snow, Kosinski, & Gandek, 1993); *SF-36 Physical and Mental Health Summary Scales: A User's Manual* (Ware, Kosinski, & Keller, 1994); *How to Score Version 2 of the SF-36 Health Survey* (Ware, Kosinski, & Dewey, 2000); and *SF-36 Physical and Mental Health Summary Scales: A Manual for Users of Version 1* (Ware & Kosinski, 2001b). Other useful documents included: *How to Score the Revised MOS Short-Form Health Scales* (Ware, 1988); *Scoring Exercise for the SF-36 Health Survey* (Medical Outcomes Trust, 1994a); *How to Score the SF-12 Physical and Mental Health Summary Scales* (Ware, Kosinski, & Keller, 1995); *How to Score and Interpret Single-Item Health Status Measures: A Manual for Users of the SF-8 Health Survey* (Ware, Kosinski, Dewey, & Gandek, 2001); and *User's Manual for the SF-12v2 Health Survey, Second Edition* (Ware et al., 2010). Also noteworthy is the contribution of Cathy Sherbourne, PhD, who assisted in the preparation of the first in a series of peer-reviewed articles, summarizing the conceptual framework underlying the selection of items for the SF-36 (Ware & Sherbourne, 1992). Other articles in this series on the SF-36, published in *Medical Care*, focused on psychometric and clinical tests of validity (McHorney, Ware, & Raczek, 1993) and tests of data quality, scaling assumptions, and reliability across diverse groups of respondents (McHorney, Ware, Lu, & Sherbourne, 1994).

The initial development and validation of the SF-36 were supported by grant #89-6515, awarded in 1989 by the Henry J. Kaiser Family Foundation to The Health Institute at Tufts-New England Medical Center (now

known as Tufts Medical Center; John E. Ware, Jr., PhD, Principal Investigator). In 1990, the Hartford Foundation awarded additional funding to Dr. Ware at The Health Institute, which supported the dissemination of the SF-36 and education regarding its use. In 1994, the Henry J. Kaiser Family Foundation provided funding in support of the Medical Outcomes Trust, a nonprofit organization established in 1992 to retain the SF-36. Development of the SF-36 Physical and Mental Component Summary measures (PCS and MCS, respectively) was supported from 1991 through 1995 by unrestricted research grants for the International Quality of Life Assessment (IQOLA) Project that were awarded to The Health Institute at Tufts-New England Medical Center by Glaxo Research Institute (now GlaxoSmithKline) and Schering-Plough Corporation (now part of Merck & Co.), as well as by The Health Institute at Tufts-New England Medical Center from its own research funds. The revised version of the SF-36, the SF-36v2, was developed by QualityMetric Incorporated from its own research funds and with assistance from the Health Assessment Lab.

The first national norming of the SF-36 in the U.S. general population (1990) and its first administration in the Medical Outcomes Study (MOS) during the 4-year follow-up survey were both sponsored by the Functional Outcomes Program of the Henry J. Kaiser Family Foundation and The Health Institute at Tufts-New England Medical Center. The second national norming of the SF-36 and SF-36v2 in the U.S. general population (1998) was supported, in part, by the Musculoskeletal Education and Research Institute of the American Academy of Orthopedic Surgeons and by QualityMetric Incorporated from its own research funds. The 2009 norming of the SF-36v2 was supported by QualityMetric Incorporated from its own research funds.

In 1981, Alvin R. Tarlov, MD, then at the University of Chicago and later at Tufts-New England Medical

Center; Edward B. Perrin, PhD, at the University of Washington; Michael Zubkoff, PhD, and Eugene C. Nelson, ScD, at Dartmouth Medical School; and John E. Ware, Jr., PhD, then at the RAND Corporation and later at Tufts-New England Medical Center, began planning the Medical Outcomes Study. We gratefully acknowledge their collaboration and their faithful service on the MOS Steering Committee and to the Medical Outcomes Trust. MOS planning and data collection were sponsored by The Robert Wood Johnson Foundation, the Henry J. Kaiser Family Foundation, and the Pew Charitable Trusts. Additional support for MOS data analysis was given by the Agency for Health Care Policy and Research (now known as the Agency for Healthcare Research and Quality), the National Institute on Aging, and the National Institute of Mental Health. The MOS Functioning and Well-Being Profile, which was the source of the items selected or adapted for the SF-36, was developed and evaluated by a multi-institutional MOS team. Their work and relevant prior work are documented in *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach* (Stewart & Ware, 1992). MOS questionnaires were fielded and processed by the Survey Research Group at the RAND Corporation.

The guidelines for administering the SF-36v2 were adapted from those developed by the Measuring Health Concepts Project Team at Southern Illinois University, under the direction of W. Russell Wright, PhD, and others at the School of Medicine (Ware, Snyder, McClure, & Jarrett, 1972).

We gratefully acknowledge the contributions of a number of other individuals, including: Allyson Ross Davies, PhD, Elizabeth Kantz, MSc, and others then at Tufts-New England Medical Center's Quality Assessment Department for their collaboration in the very first studies of improved SF-36 response options; James E. Dewey, PhD, and Barbara Gandek, MS, for their contributions to the first edition of this manual; Dwana Bush, MD, of Atlanta, Georgia, for her suggestions regarding clinical

practice applications of short-form patient surveys and for her contributions to the development of the patient case studies presented in Chapter 12 of this manual; John E. Brazier, PhD, University of Sheffield, for his comments regarding the SF-6D portions of this manual; and David DeBrotta and Eli Lilly and Company for their support of the randomized controlled study on the effects of administration mode on SF-36 results. We also gratefully acknowledge our IQOLA Project collaborators who helped us discover and correct the ambiguities and cultural biases inherent in some of the SF-36 items. We also acknowledge the Research Triangle Institute in Research Triangle Park, North Carolina, whose research into alternate survey layouts contributed to the selection of the new SF-36v2 format. Special recognition goes to Barbara Gandek, MS, who oversaw the QualityMetric 2009 Norming Study and to Kevin Smith, PhD; Jakob B. Bjorner, MD, PhD; and Mark Kosinski, MA, for their significant contributions to this manual, including their data analysis and writing contributions. We also thank other QualityMetric Incorporated staff members who contributed to the production of this manual, including Jen Hanlon, MPH; Rohit Goyal, MS; Lori Jovin; Kate Miller, PhD; Aaron Yaras, PhD; Regina Rendas-Baum, MS; Sheila Hetu; Kristin Wystepek; Megan Clark; Michelle White, PhD; and Martha Bayliss, MSc.

Furthermore, we acknowledge and thank the thousands of researchers worldwide who have published results from their use of the SF-36 and SF-36v2 in population studies, clinical trials, and clinical practice. Their applications and evaluations have provided constructive and useful feedback and have greatly expanded the knowledge base that enables the meaningful interpretation of both surveys. The contributions of these researchers are cited throughout this manual and at <http://www.sf-36.org>.

Finally, we acknowledge the more than 20 years of contributions made by John E. Ware, Jr., PhD, to the development of the SF-36, SF-36v2, and the other members of the Short Form "family of instruments."

Preface

This third edition of the *User's Manual for the SF-36v2 Health Survey (User's Manual)* is an update to the second edition of the manual (Ware et al., 2007) and serves three main purposes. First, it chronicles the history and development of the SF-36® Health Survey, the first of the Short Form instruments. Second, it documents the survey improvements that led to the development of the SF-36v2®. These include improvements in item wording, instructions, and response categories, as well as improvements in the format recommended by the developers for both the standard (4-week recall) and acute (1-week recall) SF-36v2 forms. Third, this manual documents the QualityMetric 2009 Norming Study project, which led to the development of the most current general population, age, and gender norms and disease-specific benchmark data available for the SF-36v2.

Along with documenting the norming project, this manual also presents the results of several analyses that employ the 2009 data and provide further evidence of the SF-36v2's reliability and validity. Moreover, the results of several analyses using both the 2009 and the 1998 normative data are presented here. These studies were conducted to determine the degree to which findings derived from the 2009 norms are comparable to those obtained from the 1998 norms.

As in the second edition, this third edition of the *User's Manual* provides extensive information about how to interpret SF-36v2 results for both individual respondents and groups of respondents. Also provided are detailed guidelines for evaluating the quality of individual respondent and group-level data, general strategies for interpreting SF-36v2 results, and data for conducting content- and criterion-based interpretations of those results based on the findings from the 2009 norming study. Case studies demonstrating the application of the recommended interpretive guidelines for group and individual respondent data are also included

in this edition of the SF-36v2 manual. Moreover, easy-to-use look-up tables are provided for determining the minimum sample sizes required to detect various levels of difference in SF-36v2 health domain scale and component summary measure scores. These tables can assist researchers in developing methodologically sound designs for research involving the use of the SF-36v2. Finally, features of each member of the adult Short Form family of instruments— the SF-36v2, SF-12v2®, SF-8™, and DYNHA® SF-36 Health Surveys are compared and contrasted to assist Short Form users in determining which of these instruments will best meet their clinical and/or research needs.

As with the second edition of the *User's Manual*, the content is organized and presented in a manner that facilitates its use for both research and clinical purposes. The information that is most useful for those who want to quickly begin using the survey is presented at the beginning of the manual. This includes information that will help users properly select, administer, score, and interpret the SF-36v2 forms. Furthermore, information regarding the survey's development, norms, and psychometric properties is presented in the second half of the manual. However, regardless of their interests and intended uses of the instrument, all users of the SF-36v2 should familiarize themselves with all the information presented in this manual.

It is important for the reader to note that since the publication of the second edition of this manual in 2007, QualityMetric introduced three changes in terminology that it had been using in its commercial and peer-reviewed publications for more than a decade. First, what were previously called “norm-based scores” are now referred to as “*T* scores.” Also, the set of procedures used to maximize the amount of useable Short Form data, previously referred to as “Missing Data Estimation (MDE),” is now called “Missing Score Estimation (MSE).” Finally, the “Reported Health Transition (HT)”

item that is part of the SF-36v2 survey is now referred to as the “Self-Evaluated Transition (SET)” item. The reason for these changes is to more precisely describe what each term represents and thus minimize misconceptions about what the term means among users of the Short Form surveys and other consumers of information derived from the Short Form surveys. Although the terminology has changed, what each term represents and how it’s used remains unchanged.

Note that additional information about the history and development of the SF-36v2 can be found elsewhere (Turner-Bowker, DeRosa, & Ware, 2007; Ware, 2000; Ware & Kosinski, 2001b; Ware et al., 2007; Ware, Kosinski, & Keller, 1994; Ware, Snow, Kosinski, & Gandek,

1993). For example, a comprehensive overview of the survey can be found in Ware’s (2004) “The SF-36 Health Survey: An Update,” a chapter in *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment, Volume 3: Adult Assessment Instruments, Third Edition* (Maruish, 2004c).

New sources of information about the development and empirical testing of the SF-36v2 are available or forthcoming from QualityMetric Incorporated, as well as from other researchers. Interested readers are encouraged to go to QualityMetric’s website at <http://www.qualitymetric.com> or to the website for users of the Short Form family of instruments at <http://www.sf-36.org> for more information.

PART I:
INTRODUCTION

1

Introduction

The SF-36v2® Health Survey (SF-36v2) is a multipurpose, short-form health survey with 36 questions that yields an eight-scale profile of functional health and well-being, as well as two psychometrically based physical and mental health summary measures and a preference-based health utility index. Like its predecessor, the SF-36® Health Survey (SF-36; Ware, Snow, Kosinski, & Gandek, 1993), the SF-36v2 is a generic measure of health status, as opposed to one that targets a specific age, disease, or treatment group. It has proven useful for conducting surveys of general and specific populations, comparing the relative burden of diseases, and differentiating the health benefits produced by a wide range of treatments.

The main purpose of this chapter is to provide a summary of the circumstances and events that led to the development of the SF-36v2. The evolution of this instrument is presented through a brief review of the major health status studies that employed the SF-36v2's predecessors and subsequently resulted in the improvements embodied by this survey. This chapter also describes the developments in assessment technology (e.g., item response theory, computerized adaptive testing, and QualityMetric Incorporated's item banks) that have allowed for better empirical demonstrations of survey improvements. Finally, this chapter presents a new conceptual framework for health status assessment that utilizes disease-specific surveys that have been standardized across measures in both content and scoring and enables comparisons with the specific impact of other diseases.

Context for Health Status Assessment

During the 1980s, one of the more important developments in the healthcare field was the recognition of the centrality of the patient's point of view in monitoring

the quality of medical care outcomes (Geigle & Jones, 1990). A *medical outcome* has come to mean the extent to which a change in a patient's behavioral functioning or well-being meets the patient's needs or expectations. This sentiment was well-expressed in medical literature during the 20th century (Codman, 1991; Lembcke, 1952, as cited in Silver, 1990). To wit, 60 years ago, Lembcke (1952) wrote:

The best measure of quality is not how well or how frequently a medical service is given, but how closely the result approaches the fundamental objectives of prolonging life, relieving distress, restoring function and preventing disability.

These historical objectives were echoed in the 1980s by those arguing that the goal of medical care for most patients is the achievement of a more effective life (McDermott, 1981) and the preservation of function and well-being (American College of Physicians, 1988; Cluff, 1981; Ellwood, 1988; Schroeder, 1987; Tarlov, 1983). While the patient is the best source of information regarding the attainment of these goals, patients' experiences of their diseases and treatments were not routinely collected in clinical research or medical practice during this era. Because this sort of information was typically not a part of the medical record, it was unavailable for routine analysis.

In the 1990s, clinical investigators evaluating new treatments and technologies, as well as physicians and other providers trying to achieve the best possible patient outcomes, began to utilize information about functional status, well-being, and other important health outcomes. Policy analysts also began to use this information to compare the costs and benefits of competing methods of organizing and financing healthcare services, as did healthcare organization managers seeking to produce the best value for each healthcare dollar. Today, the primary source of new information regarding general health

outcomes is rapidly becoming the standardized patient surveys that have been effectively serving researchers for the past several decades.

Several advances in the methods for assessing patient perspectives about functional status, well-being, and other important healthcare outcomes occurred during the 1980s and 1990s. These advances have been the subjects of numerous conferences (Department of Health and Human Services, Agency for Health Care Policy and Research, 1999; Katz, 1987; Lohr, 1989, 1992; Lohr & Ware, 1987; Patrick & Chiang, 2000; Reeve, 2004; Wenger, Mattson, Furberg, & Elinson, 1984). To illustrate, some of the more significant of these advances include: (a) improved understanding of the major dimensions of health and of the validity of specific measurement scales in relation to those dimensions (Hays & Stewart, 1990; Liang, 1986; Ware, Brook, Davies, & Lohr, 1981), (b) demonstration of the usefulness of standardized health surveys in clinical trials (Bombardier et al., 1986; Croog et al., 1986; Fowler et al., 1988), (c) evaluations of health policy (Brook et al., 1983; Ware et al., 1986; Ware, Bayliss, et al., 1996), and (d) development of general population health surveys (Bergner, Bobbitt, Carter, & Gilson, 1981; McHorney, Kosinski, & Ware, 1994; Stewart, Hays, & Ware, 1988; Stewart et al., 1989; Ware et al., 1986).

Subsequently, these advances facilitated: (a) the use of self-assessed well-being in medical practice (Nelson & Berwick, 1987), (b) the formation of professional societies such as the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) and the International Society for Quality of Life Research (ISOQOL), (c) the introduction of item response theory (IRT) to the field of health status measurement (Avlund, Kreiner, & Schultz-Larsen, 1993; Bech et al., 1981; Granger, Hamilton, Linacre, Heinemann, & Wright, 1993; McHorney, Haley, & Ware, 1997), and (d) the introduction of computerized adaptive testing (Bjorner & Ware, 1998; Revicki & Cella, 1997; Ware, Bjorner, et al., 2000; Ware et al., 2003).

Improvement of Health Status Surveys

The use of standardized surveys to assess functional status and well-being can be traced back over 300 years. Methodological interest, however, has been greatest during the last half of the 20th century (Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963). As such, most health measures used prior to the 1970s were not based upon methods of scale construction, even though these psychometric techniques had been available for

most of the past century (Guttman, 1944; Likert, 1932; Thurstone & Chave, 1929). However, in the last 50 years, increasing interest in such methods has resulted in the construction of numerous psychometrically sound health status scales (Berki & Ashcraft, 1979; DiCocco & Apple, 1958; Dupuy, 1984; Ware, 1976a; Williams & Lindem, 1976).

Both the techniques for constructing health measures and the content of the measures have changed over time. For example, health measures previously limited their focus to the presence or absence of negative health status, functional limitations, disease symptoms, and acute and chronic problems. Today, some health measures still exclusively focus on such negative content (Kaplan, 1989). During the last half of the 20th century, however, the content of most published measures of functioning and well-being has undergone well-documented changes (Maruish, 2004a, 2004b, 2004c; McDowell & Newell, 1987; McHorney, 1997; Stewart & Ware, 1992; Ware, 1987, 1995; Ware, Davies-Avery, & Brook, 1978; Ware, Johnston, Davies-Avery, & Brook, 1979).

In recent years, more sophisticated psychometric methods, specifically IRT methodology (Fischer & Molenaar, 1995; van der Linden & Hambleton, 1997) and structural equation models for categorical data (Muthen, 1984), have been applied in the analyses of health status surveys (e.g., Bjorner, Kosinski, & Ware, 2003a; Bjorner, Kosinski, & Ware, 2003b; Bjorner, Kosinski, & Ware, 2003c; Haley, McHorney, & Ware, 1994; McHorney & Cohen, 2000; Orlando, Sherbourne, & Thissen, 2000). These techniques have been and can be used to obtain a more realistic assessment of measurement precision, to achieve better analyses of dimensionality (Bjorner et al., 2003a; Bjorner & Ware, 1998), and to evaluate differential item functioning (i.e., whether the survey performs in the same way with different subgroups; see Bjorner, Kreiner, Ware, Damsgaard, & Bech, 1998; Groenvold, Bjorner, Klee, & Kreiner, 1995; Raczek et al., 1998). Moreover, IRT provides a rationale for selecting the most informative items for a particular person or group (Ware et al., 2003; Ware, Bjorner, & Kosinski, 2000) and is utilized in computerized adaptive testing (CAT; van der Linden & Glas, 2000; Wainer et al., 2000). Both IRT and CAT are further discussed in later sections of this chapter.

The Evolution of Short Form Health Status Surveys

The Health Insurance Experiment (HIE)

One of the first extensive applications of psychometric theory and methods to the development and

refinement of health status surveys took place during the Health Insurance Experiment (HIE; Brook et al., 1983; Newhouse et al., 1993; Valdez et al., 1989; Ware et al., 1986). The HIE scales were constructed to measure a broad array of functional status and well-being concepts for group-level longitudinal analyses of data from children and non-aged adults. Data collection for the HIE took place between 1974 and 1981, and the work was summarized and published in an eight-volume set of RAND Corporation technical reports (Eisen, Donald, Ware, & Brook, 1980) and in *Medical Care* (Brook, Ware, Davies-Avery, et al., 1979). Results of the HIE clearly demonstrated the potential reliability and validity of scales constructed from self-administered surveys and the ability of such scales to yield high quality data for assessing changes in health status in the general population. Results also demonstrated that, with vigorous follow-up, the use of such measures could yield high completion rates. However, the HIE left two basic questions unanswered: (a) Can methods of data collection and scale construction such as those used in the HIE work with individuals who are older and those who have more health problems, and (b) can more efficient scales be constructed? Answering these questions became the challenge for the Medical Outcomes Study.

The Medical Outcomes Study (MOS)

The Medical Outcomes Study (MOS; see Stewart & Ware, 1992; Tarlov et al., 1989; Ware et al., 1996) was a 4-year longitudinal, observational study of the variations in practice styles and of the health outcomes for chronically ill patients. The MOS began at the University of Chicago in 1981 and was continued at the RAND Corporation and Tufts-New England Medical Center, with institutional collaborators from the University of Washington and Dartmouth Medical School. Over 23,000 patients from the practices of 362 medical clinicians and 161 mental health care providers in Boston, Chicago, and Los Angeles participated in the study. The MOS provided the opportunity for a large-scale test of the feasibility of self-administered patient questionnaires and generic health scales for those with chronic conditions, including elderly individuals. Pilot studies began in the early 1980s, with data collection taking place between 1986 and 1990 and data analyses occurring through the early 1990s.

The surveys of both the HIE and the MOS were based on a multidimensional model of health; however, the MOS surveys were more comprehensive, assessing a total of 40 health concepts. Significantly, the study's standardized questionnaires included the items that were subsequently selected and adapted by the principal inves-

tigator of the MOS when developing the SF-36. While the SF-36 represents eight of the most important health concepts included in the MOS and other widely used health surveys, the MOS surveys included questions measuring additional health concepts, including cognitive functioning, sleep, health distress, social support, family and marital functioning, sexual functioning, and physical and psychophysiologic symptoms.

The International Quality of Life Assessment (IQOLA) Project

In 1991, The Health Institute at Tufts-New England Medical Center began an organized effort to expand worldwide the use of health status instruments. The goal of this undertaking, referred to as the International Quality of Life Assessment (IQOLA) Project, was to develop validated translations of a single health status questionnaire that could then be used in multinational clinical studies and other international studies of health. The SF-36 was selected as the measure to be translated and used in the IQOLA Project for several reasons. For example, it is a brief, comprehensive measure of generic health status that can be easily supplemented with other generic or disease-specific measures. In addition, research on preliminary translations suggested that it could be successfully translated into several languages.

During its first year, five countries (France, Germany, Italy, Sweden, and the Netherlands) participated in the IQOLA Project. Additional researchers from other countries joined the project in 1992 and 1993, resulting in 14 countries being represented. Interest in developing translations of the SF-36 continued such that it was translated for use in more than 70 countries by 2006. The development and validation of these translated versions contributed to improvements in item wording and response categories, thereby leading to the development of the SF-36v2. The methods and results from the SF-36 translation and adaptation studies that were conducted for the IQOLA Project are described in a series of articles published in a special issue of the *Journal of Clinical Epidemiology* (Gandek & Ware, 1998b). Visit <http://www.iqola.org> for further information about the IQOLA Project and its translation methodology.

The Medicare Health Outcomes Study (HOS)

In 1997, the U.S. Congress passed the Balanced Budget Act (BBA), which, among other provisions, directed Medicare to begin focusing on the health status of its enrollees and to begin gathering data on the effectiveness of disease management strategies in this population (Haffer et al., 2003; Stevic, Haffer, Cooper, Adams, & Michael, 2000). Toward this end, the Centers

for Medicare and Medicaid Services (CMS) worked with the National Committee for Quality Assurance (NCQA) to incorporate the Medicare population into the Healthcare Effectiveness Data and Information Set (HEDIS®), which is widely used to measure the performance of managed health care plans. The CMS was also interested in expanding the HEDIS outcome measures to include more generic outcomes (i.e., outcomes that relate to patients regardless of their underlying diagnoses).

Partly in response to the findings reported by Ware, Bayliss, Rogers, Kosinski, and Tarlov (1996), an NCQA technical expert panel determined that the SF-36 should be used as the core measure for the Medicare Health Outcomes Survey (HOS), the annual assessment of the physical and mental health of Medicare beneficiaries enrolled in managed care plans (NCQA, 2004). From 1998 to 2004, the HOS's primary outcomes were the SF-36 Physical Component Summary (PCS) and Mental Component Summary (MCS) measures (scored using 1998 U.S. general population norms) and mortality. The HOS assessment instrument also includes questions to obtain information regarding limitations in activities of daily living (ADLs) and to gather data for use in case-mix and risk adjustment.

1998 National Survey of Functional Health Status (NSFHS)

Key to the development of the SF-36v2 was the 1998 National Survey of Functional Health Status (NSFHS), with U.S. general population norms being derived from SF-36v2 and SF-36 data gathered during this study. Panel households were drawn from the sampling frames maintained by National Family Opinion (NFO) Research. These households were demographically balanced according to the U.S. Census Bureau's four regions and nine divisions, as well as in correct proportion by state within each of the nine divisions. The NFO used a two-stage area probability sample design. In the first stage, quota sampling was used based on age, sex, and income. The primary sampling units (PSUs) used were Standard Metropolitan Statistical Areas, or non-metropolitan counties stratified by region, market size, age, income, and household size before selection. At the second stage, the units of selection were households stratified by age, sex, and race.

The National Research Corporation (NRC) collected data for 12 weeks between October and December 1998 using a single wave of questionnaires mailed to randomly selected members of the NFO panel. At the end of the data collection period, the overall response rate for the survey was 67.8%. A total of 7,069 respondents completed the standard (4-week recall) form and 7,837

completed the acute (1-week recall) form, with norms being separately developed for each form. Sampling weights were applied to adjust the samples to match the age, gender, and age-by-gender distribution of the 1998 census. To maximize the amount of useable data, Missing Score Estimation (MSE; formerly referred to as Missing Data Estimation [MDE]) was employed using the QualityMetric Health Outcomes™ Scoring Software (Saris-Baglana et al., 2004). The resulting norm-based *T* scores for both the SF-36v2 and SF-36's health domain scales and component summary measures have means of 50 and standard deviations of 10. Norms for the SF-6D, a health state utility index derived from the SF-36 (Brazier, Usherwood, Harper, & Thomas, 1998; see also Chapter 2), were also developed based on a scale ranging from 0.0 (worst health state) to 1.0 (best health state). Because health status scores for some domains significantly differ across age groups and for men and women, norms were developed for the total population (by both combined and separate age groups) and separately for males and females (again by both combined and separate age groups).

Finally, as part of the data gathering effort, participants were asked to indicate whether they were suffering from one or more of 18 diseases or physically impairing conditions. This information enabled the development of specific sets of norms for each of these conditions and disease states, norms that can provide important comparison information when interpreting SF-36v2 results from individual respondents or groups of respondents (see Chapter 7).

QualityMetric 2009 Norming Study

With the passage of more than a decade since the development of the 1998 norms, the developers of the SF-36v2 determined that updated norms were necessary to ensure that the Short Form surveys remained current and relevant to their users' needs. The normative data that were collected during the QualityMetric 2009 Norming Study allowed for this important updating of the SF-36v2's norms, as well as to the norms for the SF-12v2 health domain scales and component summary measures. Note that SF-8 normative data were also gathered during the 2009 norming study.

A primary goal of the QualityMetric 2009 Norming Study was the development of updated norms for the SF-36v2, SF-12v2, and SF-8 based on a large, representative sample of the U.S. general population. Normative data for other surveys published by QualityMetric were also collected as part of this project. Simultaneously collecting normative data for these other instruments allowed not only for the updating and/or further validation of

these surveys but also for the further validation of the SF-36v2 (see Chapter 16) and the development of additional ways to interpret the meanings of SF-36v2 scores (see Chapter 9). Chapter 14 provides a detailed discussion of the QualityMetric 2009 Norming Study, including findings from an investigation of the comparability of the 2009 and 1998 norms.

Improvements in Standards for Measurement Evaluation

Over the past few decades, several technological and psychometric advances have led to improvements in the measurement of health status and quality of life. These advances have not only increased the efficiency of gathering health-related data but have also led to improvements in measurement precision itself. The following sections briefly discuss the innovations that are particularly notable.

New Standards for Health Status Measurement: The Short Form Health Surveys

The development of psychometrically sound measures of physical and mental health status has been guided by standards that have served the needs of health care researchers and clinical communities for several of decades. (Note that a brief overview of some well-accepted sets of these standards is presented in Chapter 13 of this manual.) However, the realities of late 20th-century healthcare delivery and research created a context that necessitated a redefinition of traditional measurement standards in order to meet the demands of the context in which modern health care measurement takes place.

Specifically, the adoption of new standards became necessary for two reasons. First, the old standards addressed the wrong questions for the MOS approach. Traditionally, longer measures often prove to be more reliable and more valid (Manning, Newhouse, & Ware, 1982). The best tests, however, are those most clearly approximating the intended use of the measure (Kerlinger, 1973; McHorney, Ware, & Raczek, 1993; Ware, 1990), regardless of length. As a result, the new direction in health outcomes assessment called for updated standards to address two questions: (a) What concepts should be measured, and (b) how much measurement precision is enough for each concept and for a particular purpose?

The second reason for adopting new standards was that considerations of respondent burden and data collection costs prompted a rethinking of measurement

goals and, accordingly, the criteria used to construct and evaluate standardized health surveys. Excelling in relation to traditional psychometric standards of reliability, validity, and precision was no longer adequate. Instead, the new direction called for modern psychometric measures to be sensitive to the demands (i.e., burden) they place on both the respondent and the administrator, in terms of time and cost; to demonstrate an adequate range of measurement to avoid floor and ceiling effects while maintaining acceptable validity and reliability across the range of possible scores; to be understandable to respondents and other stakeholders in the respondents' care; and to be translatable and acceptable across a wide range of languages and cultural groups. As expected, opportunities to measure health status now routinely demand the best compromise between traditionally defined psychometric rigor and the new standards of feasibility and practicality. The SF-36 was developed with both of these considerations in mind.

SF-36 Health Survey

The SF-36 was first made available in 1988 in a "developmental" form (Ware, 1988) and then in 1990 in the standard form (i.e., SF-36; Ware et al., 1993). Constructed to satisfy the minimum psychometric standards necessary for group comparisons, the eight health domains represented in the SF-36 profile were selected from the 40 domains that were included in the MOS. As previously mentioned, those chosen represent the health domains believed to be most affected by disease and health conditions and those most frequently measured by other widely used health surveys (Ware, 1995; Ware et al., 1993). The SF-36 items represent multiple operational indicators of health, including behavioral function and dysfunction, distress and well-being, objective reports and subjective ratings, and both favorable and unfavorable self-evaluations of general health status (Ware et al., 1993).

The relative shortness of the SF-36 makes it a more practical choice than the lengthier research tools that served as points of departure in the development of the survey; consequently, it requires less in terms of respondent time and the costs associated with collecting and processing data. Another benefit of SF-36 use is that, for the great majority of respondents, it can be self-administered. The current reliance on self-administration as the primary mode of data collection, even for surveys with more than 250 questions, is partially rooted in the successful use of relatively lengthy self-administered questionnaires in the MOS (Stewart & Ware, 1992). The use of self-administered surveys was adopted in the MOS on the strength of pilot studies demonstrating that

self-administration worked well with chronically ill and elderly participants.

With the SF-36 came the establishment of a new standard of evaluation: The MOS team evaluated the SF-36 scales in terms of their relative performance as judged by formal tests using external criteria, such as their validity in discriminating among diagnostic groups known to differ in morbidity and in predicting subsequent utilization of healthcare resources. Others have published the results of such tests and have also expanded their efforts to include tests of sensitivity to change over time (Katz, Larson, Phillips, Fossel, & Liang, 1992).

SF-12 Health Survey

The SF-36 became the most widely used health survey throughout the world because it is brief yet comprehensive, readily available, psychometrically sound, and of proven usefulness in measuring health status and monitoring health outcomes in both general and specific populations. However, even the SF-36 was judged to be too long for some large-scale surveys limited in the amount of health information that could be collected in only a few minutes of interviewing time or limited in the number of questions and response options that could fit on one to three pages of a self-administered questionnaire. In response to these issues, the goal for the SF-12 was to develop a one-page, 2-minute questionnaire module. 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. Because the Physical Component Summary (PCS) and Mental Component Summary (MCS) measure scores from the SF-36 had proven useful for many purposes (Ware & Kosinski, 2001b; Ware, Kosinski, Bayliss, et al., 1995; Ware, Kosinski, & Keller, 1994), the strategy for the SF-12 was to construct the shortest possible form that would reproduce those two summary measures with at least 90% accuracy.

The SF-12 is a short-form health status survey with just 12 questions, all selected from the SF-36 (Ware, Kosinski, & Keller, 1995, 1996). Like the SF-36, it is a generic measure, as opposed to one that targets a specific age, disease, or treatment group. The SF-12 was developed to be a much shorter, yet still valid, alternative to the SF-36. At the time of its development early in 1994, it was thought that only the physical and mental summary measure scores were estimable from the SF-12 and that these scores would be useful only in large-population surveys. However, SF-12 PCS and MCS scores proved to be very useful in measuring outcomes in clinical trials. Fortunately, the survey developers also sought

to represent each of the eight SF-36 health concepts with one or two questionnaire items (Ware, Kosinski, & Keller, 1995, 1996), setting the stage for scoring an eight-scale profile from SF-12 responses.

SF-36v2 Health Survey

Although the SF-36 proved to be useful for many purposes, 10 years of use in the field revealed the need and potential for improvements. For example, the IQOLA Project's efforts to translate the SF-36 form demonstrated the need for improved item wording and response choice categories. These needs, combined with the opportunity to collect updated normative data, led to a revision of the survey. Thus, in the early 1990s, studies were initiated to address the aforementioned problems associated with wording and response choices and to resolve the well-documented shortcomings of the two role-functioning scales (J. E. Ware, Jr., & M. Kosinski, personal communication, September 1996). The result of these efforts was the development of the SF-36v2.

Like its predecessor, the SF-36v2 is a multipurpose, 36-item health survey yielding a profile that comprises two health component summary measures and eight health domain scales. Both versions can be used across all adult patient and nonpatient populations for a variety of purposes, such as screening individual respondents, monitoring the results of care, comparing the relative burden of diseases, and comparing the benefits of different treatments (e.g., Baldwin et al., 2009; Bird et al., 2010; Crespi, Smith, Petersen, Zimmerman, & Ganz, 2010; Elston, Honan, Powell, Gormley, & Stein, 2010; Fernandez-Fairen, Sala, Ramirez, & Gil, 2007; Fitzgibbons et al., 2006; Greenfield et al., 2010; Hudson et al., 2009; Jenkinson & Stewart-Brown, 1999; Kim, Sim, Jeong, & Kim, 2010; Kosinski et al., 2005; Laslett, Burnet, Jones, Redmond, & McNeil, 2007; Martin et al., 2005; McCune et al., 2006; Morfeld, Bullinger, Nantke, & Braehler, 2005; Motalebzadeh, Bland, Markus, Kaski, & Jahangiri, 2006; Nicholson, Ross, Sasaki, & Weil, 2006; Ochiai, Hagino, Tonotsuka, & Haro, 2010; Poole & Mason, 2005; Razvi, Ingoe, McMillan, & Weaver, 2005; Ware, Kosinski, & Bjorner, 2004; Wrennick, Schneider, & Monga, 2005; Wyrwich et al., 2006). Relative to the SF-36, however, the SF-36v2 offers: (a) improved instructions and minimized ambiguity and bias in item wording, (b) improved layout of questions and answers, (c) increased comparability in relation to translations and cultural adaptations, (d) five-level response choices in place of dichotomous choices for the seven items in the Role-Physical and Role-Emotional scales, and (e) elimination of a response option from the items in the Mental Health and Vitality scales.

These improvements were implemented after thorough evaluation of their advantages. Made available for use by the research and clinical communities in 1996 (Ware & Kosinski, 1996), the SF-36v2—sometimes referred to as the “international” version—represents an improved measurement tool that maintains comparability with its original version in terms of purpose, content, scoring, and the psychometric rigor with which it was developed. For example, without increasing the number of items, the SF-36v2 provides substantially increased score reliability and validity and simplified language that makes the survey easier to understand and complete. Furthermore, the adoption of the *T*-score metric makes it possible to compare results across both versions of the SF-36 surveys, thereby eliminating concerns about loss of comparability. Also note that *T*-score linear transformations do not change the interpretation of significance of difference in group-level comparisons. Finally, use of the *T*-score metric results in all health domain scales and component summary measures having means of 50 and standard deviations of 10, now based on the new 2009 U.S. general population normative data (see Chapter 14).

Studies of diverse populations in both the United States and abroad provide clear evidence that the advantages of the SF-36v2 are substantial (Jenkinson, Stewart-Brown, Petersen, & Paice, 1999). To illustrate, its domains have improved reliability over the original version of the United Kingdom SF-36. Furthermore, the enhancements made to item wording and response categories have reduced the extent of floor and ceiling effects in the role-functioning scales (see Chapter 13). These advances will likely lead to better precision and greater responsiveness in longitudinal studies.

Although standardized comprehensive measures of generic functional status and well-being existed prior to the SF-36 (e.g., the Sickness Impact Profile [SIP; Bergner et al., 1981]), no instrument had received widespread adoption, nor had any one measure been shown to be suitable for use across diverse populations and healthcare settings. As a result, little was known about how healthy patients and those suffering from various chronic medical or psychiatric conditions differed from each other in terms of functional status and well-being because clinicians and researchers were unable to assess and describe such differences. Filling this gap, the SF-36v2 maintains comparability with the SF-36 and, like its predecessor, provides a common metric to compare those respondents with chronic health problems to those sampled from the general population.

Ten years after the development of the SF-36v2's 1998 norms, the developers of the Short Form surveys determined that a normative update was necessary to

ensure that the surveys remained current and relevant to the users' needs. To this end, the QualityMetric 2009 Norming Study was conducted to provide up-to-date norms for the SF-36v2, SF-12v2[®] Health Survey (SF-12v2; Ware, Kosinski, Gandek, Sundaram, Bjorner, et al., 2010), and SF-8[™] Health Survey (SF-8; Ware, Kosinski, Dewey, & Gandek, 2001) and to obtain additional validation data for these three surveys, as well as for the other QualityMetric patient-reported outcomes (PRO) surveys. The publication of these most recent norms also served as the impetus for revising this manual, which includes a re-evaluation of the SF-36v2's reliability, validity, and usefulness based on the 2009 norms.

SF-12v2 Health Survey

As discussed earlier, several developments provided the foundation for the construction of the SF-12 in 1994 and for the substantial improvements that are now reflected in the SF-12v2. These developments included findings that (a) physical and mental health factors accounted for 80 to 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 (McHorney et al., 1993; Ware, Keller, Gandek, Brazier, & Sullivan, 1995; Ware et al., 1993) and (b) the SF-36 PCS and MCS measures very rarely missed hypothesized differences in cross-sectional and longitudinal tests based on independent physical and mental criterion variables (Ware & Kosinski, 2001b; Ware, Kosinski, Bayliss, et al., 1995; Ware et al., 1994).

These results suggested that it may be possible to further reduce the number of items in the SF-36 without substantial loss of information. More recently, the creation and calibration of eight comprehensive “pools” of questionnaire items—one pool of items for each of the eight concepts measured by the SF-36 and SF-12 surveys—made it possible to evaluate the practical implications of improvements in question wording and item response categories (Ware, 2008). With old and new items in each of these pools calibrated in relation to a common standard metric, a much better criterion was available for estimating the practical implications of improvements being considered. For example, these studies revealed that the change from dichotomous to five-choice response categories in the two role functioning scales would lead to substantial increases in the ranges measured by both of these scales. Given the well-documented problems with ceiling and floor effects in studies using the SF-12, these improvements were noteworthy. The item calibrations from IRT models also provided a basis for evaluating scoring algorithms for the one-item and two-item scales representing the eight dimensions of health assessed in the SF-12. Without the

benefit of this information, the developers had initially recommended against reliance on scores estimated from these relatively coarse scales when the SF-12 was first published (Ware, Kosinski, & Keller, 1995).

With improved scoring algorithms and a better understanding of the relationship between sample size and score precision in group-level studies of health status, it became feasible to score the eight-scale profile in addition to the PCS and MCS measures using the SF-12v2. In fact, the *T*-score algorithms for the eight SF-12v2 scales yield unbiased estimates of scores for the corresponding SF-36v2 health domain scales in the U.S. general population. Norms developed from QualityMetric 2009 Norming Study data are now used to score the SF-12v2.

SF-8 Health Survey

The SF-8 was preceded by the SF-6 Health Survey (SF-6; Ware, Kosinski, Dewey, & Gandek, 2001). The SF-6 was developed primarily for use in large surveys of general and specific populations in which precision is achieved much more by utilizing a large sample than by increasing measurement reliability. It measures seven of the eight health domains (i.e., excluding Vitality) measured by the other Short Form instruments, and it was administered and evaluated in the MOS (Ware, Nelson, Sherbourne, & Stewart, 1992). The SF-8 forms were constructed nearly 10 years later and were then compared with the original SF-6.

The SF-8 includes the single best available item measuring each of the eight Short Form health concepts. With one exception, none of the SF-8 items are identical to those in any of the other Short Form surveys, although some are very similar. For each Short Form health concept, the SF-8 item selected maximizes the discrimination between higher and lower levels of health status, as defined by the corresponding Short Form health domain scale, and covers a wide range of score levels. In most cases, SF-8 items discriminate better and/or cover a wider range than the best performing SF-36/SF-36v2 item measuring the same concept.

To maintain comparability, it was not necessary to limit the pool of potential SF-8 items to the items in the SF-36/SF-36v2, thanks to advances in psychometric methods. Comparability was achieved by standardizing the metric underlying each of the health concepts. Because the SF-8 single-item scales and its summary measures are scored on the same metric as the SF-36/SF-36v2 and SF-12/SF-12v2, their scores are directly comparable. Average scores based on SF-8 measures are unbiased, albeit “noisier,” estimates of the scores for the same measures from other Short Form surveys.

The SF-8 was developed primarily for use in large surveys of general and specific populations in which precision is achieved much more by drawing a large sample than by increasing measurement reliability. However, the usefulness of the SF-8, as well as the SF-6, in clinical trials and outcomes research based on much smaller samples has already proven to be a subject of great interest and considerable research (e.g., Aoki, Fleming, Griffin, Lacey, & Edmundson, 2000; Paterson et al., 2000; Silagy, Griffin, Lacey, & Edmundson, 1998).

In comparison with the SF-36/SF-36v2, the SF-8 has a number of advantages. It is substantially shorter and yields directly comparable estimates of scores for all eight health domains and both component summary measures. Another advantage of the SF-8 is that versions of the survey have been developed and validated for three different recall periods: standard (4-week), acute (1-week), and 24-hour. Further, to ensure the usefulness of the SF-8 in multinational studies, the wording of SF-8 items and instructions were not finalized until they were successfully translated and adapted for use in more than 15 countries. Thus, the SF-8 is likely to be less culture-specific and more accessible. A major disadvantage of the SF-8 is that its scores cover a narrower range than the SF-36v2 and are less precise. Also, as of the publication of this manual, 2009 norms have not been developed for the SF-8. Thus, at this time scoring of the SF-8 is based on the survey’s 2000 norms.

The SF-36v2, SF-12v2, and SF-8 are now the key members of a “family” of fixed-length, short-form measures. Each can be administered and then scored on the norm-based *T*-score metric using QualityMetric Incorporated’s Smart Measurement™ System or its QualityMetric Health Outcomes™ Scoring Software 5.0 (Saris-Baglama et al., 2011; see also Chapter 5).

QualityMetric’s Item Banks and Computerized Adaptive Testing (CAT) Tool

While the SF-12v2 and SF-8 Health Surveys represent valid options for assessing the eight Short Form domains using fewer items than the SF-36v2, QualityMetric Incorporated’s item banks and computerized adaptive testing (CAT) system provide the option of assessing those same domains with even higher precision and greater range coverage than even the SF-36v2. In 2000, seven national norming studies were conducted to develop item banks for seven of the eight SF health domains. These seven studies included nearly 6,500 assessments completed via

the Internet and another 4,500 assessments completed by telephone interview. Internet respondents were recruited from AOL's Opinion Place (see Ware, Kosinski, Dewey, & Gandek, 2001, for a detailed description). Quotas were used to ensure that the final sample was approximately representative of the distribution of age and gender found in the U.S. general population.

In total, seven item banks, one each for seven of the eight Short Form health domain scales (i.e., excluding the General Health scale), were developed from the seven national norming studies. Each national norming study consisted of a survey containing items from one of the health domain scales, along with additional items that were selected from 52 published health status instruments measuring the same health concept as the health domain scale in question. To build the item banks, these norming studies surveyed a total of 305 items, ranging from 18 to 61 items per health domain. IRT methods were then used to calibrate and score the items from the seven item banks on a single, unidimensional scale.

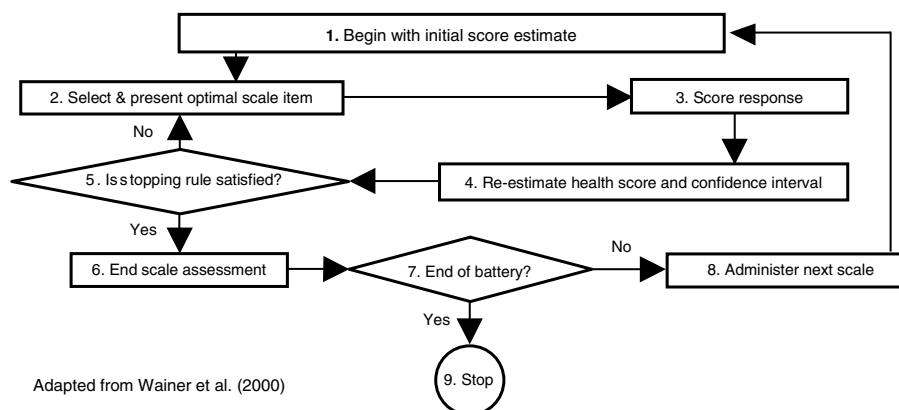
As previously noted, an item bank for the General Health (GH) scale was not included in any of the seven national norming studies conducted in 2000. Uniquely, the data for the GH item bank was obtained from the results of the Medical Outcomes Study (Stewart & Ware, 1992), which fielded all 31 items from the General Health Rating Index (GHRI; Davies & Ware, 1981; Ware, Davies-Avery, & Donald, 1978). Using IRT methods, the baseline data set ($N = 3,445$) was used to identify and calibrate a homogeneous set of 12 items.

The finalized item banks serve to cross-calibrate items from the SF-36 and SF-36v2 with items from other established measures, thus providing a deeper understanding of the breadth of the items' coverage across each domain and helping to identify their areas

of strength and weakness in the measurement of health status. The QualityMetric item banks also allow for the use of CAT technology to assess the eight health domains, resulting in even greater precision and fewer floor and ceiling problems than can be obtained when using the SF-36v2. The basic premise of a CAT system is to mimic what an experienced clinician would do: direct questions at the respondent's approximate level of health and functioning (Bjorner, Kosinski, & Ware, 2005; Ware, Bjorner, & Kosinski, 1999). For example, an adult who is able to "walk 50 feet" need not be asked a question about "walking 10 feet." CAT systems employ a simple form of artificial intelligence that selects questions tailored to the respondent, scores all respondents on a standard metric so that results can be compared, shortens or lengthens the survey to achieve the desired precision, and instantly displays survey results (see Figure 1.1; van der Linden & Glas, 2000; Wainer et al., 2000; Weiss, 1983). By altering the stopping rule, it becomes possible to match the level of score precision to the specific measurement purpose for each respondent (Bjorner et al., 2005; Ware et al., 2003). For example, more scoring precision would be needed to monitor individual progress than to assess the health status of a group of respondents.

QualityMetric Incorporated offers CAT assessment of generic and disease-specific health domains via its patented DYNHA[®] Computerized Adaptive Health Assessments engine (U.S. Patent No. 7765113B2, 2010). The DYNHA engine builds on principles from item response theory and CAT logic (Fischer et al., 1995; van der Linden et al., 1997), thus creating a set of psychometric models that describes item response probabilities as a function of item characteristics and the individual's level of health-related quality of life (HRQL).

Figure 1.1 Logic of Computerized Adaptive Testing



Adapted from *Computerized Adaptive Testing: A Primer* by H. Wainer, N. J. Dorans, R. Flaugher, R. J. Mislevy, D. Thissen, D. Eignor, B. F. Green, et al., 2000. Copyright 2000 by Lawrence Erlbaum Associates

A New Conceptual Framework for Health Status Assessment

Over the past decades, the substantial growth in the number of health status assessment tools has broadened the range of domains available for assessment and enabled researchers and clinicians to better understand the impact of disease from the patient's perspective (McHorney, 1997; Ware, 2003). However, it is difficult to compare results from different measurement tools. This is particularly true for *disease-, condition-, or procedure-specific measures*, which focus on the particulars of a specific disease or diagnostic group (e.g., diabetes, cancer), condition (e.g., congestive heart failure, low back pain), or treatment (e.g., hip or knee replacement), respectively.

In contrast to disease-specific measures, the Short Form family of instruments includes all *generic, or general, measures*; that is, they all assess health concepts that represent basic human values that are relevant to everyone's functional status and well-being, regardless of age, disease, or treatment group (Ware, 1987, 1990). The term *generic* not only implies that these measures are universally valued but also that they are not age-, disease-, condition-, or treatment-specific.

Despite their contributions to health status assessment, generic health measures are not designed or intended to serve as substitutes for traditional measures of clinical endpoints. To the contrary, this decade's greatest advances in this field are likely to come from studies that test generic health measures in parallel with clinical measures. The findings obtained from the combined use of these measures will not always be parallel; however, understanding the differences will lead to progress in this field of endeavor. The potential of such comparisons can be illustrated in the profiles of functional status and well-being for respondents with different medical and psychiatric conditions, as well as in contrast to profiles for the U.S. general population (see Chapter 14). These comparisons serve at least two important purposes. First, they test the validity of SF-36v2 health domain scales and component summary measures with regards to describing groups of respondents known to differ in functional status and well-being. Second, they facilitate understanding amongst clinicians regarding the meaning of SF-36v2 score differences, due to their familiarity with diagnostic groups.

Typically, evaluating the impact of diseases on health status has been performed using both generic and disease-specific measures. In general, disease-specific measures demonstrate greater sensitivity (Bombardier et al., 1995; Kantz, Harris, Levitsky, Ware, & Davies,

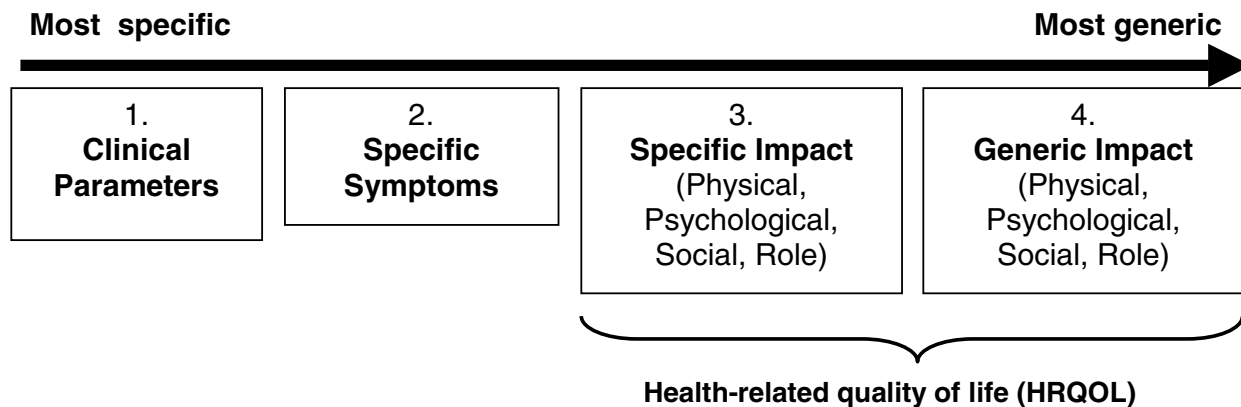
1992) and specificity than generic measures (Kantz et al., 1992), while generic measures better capture the total burden of disease (Bombardier et al., 1995; Ware, 1995). In the presence of comorbid conditions, generic measures reflect the combined effects of the primary and comorbid conditions, whereas disease-specific measures mainly reflect the primary disease (Kantz et al., 1992).

Figure 1.2 presents a conceptual framework for constructing and describing the relationships between the disease-specific and generic HRQOL measures used in clinical outcomes research. This framework makes important distinctions between domains of health and their operational definitions. To wit, note that Figure 1.2 portrays a specific–generic continuum (Ware, 1995; Ware, 2003; Wilson & Cleary, 1995) rather than simply categorizing specific and generic concepts and measures. Thus, moving from the left of the figure to the right, the measures shift from the most highly specific and objective clinical measures (Category 1), to disease-specific symptoms (Category 2), to specific measures of disease impact (Category 3), to generic measures that are applicable across chronic disease and treatment groups (Category 4). To illustrate, measures listed in Categories 3 and 4 attempt to capture specific and generic HRQOL impact with questions concerning limitations in role participation due to a specific disease versus questions about the same limitations without attribution to a specific disease, respectively.

Measures on the left (Categories 1 and 2) are the most specific and, therefore, most useful for making a diagnosis and determining the severity of a specific condition (Deyo & Patrick, 1989; Patrick & Deyo, 1989; Patrick & Erickson, 1988). In contrast, measures on the right (Categories 3 and 4) are more useful for understanding the impact (on functioning and well-being) of disease and treatment in the more distal HRQOL terms that matter most to patients. Therefore, in comparison with measures in Category 2, those in Category 3 are considered HRQOL measures because they capture the social and economic impact of disease and treatment. In comparison with Category 3, those in Category 4 (e.g., Sickness Impact Profile, SF-36v2) permit meaningful comparisons across disease and treatment groups because they are the most generic measures and are not specific to a disease or treatment (e.g., Bergner et al., 1976; Stewart et al., 1989).

As conceptualized and measured to date, the gains made in specificity when using disease-specific HRQOL measures (Category 3) have been achieved at the expense of the ability to make meaningful comparisons of burdens across diseases and of benefits across treatments. To this end, QualityMetric Incorporated launched the

Figure 1.2 Patient-Reported Outcomes (PRO) Conceptual Framework

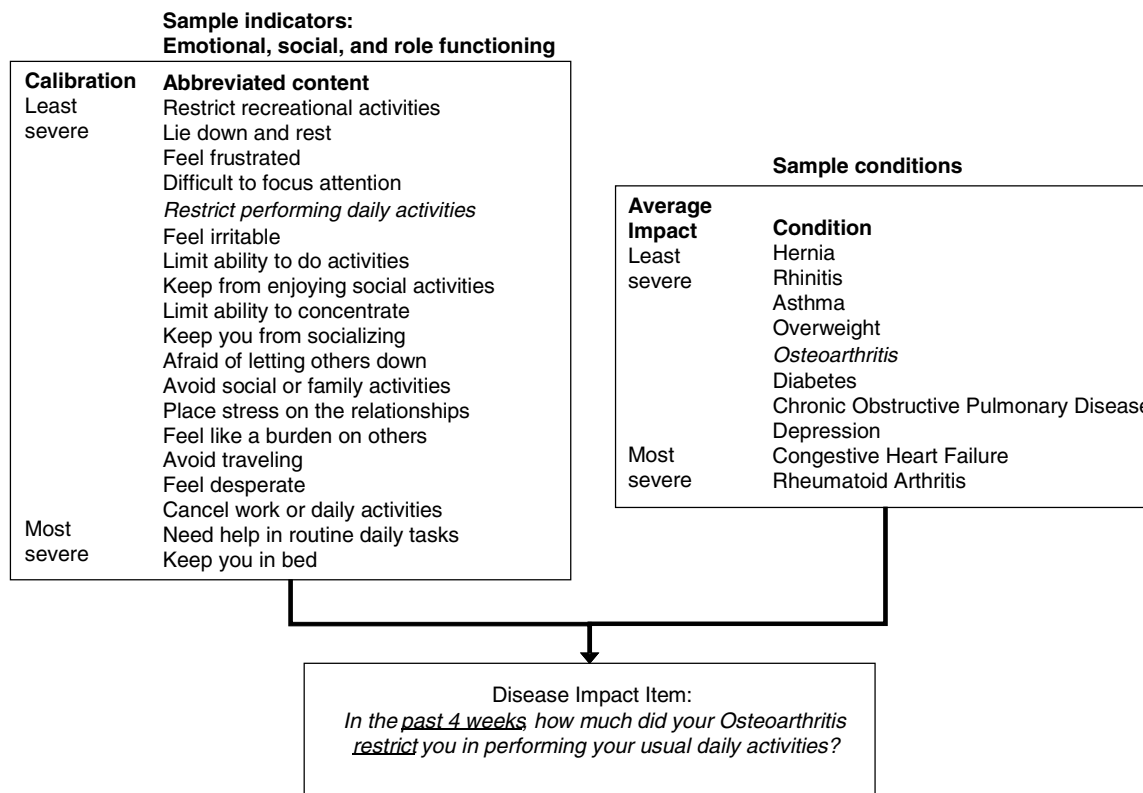


Disease Impact Project to standardize domain content and scoring algorithms across a number of tools with disease-specific attributions (e.g., limited in social activity because of diabetes, limited in social activity because of heart failure). The goal of such standardization is to achieve comparability of scores, even amongst those from specific instruments for different diseases (see Figure 1.3).

The conceptual framework illustrated in Figure 1.2 also makes useful distinctions between the content of

differing measures and helps to illustrate the importance of un-confounding measures across the four categories. For example, when symptom frequency and/or severity is assessed and scored separately (Category 2) and the associated specific impact is assessed and scored separately (Category 3), the implications of different symptoms can be meaningfully studied and interpreted in terms of their impact on HRQOL, in specific (Category 3) or generic (Category 4) terms.

Figure 1.3 Components of Disease Impact Items



Use of This Manual

The *User's Manual for the SF-36v2 Health Survey, Third Edition (User's Manual)*; Maruish [Ed.], 2011) was developed to provide those using the SF-36v2—clinicians, researchers, leaders of quality improvement organizations, and healthcare organization managers, to name a few—with all the information necessary to familiarize themselves with and to properly use the instrument, specifically with regards to the recently released 2009 U.S. general population norms. This edition of the *User's Manual* is organized such that Part I provides an introduction to the SF-36v2 and the other members of the Short Form family of health surveys (Chapters 1–3). Next, Parts II and III (Chapters 4–12) present the information most useful to those who want to quickly begin using the survey, including how to properly administer, score, and interpret the SF-36v2. Note that the edges of the pages contained in Parts II and III are screened in gray for easy location. Finally, Part

IV (Chapters 13–17) discusses the development of the SF-36v2 and its predecessors, the development of the 2009 norms, and the survey's psychometric properties. Regardless of the intended use, it is recommended that all survey users familiarize themselves with the content of this entire manual.

The *User's Manual for the SF-36v2 Health Survey, Third Edition* presents the most current information regarding the SF-36v2 at the time of its publication. With time, this manual's wealth of information will be enhanced by knowledge gained from newly published articles, books, and reports stemming from efforts to further investigate the utility and psychometric integrity of the instrument. Although QualityMetric Incorporated will strive to keep users apprised of newly published information that represents significant strides in understanding the survey and its uses, those employing the SF-36v2 for any purpose are encouraged to keep abreast of the literature on the instrument as it becomes available.

2

Concepts, Measures, and Applications

The SF-36 was developed to be a brief, broad, generic measure of eight domains, or aspects, of health status that are considered important in describing and monitoring individuals suffering from a disease or illness. It measures these domains in terms of functioning and personal evaluations, but it was not intended to be a comprehensive survey of health. A discussion of the criteria used to select the SF-36 domains and the items used to measure those domains is presented in Chapter 13 of this manual. The SF-36v2 maintains comparability with its predecessor by retaining, while improving on, the same domains, component summary measures, and items as the original version of the instrument.

The purpose of this chapter is twofold. First, it provides the SF-36v2 user with a general description of the health domain scales, the items they comprise, and the two component summary measures. Detailed information about the development and psychometric properties of the instrument is provided in Chapters 13 through 17. Guidance for interpreting the health domain scales and component summary measures is provided in Chapters 6 through 12. Second, many of the common applications of the SF-36v2 are identified and discussed, using examples from the more extensive SF-36 published literature. These examples pertain to the SF-36v2 because of the comparability of the two versions of the survey (see Chapter 13). The types of applications identified here should not be considered exhaustive; rather, they should be viewed as ways in which either or both instruments have demonstrated their value in the past. Users may find additional appropriate applications of the information that can be obtained from the SF-36v2.

Concepts and Measures

The SF-36v2 includes one scale for each of eight measured health domains: physical functioning, role

participation with physical health problems (role-physical), bodily pain, general health, vitality, social functioning, role participation with emotional health problems (role-emotional), and mental health. All health domain scales are scored such that higher scores indicate better health. These scales are the same as those developed for the SF-36, and the items that constitute them are identical in content (i.e., modified as part of the revision, as explained in Chapter 13) as those found in the original version.

Health Domain Scales

Physical Functioning (PF). The content of the 10-item PF scale reflects the importance of distinct aspects of physical functioning and the necessity of sampling a range of severe and minor physical limitations. Items represent levels and kinds of limitations between the extremes of physical activities, including lifting and carrying groceries; climbing stairs; bending, kneeling, or stooping; and walking moderate distances. One self-care item is included to represent limitations in self-care activities. The PF items capture both the presence and extent of physical limitations using a three-level response continuum. Low scores indicate significant limitations in performing physical activities, while high scores reflect little or no such limitations.

Role-Physical (RP). The four-item RP scale covers an array of physical health-related role limitations, including (a) limitations in the kind of work or other usual activities, (b) reductions in the amount of time spent on work or other usual activities, (c) difficulty performing work or other usual activities, and (d) accomplishing less. Low scores on the RP scale reflect problems with work or other activities as a result of physical problems. High scores indicate little or no problems with work or other daily activities.

Bodily Pain (BP). The BP scale comprises two items: one pertaining to the intensity of bodily pain and

one measuring the extent of interference with normal work activities due to pain. Low scores indicate high levels of pain that impact normal activities, while high scores indicate no pain and no impact on normal activities.

General Health (GH). The GH scale consists of five items, including a rating of health (*excellent to poor*) and four items addressing the respondent's views and expectations of his or her health. Low scores indicate evaluation of general health as poor and likely to get worse. High scores indicate that the respondent evaluates his or her health most favorably.

Vitality (VT). This four-item measure of vitality (i.e., energy level and fatigue) was developed to capture differences in subjective well-being. Low scores indicate feelings of tiredness and being worn out. High scores indicate feeling full of energy all or most of the time.

Social Functioning (SF). This two-item scale assesses health-related effects on quantity and quality of social activities, asking specifically about the impact of either physical or emotional problems on social activities. The degree to which physical and emotional problems interfere with normal social activities increases with decreasing SF scores. The lowest score is related to extreme or frequent interference with normal social activities due to physical and emotional problems; the highest score indicates that the individual performs normal social activities without interference from physical or emotional problems.

Role-Emotional (RE). The three-item RE scale assesses mental health-related role limitations in terms of (a) time spent on work or other usual activities, (b) amount of work or activities accomplished, and (c) the care with which work or other activities were performed. Low scores on this scale reflect problems with work or other activities as a result of emotional problems. High scores reflect no limitations due to emotional problems.

Mental Health (MH). The five-item MH scale includes one or more items from each of four major mental health dimensions (anxiety, depression, loss of behavioral/emotional control, and psychological well-being). Low scores on MH are indicative of frequent feelings of nervousness and depression, while high scores indicate feelings of peace, happiness, and calm all or most of the time.

Self-Evaluated Transition (SET). Formerly referred to as *Reported Health Transition*, this general health item asks respondents to rate the amount of change they experienced in their health, in general, over a 1-year period on the standard (4-week) form or over a 1-week period on the acute (1-week) form. This item is not used to score any of the eight health domain scales or component summary measures; however, it does provide

useful information about perceived changes in health status that occurred during the year (standard form) or week (acute form) prior to survey administration. If clinical or research needs require the measurement of reported health transition over a period other than 1 year or 1 week (e.g., during the past 3 months), the user may use this item as a template for developing a more time-relevant item that would be administered *in addition to* the standard SET item.

The content of each SF-36v2 item is summarized in Table 2.1.

Physical and Mental Component Summary (PCS and MCS) Measures

Figure 2.1 illustrates the measurement model underlying the construction of the SF-36v2 multi-item health domain scales and component summary measures. This model has three levels: (a) items, (b) health domain scales that aggregate items, and (c) component summary measures that aggregate the health domain scales. The aggregates of the health domain scales are referred to as *component* summary measures because they were derived and scored using a factor analytic method called principal *components* analysis (Harman, 1976; see also Chapter 13). Although they reflect the two broad components, or aspects, of health—physical and mental—*all* of the eight health domain scales are used to score *both* component summary measures. All but 1 of the 36 items (Item 2, Self-Evaluated Transition) is used to score the eight health domain scales.

Factor analyses of correlations among the eight health domain scales of each version of the survey have consistently identified two factors (Ware, Kosinski, Bayliss, et al., 1995; Ware et al., 2007; Ware et al., 1998; Ware, Kosinski, & Keller, 1994). Based on the strength of the pattern of their correlations with the eight scales, the two factors have been interpreted as *physical* and *mental* components of health status. Three scales (PF, RP, and BP) correlate most highly with the physical component and contribute most to scoring of the Physical Component Summary (PCS) measure. The mental component correlates most highly with the MH, RE, and SF scales, which contribute most to the scoring of the Mental Component Summary (MCS) measure. Three of the scales have noteworthy correlations with both components: the VT correlates substantially with both but higher with the mental component, GH correlates with both but higher with the physical component, and SF correlates much higher with the mental component.

The PCS and MCS measures were constructed and scored to achieve a number of advantages in addition to

Table 2.1*Abbreviated Item Content for the SF-36v2 Health Domain Scales*

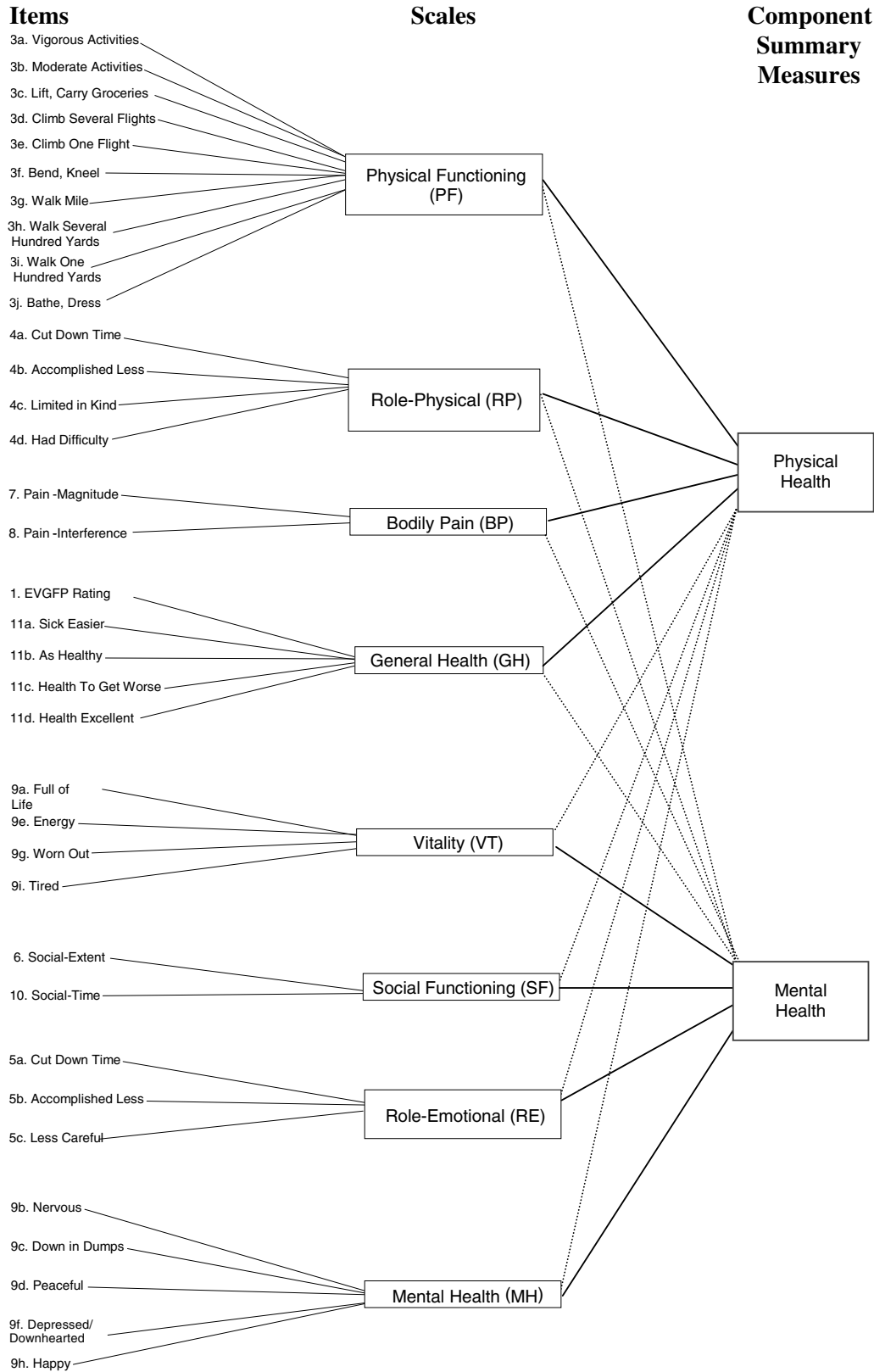
Scale	Item	Abbreviated Item Content
Physical Functioning (PF)	3a	Vigorous activities, such as running, lifting heavy objects, or participating in strenuous sports
	3b	Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf
	3c	Lifting or carrying groceries
	3d	Climbing several flights of stairs
	3e	Climbing one flight of stairs
	3f	Bending, kneeling, or stooping
	3g	Walking more than a mile
	3h	Walking several hundred yards
	3i	Walking one hundred yards
	3j	Bathing or dressing oneself
Role-Physical (RP)	4a	Cut down the amount of time spent on work or other activities
	4b	Accomplished less than you would like
	4c	Limited in kind of work or other activities
	4d	Had difficulty performing work or other activities (e.g., it took extra effort)
Bodily Pain (BP)	7	Intensity of bodily pain
	8	Extent pain interfered with normal work
General Health (GH)	1	Is your health: excellent, very good, good, fair, poor
	11a	Seem to get sick a little easier than other people
	11b	As healthy as anybody I know
	11c	Expect my health to get worse
	11d	Health is excellent
Vitality (VT)	9a	Feel full of life
	9e	Have a lot of energy
	9g	Feel worn out
	9i	Feel tired
Social Functioning (SF)	6	Extent health problems interfered with normal social activities
	10	Frequency health problems interfered with social activities
Role-Emotional (RE)	5a	Cut down the amount of time spent on work or other activities
	5b	Accomplished less than you would like
	5c	Did work or other activities less carefully than usual
Mental Health (MH)	9b	Been very nervous
	9c	Felt so down in the dumps that nothing could cheer you up
	9d	Felt calm and peaceful
	9f	Felt downhearted and depressed
	9h	Been happy
Self-Evaluated Transition (SET)	2	How health is now compared to 1 year ago

reducing the eight-scale profile to two component summary measures without substantial loss of information. Features of the PCS and MCS scores for the standard and acute SF-36v2 forms—including their reliability, confidence intervals (CI), skewness (percentage ceiling and floor), and number of levels observed in a 2009 U.S. general population sample (see Chapter 14)—are summarized in Table 2.2 (see also Table 7.1). These results confirmed some of the theoretical advantages of the two component summary measures as compared to the eight health domain scales, including a very large increase in

the number of levels defined, smaller confidence intervals relative to each of the eight health domain scales, and the elimination of both floor and ceiling effects. A practical advantage is the reduction of the number of statistical comparisons required in an outcome study or clinical trial.

Very low scores on the PCS measure indicate limitations in physical functioning, limitations in role participation due to physical problems, a high degree of bodily pain, and/or poor general health. A very high score on PCS indicates little or no measured physical limitations, disabilities, or decrements in well-being; a high energy

Figure 2.1 SF-36v2 Measurement Model



Note. All health domain scales contribute to the scoring of both the Physical and Mental Component Summary measures. Scales contributing most to the scoring of the summary measures are indicated by a connecting solid line (—). Scales contributing to the scoring of the summary measures to a lesser degree are indicated by a dotted line (.....).

level; and/or good general health. For the MCS measure, a very low score is indicative of frequent psychological distress, social and role disability due to emotional problems, and/or poor general health. A very high score on MCS indicates frequent positive affect, little or no psychological distress or limitations in usual social/role activities due to emotional problems, and/or good general health. A strength of the PCS and MCS measures is their value in distinguishing a physical health outcome from a mental health outcome (Ware & Kosinski, 2001a; Ware, Kosinski, Bayliss, et al., 1995).

Table 2.2

Comparison of Features of SF-36v2 Health Domain Scales and Component Summary Measures Based on 2009 U.S. General Population Data

	Standard Form			Acute Form		
	PCS	MCS	Scales*	PCS	MCS	Scales*
Reliability	.96	.93	.82–.96	.97	.93	.81–.96
95% CI value (\pm)	3.9	5.3	3.9–8.3	3.5	5.4	3.9–8.5
% Floor	0	0	0.2–2.4	0	0	0.1–1.8
% Ceiling	0	0	2.0–61.2	0	0	2.6–67.5
Observed levels	486	494	8–21	486	494	8–21

* Statistics are presented as the range of results found across the eight SF-36v2 health domain scales (standard and acute forms) in the 2009 U.S. general population.

Profile of Scores

The SF-36v2 was constructed to achieve at least the minimum standards of precision necessary for group comparisons in eight conceptual areas. It was also constructed to yield a profile of scores that would be useful in understanding population differences in physical and mental health status, the health burden of chronic diseases and other medical conditions, and treatment effects on general health status. Figure 2.2 illustrates the survey's profile of scores and calls attention to important features of the two component summary measures and the eight health domain scales in this regard.

Unlike previous presentations of the profile, the SF-36v2 profile now begins with a presentation of the results of the PCS and MCS measures. The recent incorporation of these measures at the beginning of the standard Short Form survey profile (including profiles for the SF-12v2 and SF-8) emphasizes the importance of considering the findings from these more general measures of health status in the interpretation of results from any of the surveys in the SF family of instruments. It also facilitates interpretation by immediately establishing what the general burden of illness or effects of treatment are (i.e., physical or mental) before examining the more specific health domain scales. The PCS and

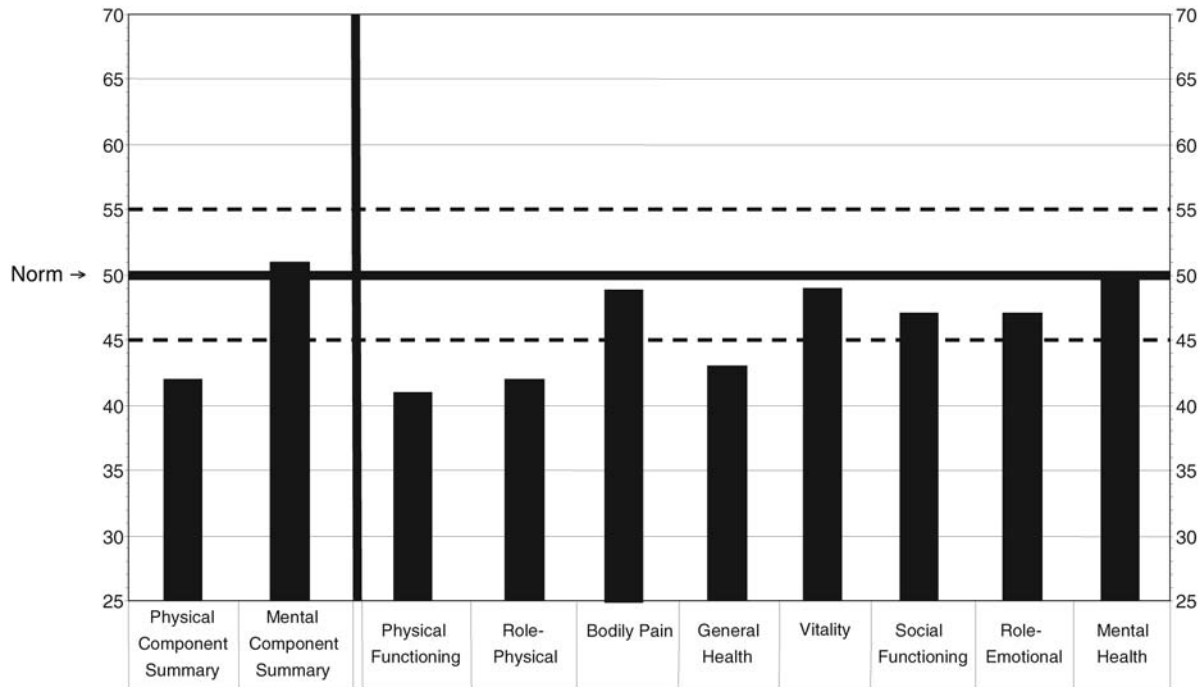
MCS scores provide, as their labels suggest, a summary of the respondent's health status from both a broad physical health perspective and a broad mental health perspective, respectively. Results on the PCS and MCS measures should serve as a starting place for determining whether functional limitations exist in either of the two major components of health; if so, the health domains contributing greatest to the affected dimension(s) and the items they comprise require further examination to ascertain their potential contribution to the respondent's impaired functioning. This drill-down approach to profile interpretation is discussed in detail in Chapter 7.

It is important to note that the eight health domain scales are ordered, left to right, from the best physical health measure (PF) to the best mental health measure (MH). The empirical evidence for this ordering of the scales is discussed in Chapter 13. This ordering further facilitates interpretation of the profile, with domain scales on the left side of the health domain profile reflecting physical health status and the domain scales on the right side reflecting mental health status.

In reviewing an SF-36v2 profile, users should be aware of an important feature of the range of measurement for each of the eight health domain scales. Five scales (PF, RP, BP, SF, and RE) define health status as the *absence of limitations or disability*. For these scales, the highest possible score is achieved when no limitations or disabilities are observed. Three of the scales (GH, VT, and MH) are bipolar in nature, measuring a much wider range of *positive and negative health states*. For these scales, a mid-range score is earned when respondents report no limitations or disability. A high score on these bipolar scales is earned only when respondents report positive states *and* favorably evaluate their health.

SF-6D Health Utility Index

Although not originally designed for use in economic evaluations, research has shown that a meaningful health state classification can be created by applying a scoring method that focuses on 11 items chosen from seven of the health domains covered by both versions of the SF-36 (Physical Functioning, Role Limitation [combined Physical and Emotional], Social Functioning, Bodily Pain, Mental Health, and Vitality). The resulting SF-6D (Brazier, Roberts, & Deverill, 2002; Brazier, Usherwood, Harper, & Thomas, 1998) is the first preference-based index constructed from a psychometric measure of health status. Scored from 0.0 (worst measured health state) to 1.0 (best measured health state), it uses a six-domain classification of health states—totaling 18,000 states in all—and can be used in the determination of the cost-effectiveness of various health care interventions

Figure 2.2 Sample SF-36v2 Profile of Scores

Note. The dashed lines (---) indicate the upper (55) and lower (45) bounds of *T* scores considered to be in the average range of functioning for individual respondents.

and *quality adjusted life years* (QALYs). Currently, the SF-6D is the only Short Form measure that provides a description of health and an economic evaluation. Hammer, Lawrence, Anderson, Kaplan, and Fryback (2006) published SF-6D age- and sex-stratified mean values and confidence intervals for the noninstitutionalized U.S. adult population based on SF-12 results ($N = 22,523$) from the 2001 Medical Expenditures Panel Survey (MEPS). The development of the SF-6D is briefly described in Chapter 13 of this manual.

Applications

The Short Form family of instruments is widely recognized as being among the leading patient-reported outcomes (PRO) measures. Used in studies, many reporting clinical trial results, documented in over 17,000 published articles as of July 2011, their reliability and validity in assessing the burden of disease and the effects of treatment have been demonstrated for patients with many different conditions. Translated or adapted into more than 140 languages, the Short Form surveys represent an international benchmark for health outcomes measurement and have been used as efficacy endpoints in clinical trials.

Until the early 1990s, most clinical trials, disease management programs, population monitoring efforts,

and health research studies defined the results or outcomes of interest relatively narrowly; that is, in terms of clinical variables. When patient-reported outcomes were considered, definitions tended to focus on disease-specific indicators. Increasingly, the variety of uses and users of patient-reported health assessments has expanded the definition of outcomes to include measures of both generic and disease-specific concepts. Used together, generic and disease-specific health assessments provide a comprehensive definition of health in its multiple dimensions as experienced by the individual (see Chapter 1).

Like the other members of the Short Form family of instruments, the SF-36v2 can be used alone or in combination with disease-specific PRO measures in several ways and for several purposes. It is this utility, along with its brevity, normative data, and demonstrated psychometric grounding, that makes the SF-36v2 a valuable tool in both clinical and research settings.

Evaluating and Monitoring Individual Patients in Clinical Practice

Although primarily intended for use in population studies, the SF-36v2 has proven valuable to physicians and other health care providers as a means of evaluating and monitoring individuals seeking treatment for physical or mental health problems. Unlike standard means of

assessing health status (e.g., physician examination, lab tests, mental status examinations), the survey provides a broad overview of a patient's health status and its effect on his or her functioning. Its incorporation into a standard office procedure is facilitated by the fact that it is a brief measure of patient self-report.

When administered at the beginning of an episode of care, the SF-36v2 can be used to help identify aspects of a respondent's health (e.g., functional impairment or distress) that may not otherwise be detected. The results of the initial administration can also serve as a baseline measure of health status that can then be compared to results obtained from one or more readministrations of the survey during the course of treatment, thus providing objective means of documenting the outcomes of said treatment. The results from one episode of care can also be used as comparison data for subsequent episodes of care. In addition, scores on the component summary measures can be used to roughly stratify patients according to who is more likely to utilize healthcare services (Ware & Kosinski, 2001a) or consume more health care dollars (Fleischman, Cohen, Manning, & Kosinski, 2006). Wetzler, Lum, and Bush (2000) provided a detailed discussion of the use of the SF-36 in primary care settings for various decision-making purposes related to patients presenting with possible behavioral healthcare problems. Meyer et al. (1994) reported results for individual hemodialysis patients that illustrated the feasibility and usefulness of periodic health assessments, including administration of the SF-36, in managing patients during the progression from advanced renal failure to end-stage renal disease. Like its predecessor, the SF-36v2 can assist in determining the need for and/or the most appropriate intervention, developing specific treatment recommendations, and predicting treatment outcomes. Moreover, the MH scale (Berwick, Murphy, Goldman, Ware, Barsky, & Weinstein, 1991) and the MCS measure (Ware & Kosinski, 2001b) can be used as screening tools for depression. Case studies demonstrating the application of individual respondent SF-36/SF-36v2 results in day-to-day clinical practice can be found in Wetzler et al. (2000) and in Chapter 12 of this manual.

Results from SF-36v2 studies can also be used to determine whether one treatment option is likely to have a more significant impact on a respondent's health status or quality of life. For example, Perry et al. (2003) found that patients undergoing laparoscopic donor nephrectomy had significantly higher postoperative PF, BP, and RE scores than those undergoing mini-incision open donor nephrectomy. At the same time, both groups scored at or above the average age-matched national

norms. Camilleri-Brennan and Steele (2002) found no significant differences on any of the SF-36v2 health domain scales between patients with low rectal cancer with an anterior resection and those with an abdominoperineal resection. These and other findings led the investigators to conclude that there was no significant difference in quality of life between patients undergoing one or the other treatment. Lanman and Hopkins (2004) investigated changes in the quality of life of patients with cervical disc disease treated with an anterior cervical spine fusion combined with a bioabsorbable interbody spacer. They reported 3-month postoperative score increases for all SF-36v2 health domain scales except GH, with the greatest increases occurring on the SF scale (7.4 points), PF scale (5.7 points), and RE scale (4.3 points).

In addition, Ko et al. (2002) found no significant differences in SF-36 health domain scale or component summary measure scores for groups of patients with familial adenomatous polyposis who underwent either a permanent ileostomy or a procedure to restore bowel continuity. In another study, Russell, Conner-Spady, Mintz, Mallon, and Maksymowych (2003) demonstrated the responsiveness of the SF-36 and other measures to changes in two groups of patients with rheumatoid arthritis—one group considered stable and the other group having persistent and unacceptably high disease levels—beginning treatment with a drug (infliximab) previously shown to yield a good response. The SF-36 was found to be responsive to the infliximab patients' pain and global assessment after 14 weeks of treatment.

Monitoring Populations

Health plans, employers, and researchers are continually challenged to find efficient and comprehensive ways of measuring the health of various populations. The measures they use must be well understood and accepted. Moreover, these measures need to reflect multiple aspects of health over a wide range, permit comparisons within and across groups, and demonstrate sensitivity to changes in health over time. Ideally, such measures would meet all these requirements with as few items as possible, thereby minimizing respondent burden and data collection costs.

The SF-36v2's brevity lends itself to comprehensive population monitoring. As one of the leading measures of general health status, the effectiveness of it and other members of the Short Form family of instruments in monitoring functioning and well-being, assessing disease burden, and comparing the health of different populations and patient groups has been reported in a total of more than 17,000 publications as of July 2011. The survey's usefulness in assessing the burden of disease is

documented in these publications describing more than 150 diseases and conditions, with at least 16 conditions *each* being addressed in more than 100 publications. A prime example of how the Short Form surveys can be used in population monitoring is the yearly Medicare Health Outcomes Survey (HOS; Gandek, Sinclair, Kosinski & Ware, 2004; Ware, Gandek, Sinclair, & Kosinski, 2004). From 1998 to 2004, the HOS consisted of the SF-36 survey along with questions about activities of daily living (ADLs) and case-mix and risk-adjustment questions for Medicare beneficiaries enrolled in managed care programs. (Note that the SF-12v2 replaced the SF-36 in the HOS beginning in 2005.) All Medicare managed care plans must participate in the annual HOS, in which the MCS and PCS measures, along with mortality, are the primary outcomes measures used to assess enrollees' health.

Estimating the Burden of Disease

The SF-36v2 and other standardized assessment methods offer a number of advantages to care providers. For example, it can be used to obtain information about functioning and well-being directly from patients in a standardized manner. By standardizing questions, answers, and scoring, reliable and valid comparisons can be made to determine the relative burden of different conditions in several domains of health.

The value of general and specific population norms, which was well demonstrated for the Sickness Impact Profile (SIP; Bergner, Bobbitt, Carter, & Gilson, 1981), later for the MOS SF-20 (Stewart, Hays, & Ware, 1988; Stewart et al., 1989), and for other measures as well, has also been demonstrated for the SF-36 and its revised version. Whereas some of the initial descriptive studies using the SF-36 were performed primarily to validate scale scores (McHorney et al., 1992), the Short Form survey scales appear to be increasingly accepted as valid health measures for the purposes of documenting disease burden. Disease-specific benchmarks, developed from the disease or physically impaired subsamples of the 2009 SF-36v2 normative group, provide estimates of the burden of disease for each of 40 disease or condition groups on each of the SF-36v2 scales and measures and are available from QualityMetric or its authorized resellers.

As previously mentioned, for each of at least 16 conditions, there are at least 100 articles that have been published on the burden of illness as measured by the Short Form family of instruments. Recent SF-36v2 articles reporting the burden of a disease/condition or its treatment include those for anterior cruciate ligament injury (Ochiai, Hagino, Tonotsuka, & Haro,

2010), systemic sclerosis (Hudson et al., 2009), cancer survivors (Greenfield et al., 2010), subclinical hypothyroidism (Razvi, Ingoe, McMillan, & Weaver, 2005), lung transplant recipients (Girard et al., 2006), multiple sclerosis (Forbes, While, Mathes, & Griffiths, 2006), inguinal hernia (Fitzgibbons et al., 2006), sacroiliac syndrome (Cheng & Ferrante, 2006), colorectal cancer survivors with intestinal ostomies (Baldwin et al., 2009), non-Hodgkin lymphoma and breast cancer survivors (Crespi, Smith, Petersen, Zimmerman, & Ganz, 2010), fibromyalgia (Bennett et al., 2005), low back pain (Kosinski et al., 2005), patients with asthma and/or COPD (Abramson et al., 2010), and shoulder pain in diabetics (Laslett, Burnet, Jones, Redmond, & McNeil, 2007).

Evaluating Treatment Effects in Clinical Trials

As people live longer, healthcare focuses less on mortality than on improving how people feel and function, often in the face of multiple chronic diseases or conditions. Many drugs in the discovery and development pipeline hold the promise of reducing the impact of chronic health problems on everyday life. Medical researchers conducting clinical trials now recognize the need to define benefits more broadly than traditional clinical endpoints allow by including PROs in clinical trials. Additional clinical evidence based on PROs also commands increasing attention from the FDA, making it critical to the drug review and approval process. The FDA and the National Institutes of Health (NIH) have launched an effort to encourage the use of PROs, standardize their assessment, and, when warranted, grant indications for drugs based on patient-reported evidence of improved functioning and well-being.

Given the high costs associated with drug development and testing, clinical trials depend on reliable and scientifically valid health outcomes measurements that are acknowledged and accepted by the FDA. In turn, the FDA has issued guidelines for the use of PRO measures in medical product development to support labeling claims (U.S. Department of Health and Human Services, 2009). Conversely, clinical trials that meet regulatory roadblocks due to insufficient data can incur costly corrective actions. Pharmaceutical companies thus require the use of well-validated, documented, and accepted PRO measures that can capture, with high degrees of reliability and sensitivity, differences between alternative drugs, drugs versus placebos, and drug dosages over relatively short periods of time.

The SF-36 and SF-36v2 are becoming widely recognized as leading PRO measures in clinical trials. They, along with other members of the SF family of instruments, have been cited in a total of more than

2,000 published articles as of January 2011 reporting randomized controlled trial results. Their reliability and validity in assessing the burden of disease and the effects of treatment have been demonstrated for patients with many different conditions. With more than 140 translations and adaptations available, the SF-36 and SF-36v2 represent international benchmarks for health outcomes measurement and have been used as efficacy endpoints in clinical trials.

When included in a clinical trial protocol, the SF instruments can quantify a respondent's experience of improved health-related quality of life (HRQOL), deliver proof of efficacy that goes beyond traditional clinical endpoints, and provide a scientifically valid body of evidence to facilitate timely regulatory approval. For example, Nicholson, Ross, Sasaki, and Weil (2006) included SF-36v2 PCS and MCS scores as endpoints in their Phase IV prospective, randomized trial comparing the efficacy, tolerability, and safety of polymer-coated extended-release morphine sulfate (P-ERMS) and controlled-release oxycodone hydrochloride (CRO) in the treatment of patients with moderate to severe nonmalignant pain. Comparison of baseline and 24-week scores revealed significant change ($p < .05$) in PCS for both treatment groups, whereas only the CRO group showed significant 24-week change ($p < .05$) on the MCS measure.

Fitzgibbons et al. (2006) included 2-year change in the SF-36v2 PCS score as one of their primary outcomes in a study of men with inguinal hernia undergoing either standard open tension-free repair with mesh ($n = 356$) or "watchful waiting" ($n = 364$). A total of 317 and 336 of the respondents, respectively, completed the 2-year follow-up assessment, which demonstrated that the two groups did not significantly differ in amount of baseline-to-follow-up change on the PCS measure.

Strand et al. (1999) used the SF-36 to assess improvement in function and HRQOL in patients with rheumatoid arthritis assigned to a leflunomide, methotrexate, or a placebo treatment group for 12 months. The baseline scores were found to be significantly lower than the U.S. norms (0–100 scale). Substantial improvement on PCS, PF, BP, GH, VT, and SF were noted for the leflunomide group, with the PCS change being significantly greater than that found for the methotrexate and placebo groups. The leflunomide group also had a greater percentage of respondents showing two levels of improvement (20% and 50%) on this same measure. In a randomized, open-label, 1-year trial, Raynauld et al. (2002) found that SF-36 PCS scores increased significantly ($p < .0001$) at 12-months postbaseline for a group of 127 patients with knee osteoarthritis receiving

appropriate care in addition to an injection of hylan G-F 20 (a viscosupplementation product) each of the first 3 weeks of the study. No significant change was noted on the SF-36 or any of the other quality-of-life measures used for a control group of 128 patients.

Among some of the more recently published clinical trials that employed the SF-36v2 are a study investigating the effect of active resistive exercise on breast cancer-related lymphedema (Kim, Sim, Jeong, & Kim, 2010); an automated, interactive telephone intervention to improve type 2 diabetes self-management (Bird et al., 2010); and the effect of acoustic cueing on the quality of life of people with moderate to severe Parkinson's disease (Elston, Honan, Powell, Gormley, & Stein, 2010).

In addition, the SF-36v2 can be delivered in fixed-length formats for self- or interviewer-administration, by way of paper-and-pencil forms, smartphones, tablets, and online (see Chapter 4). The ability to choose from among several administration options is another feature that enables the survey to meet the needs of clinical trials that require practical and precise measures for risk screening and sensitive, patient-reported measures of outcomes. Further research is underway to evaluate the comparability of scores across administration modalities.

Disease Management

Health plan providers and others concerned with disease management face significant measurement challenges. To control costs without harm to health, they must have patient-specific information that predicts risk, identifies healthcare needs, and quantifies the outcomes that matter most to patients. Leaders in disease management recognize that no single metric meets all these requirements. While medical claims data provide a convenient and retrospective view of utilization that contributes to broad-based program planning, they offer little to help understand the impact of disease on a patient's physical and mental health or identify who is likely to benefit most from disease management strategies. Rather than relying solely on claims data, experts recommend an integrated measurement strategy that combines data from multiple sources, including the patient. Increasingly, disease management providers are incorporating PRO surveys into their measurement systems. Data from such surveys add significant value because they improve risk prediction, service planning, and outcomes monitoring efforts, as well as ensuring that program planning and evaluation efforts incorporate the patient's perspective.

The SF-36v2 can provide practical solutions to disease management's most pressing measurement challenges. Its reliability and validity in assessing the bur-

den of disease has been demonstrated for several patient populations. Many studies document its ability to predict hospitalization, total medical expenditures, job loss and work productivity, future health, risk of depression, use of mental health care, and mortality. For example, Haffer, Bowen, Shannon, and Fowler (2003) used the SF-36 to assess participants with one or more of several chronic conditions in the Medicare HOS at baseline and again 2 years later to demonstrate the need for disease management programs for chronically ill Medicare enrollees. Sidorov, Shull, Girolami, and Mensch (2003) measured the impact of a disease management program on the quality of life of a group of congestive heart failure (CHF) patients using the SF-36. In a broader study, Walker, Landis, Stern, and Vance (2003) used PCS and MCS measures derived from the SF-36v2 and SF-8 to demonstrate changes in the quality of life of large samples of patients with coronary artery disease, chronic obstructive pulmonary disease, and heart failure who were involved in disease management programs.

In addition, disease-specific surveys can be paired with the SF-36v2 to capture a more comprehensive picture of HRQOL. When used with one or more disease-specific measures, it provides information necessary to screen patients with common chronic conditions—such as asthma, congestive heart failure, diabetes, migraine headaches, and osteoarthritis—and to monitor and compare their outcomes over time.

Risk Prediction and Cost-Effectiveness

Health plans, disease management programs, and employers use predictive models to forecast health expenditures and to identify those who may benefit from proactive health interventions and prevention programs. Most predictive models rely on laboratory and claims data. These models often fail to identify many at-risk individuals because they miss previously healthy patients who have developed serious conditions and stable patients whose conditions are beginning to worsen. Such models may underestimate future expenditures and miss individuals in early stages of disease or illness episodes that could benefit from disease management interventions leading to reduced downstream complications and costs.

Increasingly, health planners are recognizing that when generic HRQOL data from patients' self-assessments of physical and mental health are added to predictive models, their predictive power substantially improves, yielding information that helps providers better anticipate and manage health problems. The SF-36v2 is among the best validated and widely used HRQOL measures available today. It can be used as a

baseline in risk stratification and, when repeated over time, for health outcomes monitoring. Because it can be completed quickly, health plan administrators and other users can collect self-reported health assessment data efficiently and inexpensively. Moreover, because it measures both physical and mental health over a very wide range, it can be used for risk prediction with any population.

Including the SF-36v2 scales and measures in predictive models can improve forecasts of future expenditures, resource utilization, health outcomes, likelihood of hospitalization, risk of depression, use of mental health specialty care, job loss, return to work and work productivity, future health, and mortality. For example, Hornbrook and Goodman (1995) found results from the SF-36 PF, RP, and GH scales and the Reported Health Transition (HT) item, now called the Self-Evaluated Transition (SET) item, to be better predictors of total annual health care expenditures for a large sample of HMO subscribers than demographic and clinical variables (e.g., age, existing condition) alone. Thus, using the SF-36v2 in baseline health assessments can help to more accurately quantify patients' healthcare needs and develop effective care plans. Administering it at selected intervals, such as before and after a disease management intervention, will allow the user to quantify physical and mental health outcomes and to evaluate the effectiveness of interventions.

Patient-Provider Relations

Containing costs is one of the biggest challenges facing health care providers today. As employers have begun to shift a greater portion of health care costs to employees and their families, interest in consumer-driven health care has markedly increased. As a result, consumers are taking greater control of their health care and becoming more actively engaged in making important treatment decisions. Health plans and disease management companies have responded by making every effort to keep members informed and educated.

This trend in health care consumerism is also giving rise to increased use of technology. As consumers search the Internet for medical information and data, online health care is gradually being personalized, with managed care organizations viewing their websites as core components of their businesses.

Another result of high health care costs can be seen in the amount of time that clinicians spend each day with patients. For example, Gottschalk and Flocke (2005) found the average face-to-face patient care time for a sample of family physicians was 10.7 minutes, with an additional 2.6 minutes being spent on visit-specific

work outside of the examination room. These results are far more conservative than those found in the 2003 National Ambulatory Medical Care Study (NAMCS; Hing, Cherry, & Woodwell, 2005). Overall, Gottschalk and Flocke's sample spent only 54.9% of the workday involved in actual face-to-face patient care, with additional visit-specific work outside of the examination room occupying 14.5% of their time and work related to other patients *not* being seen in the office at that time accounting for another 22.9%. Predictably, limited contact between patients and providers has only increased the need and demand for member-focused services that promote information flow and foster improved care delivery through consumer involvement.

When incorporated as part of the standard care process, the SF-36v2 can improve and enhance communication by providing information that enables health care providers to make the best use of the limited time they have to see patients. As previously discussed, the survey's results can be used to establish an objective baseline measure of health status against which health problems can be identified, effects of treatment monitored, and outcomes of that treatment quantitatively assessed. Moreover, employing an Internet-based method of administering the SF-36v2 can afford the busy provider the additional benefits of providing immediate feedback for members, a rich set of reporting facilities for the clinician, and aggregated survey results for groups of patients.

An example of how the survey can improve patient communication and management is provided by Wagner et al. (1997). The authors conducted a controlled study in which 163 consecutive epilepsy patients were administered the SF-36 during a prestudy assessment and then again prior to subsequent office visits, beginning within 6 months of the prestudy assessment and continuing for 6 months thereafter. During the follow-up visit, 126 of the study participants (70%) were randomly assigned to the intervention condition in which their physicians had access to their assessment results at the time of the encounter. The remaining 37 participants (30%) were assigned to a control condition in which their physicians did not have access to their SF-36 results. After each encounter, patients in both conditions completed a satisfaction questionnaire and, in the case of intervention patients, physicians completed a questionnaire regarding the usefulness of the SF-36 information during the encounter. Although the two groups of patients did not differ significantly in their attitudes toward or satisfaction with their care, the physicians reported that the survey results provided new information in 63% of the encounters, prompted change in therapy in 12%,

was useful for patient communication in 14%, and was useful in management in 8%. They also found that the worse the survey results, the greater utility of the SF-36 for patient communication and management.

The SF-36v2 also can be used to measure the effects of other attempts at improving communication between patients and their health care providers. Using a randomized crossover design, Detmar, Muller, Schornagel, Wever, and Aaronson (2002) studied the effects of providing HRQOL feedback to physicians and their oncology patients undergoing palliative care. For the purpose of this study, results from the patient self-administered Quality of Life Questionnaire-Core 30 (QLQ-C30 [version 3.0]; Fayers, Aaronson, Bjordal, Curran, & Groenvold, 1999), which was administered before each of four visits, were used. Among several variables investigated was change in patients' scores on the SF-36. The two cohorts of intervention patients and two cohorts of control patients did not differ significantly in health domain scores between the first and fourth visits; however, a significantly greater percentage of the intervention patients exhibited an improvement of 0.5 standard deviations or greater on both the MH scale (43% vs. 30%; $p = .04$) and the RE scale (22% vs. 11%; $p = .05$), suggesting positive emotional effects were brought about as a result of the intervention.

Direct-to-Consumer Information

As pharmaceutical companies shift from marketing their drugs and devices to physicians toward broader-based efforts to position their products as solutions to health problems, they increasingly engage in providing information directly to potential consumers. Direct-to-consumer (DTC) information comes in many forms, from marketing/advertising to outreach/educational campaigns.

As indicated in surveys by the FDA and the National Consumer League, companies provide DTC information to achieve a number of objectives. In addition to promoting specific products, these objectives include educating the public about medical conditions, their symptoms and effects, and potential treatment options; prompting recognition or detection of personal health problems that may benefit from clinical consultation, thereby encouraging more appropriate care-seeking, case-finding, and physician-patient dialogue; and promoting self-care and compliance with treatment regimens. At the same time, consumers are actively searching for relevant information to help them understand health problems, recognize risks and side effects, communicate better with their clinicians, and participate in managing symptoms and treatments.

Critical to the success of DTC information campaigns is consumer recognition that the information provided has immediate relevance to them. Increasingly, DTC materials include short, self-report health assessments, the results from which link directly to guidelines regarding likelihood of diagnosis and/or recommended self-care, physician consultation, and treatment options. To be most effective, such assessments should meet scientific standards of reliability and validity and have demonstrated acceptance and relevance among consumers and clinicians. A recent example is the promotion of the Asthma Control Test™ (ACT™; Kosinski, Bayliss, Turner-Bowker, & Fortin, 2004) as part of a popular media (e.g., newspapers, magazines, television, Internet) advertising campaign for an asthma medication. In these ads, asthma patients are encouraged to complete the five-item questionnaire about how well their asthma is being controlled and to discuss the results with their health care provider.

When health assessments meet measurement standards and are selected or developed with their planned use in mind, benefiting populations can be identified, key data can be collected, and recommendations can be provided, all with a solid return on investment. Those employing the SF-36v2 as part of a DTC assessment have the added benefit of being able to administer the survey in fixed-form format either in print, by smartphone or tablet, or online via the Internet. Also, one or more disease-specific measures can be administered along with the survey to provide consumers and their clinicians with the information required to screen and monitor common chronic conditions such as asthma, congestive heart failure, or depression.

In addition, as pharmaceutical companies strive to cost-effectively target specific consumer populations, health assessments delivered online can help to identify potential users/consumers and better match treatments to their needs. For DTC campaigns relying chiefly on Internet-based material, the results of their efforts may be maximized by making online administration available to potential customers, such as in the case of the availability of the ACT via the Internet. Overall, using the SF-36v2 as part of a DTC marketing effort can help garner consumer acceptance by providing a first-stage screen for conditions having substantial impact on generic domains and, when used longitudinally, gathering proof of improved outcomes.

Survey Validation

Because of their solid psychometric foundations and frequent incorporation into studies published in peer-reviewed journals, the SF-36v2 and the other Short

Form surveys are considered by many to be the “gold standard” of HRQOL surveys. As such, the Short Form component summary measures and health domain scales are often used as criteria for validating new or existing disease-specific and generic HRQOL measures. For example, Hawthorne, Kaye, Gruen, Houseman, and Bauer (2011) used correlations with the SF-36v2 PCS and MCS measures as means to support the construct validity of scales from the Quality of Life after Brain Injury measure. Also, Gersh, Arnold, and Gibson (2011) used the SF-36v2 RP and MH scales to measure disability and mood disturbance, respectively, in a study investigating the utility of the Pain Stages of Change Questionnaire (PSOCQ) to assess treatment completion and to determine if PSOCQ scores correlate with clinical outcomes with a group of chronic pain patients. Yoshida et al. (2011) used the SF-36v2 as a criterion measure in their validation study of the Brief Scale for Psychiatric Problems in Orthopaedic Patients assessment. In another example, Hirsch et al. (2008) used the SF-36v2 to evaluate the validity of the Gout Impact section of the Gout Assessment Questionnaire.

A Final Comment on Applications

Debate about the uses of health outcomes assessment methods is spreading beyond the arcane realm of methodologists (Maruish, 2002, 2004a; Ogles, Lambert, & Fields, 2002; Ware, 1990b, 1993). Policy analysts and health care managers—intent on getting the best value for their dollars—have joined the intellectual fray. Clinical investigators evaluating new treatments and technologies, as well as practicing clinicians seeking better patient outcomes, are also demanding useful assessment methods.

Despite advances in measurement tools, the current state of health care monitoring is woefully deficient. To wit, national health surveys, management information systems used by health care delivery organizations, databases analyzed in most clinical trials, and inpatient or outpatient medical records do not include comprehensive health assessments. However, federal health agencies are increasingly recognizing the importance of standardizing the content of tools to measure health concepts and are coordinating their efforts in this regard. One such example is the Patient-Reported Outcomes Measurement Information System (PROMIS) project, supported by the National Institutes of Health (NIH). This trans-NIH initiative involves a cooperative network of six primary research sites and a statistical coordinating center whose goal is to help define the next generation of health

outcomes measurement by improving upon existing measures through better psychometrics, CAT software, and the use of the Internet for alternative connections and standardized scoring (see <http://commonfund.nih.gov/promis> for more information).

To meet the needs of the 21st century patient, information about general health outcomes must be added to the nation's healthcare database. Minimum standards of comprehensiveness should be adopted to monitor the health of the general population and to evaluate health care policies. A core set of measures assessing generic

health outcomes should be standardized and adopted to compare the relative burden of medical and psychiatric conditions and relative treatment benefits. It is now practical to include a standardized core set of general health measures across applications (e.g., general population surveys, clinical trials) while supplementing this core according to the particular needs of a given study. The resulting comparison data would greatly advance the understanding of health measure interpretation for all applications. Adoption of a standardized core set of health measures should be a high priority.

3

The Short Form Family of Health Survey Instruments

The “developmental” version of the SF-36, published in 1988 (Ware), represented a significant advance in the short-form instrumentation available to measure the self-reported health status of patient and nonpatient populations. Since that time, one revised version of the SF-36 and three abbreviated Short Form surveys have been made available. The SF-36v2, the most up-to-date fixed-form version of the SF-36 that is currently available, incorporates the use of more comprehensive normative data with the knowledge and advancements gained from over a decade of applications in research and clinical settings. It is recommended for all new studies requiring one of the two 36-item measures. However, all members of the Short Form family of instruments for adults—the SF-8, SF-12, SF-12v2, SF-36, SF-36v2, and DYNHA Computerized Adaptive Health Assessments—are cross-calibrated and scored on the same norm-based *T*-score metric to maximize their comparability and all have demonstrated their usefulness in assessing health status. Note that the SF-10™ Health Survey for Children is also a member of the Short Form family but is not calibrated with the adult surveys.

Although the original SF-36 and SF-12 (which is comprised of a subset of SF-36 items) proved to be useful for many purposes, years of experience revealed the potential for improvements. The need to improve item wording and response choices, demonstrated by the International Quality of Life Assessment (IQOLA) Project (see Chapter 1) and the translation of the SF-36 forms, and the opportunity to update normative data led to the revision and norming of the new SF-36 survey—the SF-36v2—in 1998. The SF-36v2 was re-normed in 2009, providing more current U.S. general population comparison data than are available for the SF-36. Because the SF-36v2 is now considered superior to the original instrument for the aforementioned reasons, QualityMetric Incorporated has discontinued the licensing of data collection and scoring services for

the original SF-36 and SF-12 forms, including the sale of supporting materials for these surveys.

The purpose of this chapter is to provide a broad overview and comparison of the SF-36v2 with regard to the SF-12v2 and the SF-8. The following sections describe the features of each survey and discuss considerations for deciding which one to use. In addition, this chapter discusses the general considerations for matching a Short Form survey to an application and provides a direct, survey-to-survey comparison summary for application-matching purposes. Finally, the efforts to translate the Short Form surveys for multinational use are discussed briefly.

The Short Form Instruments

There are many commonalities among the members of the Short Form family of instruments. A brief description of each of the five available surveys is presented in the following sections.

The SF-36v2 Health Survey

Based on the SF-36, the SF-36v2 (Ware, 2000, 2004; Ware & Kosinski, 1996; Ware et al., 2007; Ware, Kosinski, & Dewey, 2000) contains 36 items used to measure eight domains of health-related quality of life (HRQOL): Physical Functioning, Role-Physical (i.e., role limitations due to physical health), Bodily Pain, General Health, Vitality, Social Functioning, Role-Emotional (i.e., role limitations due to mental/emotional health), and Mental Health. The information obtained from these eight health domains can be further aggregated into the Physical Component Summary (PCS) measure and the Mental Component Summary (MCS) measure. Data from the survey have proven its usefulness in measuring health status and outcomes in both general and specific populations. Information about the SF-36v2, including citations for the most recently published studies and the

developers' responses to frequently asked questions, are available online at <http://www.sf-36.org>.

As previously mentioned, the SF-36v2 offers significant improvements in the measurement of HRQOL compared to the SF-36. These advances include:

- Improved instructions and questionnaire items, designed to simplify the wording and make the language more familiar.
- Improved layout for questions and response choices, making them easier to read and complete, thereby reducing the frequency of missing responses.
- Greater comparability with the widely used translations and cultural adaptations.
- Five-level response choices, replacing *yes/no* response choices, for items in the Role-Physical and Role-Emotional health domain scales, extending the range of functioning measured and increasing score precision.
- Five-level response choices, replacing six-level response categories, designed to eliminate the ambiguous response choice (*A good bit of the time*) in the Mental Health and Vitality health domain scales.
- Norm-based scoring, in the form of *T* scores, for the health domain scales. Note that the component summary measures have always been scored using *T* scores.
- Up-to-date 2009 *T*-score norms for both the standard (4-week) and acute (1-week) forms.

These improvements are discussed in detail in Chapter 13 of this manual.

The SF-12v2 Health Survey

Based on the SF-12, the SF-12v2 (Ware et al., 2010) offers significant advantages in the measurement of health status. Its 12 items were taken directly from the SF-36v2; as a result, the improvements found in the SF-12v2 are similar to those made to the SF-36v2. In addition to the substantial gains in the range and precision of measurement achieved in comparison with the SF-12, the eight health domain scales can be scored on the SF-12v2 as well. Thus, it has proved to be a viable alternative to the SF-36v2 for those seeking a very brief but comprehensive measure of health status. Detailed information about the development of the SF-12v2 can be found in Ware et al.

The SF-8 Health Survey

The SF-8 (Ware, Kosinski, Dewey, & Gandek, 2001) contains 8 items, only one of which is identical to any of

the items in the SF-36v2. Although the SF-8 items are not a direct subset of SF-36v2 items, both the SF-8 and the SF-36v2 measure the same eight health domains. Whereas the SF-36v2 uses between 2 and 10 items to measure each health domain, the SF-8 uses just one item for each health domain, making it less burdensome to complete and a good alternative to the SF-36v2 and the SF-12v2 for large-scale population survey efforts. Similar to the SF-36v2 and the SF-12v2, the PCS and MCS measures can be calculated from SF-8 results. The one disadvantage is that its scores generally cover a narrower range of the measured constructs, are more coarse (i.e., define fewer levels) for some scales, and are less precise. Therefore, the SF-8 is not the Short Form survey of choice when one is interested in respondent-level interpretations of scores, in conducting studies with smaller sample sizes where enhanced precision is especially important, or in performing investigations requiring more statistical power.

Computerized Adaptive Testing (CAT) and the DYNHA Computerized Adaptive Health Assessments

For the most demanding applications of health status surveys, brief fixed-form tools are no longer the most efficient, practical, or precise measures available. Ongoing research is demonstrating that software based on computerized adaptive testing (CAT) logic delivers the best of both worlds: increasingly practical and precise measures that cover the very wide range of levels of health and well-being required to monitor and compare generic health outcomes across diverse populations, all while being administered with only the minimum of necessary items. By matching questions to each respondent's health level, CAT can also estimate scores much more efficiently than fixed-form surveys.

The core general health measures in QualityMetric's DYNHA software are based on the Short Form family of instruments. This software uses item response theory (IRT) models to calibrate item pools (using items taken from the SF-36v2 and other widely used questionnaires) and to select the best items for each respondent, items that are then scored using the same *T* scores as the SF-36v2. The resulting CAT survey scores are quite accurate over a very wide range of measurement. This approach to survey administration offers efficiency, comparability of results using *T*-score norms, and availability of interpretation guidelines based on the Short Form surveys.

A prototype of computerized dynamic health assessments is available online at <http://www.amIhealthy.com>.

The SF-10 Health Survey for Children

The SF-10 Health Survey for Children (Saris-Baglana et al., 2007) is a 10-item, parent-completed Short Form survey designed to measure the physical and psychosocial functioning of children aged 5 through 17 years. This survey was designed to be an alternative to the short-form Child Health Questionnaire (CHQ™; Landgraf, Abetz, & Ware, 1999). The CHQ was developed in the early 1990s from the findings of the Child Health Assessment Project at Tuft's New England Medical Center's Health Institute and in response to the need for a comprehensive generic measure of functional health status and well-being in children and adolescents.

Much like the SF-12v2 and SF-8, the SF-10 instrument was developed to be brief, reliable, and valid, yet still comprehensive in its coverage of content areas relevant to children's physical and psychosocial functioning and well-being. Specifically, the developers' objective was to reproduce the CHQ's Physical Summary (PhS) and Psychosocial Summary (PsS) scores (referred to as PHS-10 and PSS-10, respectively, in the SF-10) using only one or two items from eight of the 10 domains represented. As previously mentioned, the SF-10 was developed as an alternate form to the CHQ that would enable the reproduction of the PhS and PsS scores of the 50-item CHQ using significantly fewer items.

A brief instrument like the SF-10 offers many advantages for practical application; however, it is not as precise as the longer-form CHQ and generally covers a narrower range for each of the construct areas assessed. The SF-10 is intended for use in population-based studies, in studies involving large sample sizes, and in group-level comparisons where precision is less of a concern due to large sample sizes. Short-form measures like the SF-10 work well in large studies because precision and the statistical power of hypothesis testing are achieved more by utilizing a larger representative sample than by increasing measurement reliability through the administration of many items. When used in population studies, the SF-10 yields results that are comparable to those that can be obtained with the longer-form CHQ.

Deciding Which Short Form Survey to Use

Choosing among the forms and versions of the SF family of health survey instruments depends on the requirements of the intended application, among other considerations. Score interpretation and the need for norms are not major considerations because the underlying

metrics (i.e., *T* scores) used in the scoring of all the Short Form surveys have been standardized across the summary measures. In most cases, choosing a survey involves a tradeoff between precision and respondent burden and whether Internet-based dynamic administrations are possible. The following sections discuss considerations for selecting a survey, focusing on the Short Form instruments developed for use with adults.

Features of the Short Form Surveys

Content. All of the adult Short Form surveys measure the same eight health domains: Physical Functioning (PF), Role-Physical (RP), Bodily Pain (BP), General Health (GH), Vitality (VT), Social Functioning (SF), Role-Emotional (RE), and Mental Health (MH). Because more items permit better representation of each health domain, the domains are best represented in the SF-36v2, followed by the SF-12v2, and then the SF-8. The SF-36v2 and SF-12v2 have 12 items in common, whereas the SF-8 has only one item in common with the SF-36v2 and no items in common with the SF-12v2. Content is very similar across all the surveys, however, and measures of corresponding concepts achieve a very high correlation across all forms. Finally, the SF-8, SF-12v2, and SF-36v2 all yield scores for the eight health domains and the two component summary measures (PCS and MCS).

Recall period. In each survey, most items ask respondents to consider a specific period of time, or *recall period*, when responding. Both the SF-36v2 and SF-12v2 are available in two forms, each covering a specific recall period. The *standard*, or 4-week recall, form asks the respondent to answer the Short Form questions as they pertain to the way he or she felt or acted *during the past 4 weeks*. The *acute*, or 1-week recall, form asks the respondent to answer the Short Form questions as they pertain to the way he or she felt or acted *during the past week*. The SF-8 is available in three validated forms, each with a differing recall period: a standard form (4-week recall), an acute form (1-week recall), and a second acute form (24-hour recall; Ware, Kosinski, Dewey, & Gandek, 2001).

The standard 4-week recall period was adopted for the Short Form surveys to maintain comparability with the long-form Medical Outcome Study (MOS) measures from which it was derived. The 4-week recall period was adopted for the MOS long-form measures because it was thought that focusing on the previous 4 weeks would capture a more representative and reproducible sample of recent health, not unduly affected by daily or momentary fluctuations (Fowler, 1984; Stewart & Ware, 1992). Use of the SF-36v2's standard (4-week

recall) form is appropriate when the instrument will be administered only once to the respondent or when at least 4 weeks will pass between readministrations. In most cases, the standard form will meet a clinician's needs concerning patient monitoring and a researcher's needs regarding many types of investigations, particularly those of a longitudinal nature. However, there are many instances in which a 4-week recall period is not appropriate, particularly in studies that require relatively short intervals between follow-up assessments because changes in health status occur more rapidly.

The acute form of the SF surveys was designed for applications in which health status would be measured weekly or biweekly. To create the acute form, the recall period for six SF scales (RP, BP, VT, SF, RE, and MH) was simply changed from "the past 4 weeks" to "the past week." For example, the question, "During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?" was changed to, "During the past week, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting friends, relatives, etc.)?" Two scales, PF and GH, do not have a recall period, so they are identical across acute and standard forms. The acute (1-week recall) form provides a better description of a respondent's health status during the most recent week than the standard form. Also, when more frequent readministrations are required, the acute form is most appropriate. For example, the acute form is recommended when a clinician or researcher wants to closely monitor the effects of a physical (e.g., pharmacological) or behavioral (e.g., psychotherapeutic) intervention on a patient or group of patients when such effects are likely to occur rapidly (e.g., asthma therapy). However, at least 1 week must pass between acute form administrations in order to obtain valid information.

Generally, the results from administrations of the standard and acute forms substantially agree. However, users may find that results from the acute form differ from those obtained from the standard form. For example, Keller et al. (1997) found that the effect of the form did approach significance ($p = .08$) with two small samples of asthma patients participating in a controlled study of the effects of inhaled corticosteroid on HRQOL. In addition, univariate analyses revealed more favorable results (i.e., higher scores on the 0–100 scoring metric) using the acute form, with RE averaging nearly 7 points higher ($p = .05$), RP averaging nearly 5 points higher, and SF averaging nearly 3 points higher. It is important to note, however, that this study was conducted within the context of a randomized clinical trial where changes

in health status can occur relatively quickly; therefore, these results still need to be replicated with other acutely ill patient samples. Also be aware that the Keller et al. findings could not be replicated using data from the 1998 SF-36v2 normative sample, which found that health domain scale scores from the standard and acute forms were very similar.

The 24-hour recall version of the SF-8 was developed to increase the survey's responsiveness to very acute changes in health status, such as those that may occur within 2 to 3 days. Thus, it is an SF solution for situations requiring group-level health status assessment more frequently than once a week.

Respondent burden. Shorter surveys can be completed more quickly and require less space in printed questionnaires. On average, the SF-8 can be completed in 1 to 2 minutes, the SF-12v2 in 2 to 3 minutes, and the SF-36v2 in 5 to 10 minutes. Survey length and respondent burden may be an issue in some clinical settings or when a survey is administered as part of a large battery of instruments. Consequently, the SF-12v2 quickly became the tool of choice among fixed-form population surveys because its RP and RE health domain scales cover wider ranges of health levels more accurately with *fewer* items than their three- and four-item counterparts on the SF-36v2. This improvement in precision, in conjunction with a reduction in respondent burden, is noteworthy in light of the importance of the role-participation domains and the increasing importance of practical considerations in selecting health measures for widespread use.

Precision. Like respondent burden, precision in part varies directly with the numbers of items and response choices. Overall, the SF-8 scales are the coarsest, offering the least amount of precision and generally covering a narrower range of each of the eight health domains. The SF-12v2 provides more precision than the SF-8 in half of the domains, but less precision than the SF-36v2 in all the domains. Generally, scales with more levels provide greater measurement precision (see Table 3.1). The improvements embodied in the SF-36v2 and SF-12v2 significantly increased the precision of both of these surveys over their predecessors. Across all domains, the SF-36v2 health domain scales have as many or more levels, and thus greater measurement precision, than any of the SF-12v2 or SF-8 scales. This is an important feature to consider when sample sizes are small and measurement precision is paramount.

Note that the component summary measures of each of the adult Short Form instruments provide the greatest number of levels of measurement and, thus, more measurement precision than each of their respective form's health domain scales. For this reason, even the SF-8

Table 3.1

Comparison of the Number of Items and Levels of Measurement for Each Component Summary Measure and Health Domain Scale for the SF-8, SF-12v2, and SF-36v2

	SF-8		SF-12v2		SF-36v2	
	Items	Levels	Items	Levels	Items	Levels
PCS	8	382	12	441	36	486
MCS	8	386	12	438	36	494
PF	1	5	2	5	10	21
RP	1	5	2	9	4	17
BP	1	6	1	6	2	11
GH	1	6	1	5	5	21
VT	1	5	1	5	4	17
SF	1	5	1	5	2	9
RE	1	5	2	9	3	13
MH	1	5	2	9	5	21

component summary measures may provide sufficient measurement precision for studies involving small sample sizes.

Treatment of missing data. Two procedures have been developed for estimating Short Form survey scores when there are missing data: the *Half-Scale Rule* and *Full Missing Score Estimation (Full MSE)*; see Chapter 6). These procedures can be applied to data from any of the Short Form surveys; however, the most robust treatment of missing data occurs with the SF-36v2, followed by the SF-12v2, and, then the SF-8. Note that the Full MSE method requires the use of the QualityMetric Health Outcomes Scoring Software 5.0 (Saris-Baglama et al., 2011; see Chapter 5).

Data quality evaluation. Several measures and procedures have been developed or are otherwise available for evaluating the quality of data obtained from the administration of the Short Form surveys, including completeness of data, responses within range, confir-

mation of the two-component structure, percentage of estimable component scores, convergent validity, discriminant validity, consistent responses, percentage of estimable scale scores, item internal consistency, item discriminant validity, and scale reliability. Note that each of these data quality evaluation methods cannot be used with every Short Form instrument (see Table 3.2; see also Chapter 6).

Ceiling and floor effects. Additional considerations when choosing a Short Form survey are ceiling and floor effects. With the exception of the RP and RE scales, the range of observed scores is greatest among the SF-36v2 health domain scales, compared to the SF-12v2 and SF-8 scales, although the differences are not great. The implication is that the SF-36v2 health domain scales define a wider range of each measured construct than do the SF-12v2 and SF-8 scales. Therefore, the ceiling and floor effects found with SF-36v2 scales are less problematic than those found with the SF-12v2 and SF-8 scales.

Norms. Norms for both the SF-36v2 and SF-12v2 are based on a 2009 U.S. general population sample, while the SF-8 norms are based on a 2000 U.S. general population sample. Although the international norms available for the SF-36v2 are not as abundant as those for its predecessor, the number of SF-36v2 translations is continually growing.

Norm-based scoring and interpretation. Norm-based scoring, in the form of *T* scores, and interpretation guidelines are available for each of the three adult Short Form surveys (see Chapter 14).

Availability of health domain scales. Interest in the ability to score the eight health domains is no longer a reason to favor the SF-36v2 over a SF-12 form, as has previously been the case. In contrast to the SF-12, which

Table 3.2

Short Form Data Quality Indicators, by Survey

Indicator	SF-8	SF-12v2	SF-36v2
Completeness of data	•	•	•
Responses within range	•	•	•
Confirmation of the two-component structure	•	•	•
Percentage of estimable component scores	^a	^a	^a
Convergent validity	•	•	^b
Discriminant validity	•	•	^c
Consistent responses			•
Percentage of estimable scale scores	•	•	•
Item internal consistency			•
Item discriminant validity			•
Scale internal consistency reliability			•

^aAssessed as part of estimable scale scores.

^bAssessed as part of item internal consistency.

^cAssessed as part of item discriminant validity.

yielded score estimates for only the two component summary measures (Ware, Kosinski, & Keller, 1995, 1996), the SF-12v2 has the advantage of yielding scores for all eight health domains in addition to scores for the physical and mental component summary measures. The SF-8 provides scores on all health domain scales and component summary measures as well.

Translations. Beginning in 1991 with the SF-36, the IQOLA Project adopted a multistage translation procedure designed to assure that translations of the instrument were not only conceptually equivalent to the U.S. source-form but also linguistically and culturally relevant (Aaronson et al., 1992; Bullinger et al., 1998). As of August 2011, more than 140 translations and English-language adaptations of the Short Form instruments had been completed pursuant to the International Quality of Life Assessment (IQOLA) Project (see Chapter 1), and other translation projects are currently underway. A list of translated versions of all the Short Form instruments is available at <http://www.qualitymetric.com>.

Chapter 13 of this manual provides a more detailed discussion of the SF-36v2 translations. Additional information about translations of the SF instruments, as well as information related to products, services, and licensing, can be found online at <http://www.qualitymetric.com>.

Documentation. Up-to-date manuals and/or guides that document survey development, scoring processes, and interpretation guidelines are available for the SF-36v2, SF-12v2, and SF-8.

Published literature. As of July 2011, over 17,000 articles and other publications about the Short Form surveys had been identified. Although most of these publications are about the SF-36, the number of published articles on the SF-36v2 and SF-12v2 is expected to quickly accelerate within the next few years. The most up-to-date information regarding published literature about all of the Short Form surveys can be found online at <http://www.qualitymetric.com> and <http://www.sf-36.org>.

Matching a Form to an Application: General Considerations

A number of factors should be considered when deciding which survey to use for a particular application. This decision hinges, in large part, on making a tradeoff between respondent burden and score precision. This and other considerations are addressed in the following sections.

Assessing and monitoring individual patients for clinical purposes. Originally, the SF-36 was used in population health surveys. Its brevity, however, made it and the SF-36v2 increasingly attractive for use in clinical

trials and for individual patient evaluation purposes in clinical practice.

Selecting a health status measure for assessing and monitoring individual patients for clinical purposes often requires a compromise between the burden placed on patients and medical staff to obtain the information and the usefulness of that information. Gathering health domain and component summary information is much less burdensome when employing the SF-12v2 instead of the SF-36v2, and it is even less burdensome when using the SF-8. At the same time, the SF-12v2 and SF-8 cover a narrower range of functioning and are less precise than the SF-36v2. Thus, the two shorter instruments provide less quantitative and reliable information about a patient's health status at any given point in time and about the amount of change in that status over time. Therefore, use of the SF-12v2 or SF-8 for assessing and/or monitoring individuals is discouraged. Instead, the DYNHA-administered SF-36 is recommended for this purpose; however, if a fixed-form instrument is required, then the SF-36v2 is recommended. Use of the SF-36v2 provides greater utility and breadth of coverage for both the component summary measures and health domain scales. For example, the SF-36v2's five-item MH scale, initially developed as the Mental Health Inventory (MHI-5; Berwick et al., 1991; Veit & Ware, 1983), has been found to be a psychometrically sound alternative to longer instruments for the screening of anxiety and affective disorders (Berwick et al., 1991). Its usefulness with individual patient evaluations has also been established in case study demonstrations (e.g., see Wetzler, Lum, & Bush, 2000; see also Chapter 12).

It is important to note that some experts in the field would contend that the psychometric properties of the SF-36v2 are not adequate for use in individual assessments. For example, McHorney and Tarlov (1995) argued that the SF-36 did not meet all of their six criteria for individual patient applications. These criteria were: (a) practical features (e.g., takes less than 15 minutes to complete), (b) breadth of health measured (e.g., includes scales for measuring physical and mental status), (c) depth of health measured (e.g., allows for adequate floor and ceiling), (d) cross-sectional measurement precision (e.g., internal consistency reliability greater than or equal to .90), (e) longitudinal-monitoring measurement precision (e.g., 2- to 4-week test-retest reliability greater than or equal to .90), and (f) validity (e.g., convergent and divergent validity, sensitivity to change).

According to the data available at the time, McHorney and Tarlov argued that the original SF-36 did not meet the aforementioned criteria for ceiling effects and reliability (internal consistency and test-retest).

However, these requirements may be too stringent and unrealistic. By these standards, the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), arguably the most widely used and researched objective abnormal personality assessment in the world, would not be considered appropriate for individual testing purposes because of the reliability of its scales (Butcher, Graham, Ben-Porath, Tellegen, Dahlstrom, et al., 2001, Table E-4). Regarding the SF-36 survey, the floor effects were particularly problematic for the RP and RE scale; however, these effects were significantly reduced when these scales were revised for the SF-36v2.

Furthermore, McHorney and Tarlov's required "practical features" can't realistically be achieved without some sacrifice of their other required features, whether it comes in the form of lowered validity or reliability or of limitations in the breadth or depth of measurement. In some cases, as with the SF-36v2's MH scale previously mentioned, brevity may not always require such a compromise. In short, many experts would argue that the SF-36v2 is much more than "adequate" or "acceptable" for individual patient assessment, especially in light of the demands that health care systems place on such instruments (e.g., brevity, ease of use) if they are to be incorporated into the daily work flow of care providers (e.g., Maruish, 2002).

Perhaps more importantly, providers considering the SF-36v2 must decide whether patient evaluations are better served with or without the information that this survey provides. It is the contention of its developers that SF-36v2 results for an individual patient will always contribute to the evaluation of that patient by providing either new information or information that supports or clarifies the provider's clinical impressions. Further discussion on and illustration of the use of the SF-36v2 for clinical purposes can be found in Chapters 2 and 12, respectively.

Detecting small group differences. Because a high standard of score reliability (.90 or higher) is recommended to achieve satisfactory statistical power, single-item health scales like those in the SF-8 are likely to be inadequate or wholly unable to detect only very large differences. In such situations, use of the DYNHA engine would provide the best solution. However, the SF-36v2 and SF-12v2 are recommended for efforts focused on detecting small group differences when DYNHA is not an administration option. The improved precision afforded by the two longer measures can be observed through narrower confidence intervals around score estimates.

Large population surveys and samples. The SF-36v2, SF-12v2, or SF-8 can each be considered for use in the

largest population surveys and for studies involving large samples and group-level comparisons. Single-item measures, such as those used for all the SF-8 scales and four of the SF-12v2 scales, work well in these situations because the precision of mean scores is determined more by sample size than by increasing measurement reliability. Although concerns have been expressed in the past about single-item measures, several of these concerns are addressed by the use of norm-based scoring algorithms (see Ware, Kosinski, Dewey, & Gandek, 2001), making the SF-8 an appropriate choice for large surveys of representative samples. Furthermore, because statistical power is, in part, a function of sample size, the SF-8 may be the more viable and practical tool for use in large population studies.

Ongoing studies. The authors recommend against adopting either the SF-36v2 or SF-12v2 in "midstream;" that is, during the course of a longitudinal study that began with the use of the SF-36 or SF-12, respectively. Unless there are many years remaining in a longitudinal panel study, the threat to validity and the cause for concern perceived by others may be too great to justify such a change. In these cases, parallel administrations of items from the both versions of the chosen survey may provide the additional data necessary to determine whether estimates of scores generalize across the two versions of the instrument. Although QualityMetric Incorporated has discontinued the licensing of data collection and scoring services for the SF-36 and SF-12 surveys, such services for in-process studies or projects involving either instrument are still available from QualityMetric Incorporated.

Another potential concern with regard to ongoing studies has to do with adopting the SF-36v2 2009 scoring algorithms and norms during the course of a study that initially employed the 1998 scoring algorithms and norms. More generally, the issue is whether SF-36v2 data based on 2009 and 1998 scoring algorithms and norms can be or should be combined or compared within a single study or across studies. This issue is addressed in Chapter 14.

Cross-cultural studies. An important feature of the Short Form surveys is the availability of translated versions for use in non-English speaking countries or with U.S. samples for which English is not the first or primary language. Translations and/or English-language adaptations are available for the SF-36v2, SF-12v2, and SF-8; moreover, there are efforts to continue developing additional translations and adaptations for these surveys. Users requiring a translated version of one of the Short Form surveys can consult QualityMetric Incorporated's website (<http://www.qualitymetric.com>) for a current

Table 3.3

Summary of Fixed-Form Short Form Health Survey Similarities and Differences

Characteristic	SF-36v2	SF-12v2	SF-8
Improved item wording ^a	•	•	•
Increased range	•	•	
Improved format ^a	•	•	•
Standard form (4-week recall)	•	•	•
Acute form (1-week recall)	•	•	•
Acute form (24-hour recall)			•
Eight-scale profile	•	•	•
Component summary measures	•	•	•
2009 U.S. general population norms	•	•	
Translated versions	•	•	•
Use for individual patient assessment	•	• ^{b,c}	• ^{b,c}
Use for detection of small differences in group data	•	•	
Use for large samples	•	•	•
Use with population surveys	•	•	•

^aImprovement over SF-36/SF-12.

^bUse of the PCS and MCS summary scores is most appropriate for this application.

^cHealth domain scales are appropriate for use with individuals only when very large score differences are expected.

list of translated versions available for each instrument. Short Form users should contact QualityMetric if a desired translation for a particular Short Form is not available.

A summary of the general similarities and differences amongst the three Short Form surveys can be found in Table 3.3.

Matching a Form to an Application: Specific Form-to-Form Considerations

SF-36v2 versus SF-12v2. The SF-12v2 is the instrument of choice for surveys that require a shorter instrument than the SF-36v2. Large population health surveys can take advantage of its relative brevity while having confidence that, with only rare exceptions, group differences and changes in health status over time will be detected and that scores and interpretive guidelines will be directly comparable with those from the SF-36v2. The fact that the SF-12v2 comprises a subset of the SF-36v2 items is a noteworthy advantage if a study's objectives are the maximum comparability of results and the equivalence of population norms and other interpretive guidelines developed for the longer instrument. Most publications documenting previous "head-to-head" comparisons between the SF-12 and SF-36, including studies of responsiveness, reached the same conclusions about the PCS and MCS measures (see Ware, Kosinski, Turner-Bowker, & Gandek, 2002). Among the most common criticisms noted in published reports from such studies are the observed ceiling and floor effects, particularly for the two SF-12

role-participation scales. However, the survey's developers did not intend for the eight health domain scales to be scored from SF-12 item responses because of their coarseness and observed ceiling and floor effects. Thus, the SF-12v2 represents a substantial improvement in that regard and provides a means of scoring both the health domain scales and the component summary measures.

SF-12v2 versus SF-8. The SF-8 provides an even shorter survey option for purposes of estimating the health domain scale and component summary measure scores in the largest of population health surveys. However, unlike the SF-12v2, items in the SF-8 are *not* a subset of those in the SF-36v2, which may be a disadvantage depending on the purpose of the study and the degree of direct comparability demanded (see Ware, Kosinski, Dewey, & Gandek, 2001). Scores for all SF-8 health domains are estimated from single-item measures, as are scores for four of the SF-12v2 scales. As previously noted, such single-item measures perform best in very large surveys of general and specific populations because precision is achieved much more by drawing upon the large representative sample than by increasing measurement reliability. The SF-12v2 is also the instrument of choice for studies that require greater precision over a wider range of levels of health.

Concerns about single-item measures still apply (McHorney, Ware, Rogers, Raczek, & Lu, 1992; Ware, Kosinski, & Keller, 1996); however, these concerns have diminished due to advances in item-response categories and improvements in scoring algorithms for single-item

scales. Also, there is a better understanding of the conditions under which the standard error of the measurement of an *individual*, as opposed to the standard error of a *group mean*, is worth a substantial increase in respondent burden. The usefulness of well-constructed, single-item measures in group-level clinical trials and outcomes research projects is a subject of considerable ongoing interest and research (e.g., Aoki, Fleming, Griffin, Lacey, & Edmundson, 2000; Patterson et al., 2000; Silagy, Griffin, Lacey, & Edmundson, 1998; Ware, Kosinski, Dewey, & Gandek, 2001).

Short Form fixed-form measures versus CAT.

The highest level of score accuracy is often required for those survey applications focusing on individual scale scores or those needing to detect the smallest of important changes in health status in very small group-level analyses. For the most demanding applications, users no longer need to rely on short or long fixed-form instruments to achieve more practical or more precise measures. Research in progress suggests that software based on CAT logic, such as is employed by the DYNHA system, provides the best solution.

**PART II:
DATA COLLECTION
AND SCORING**



4

Survey Administration

This chapter presents guidelines for administering the SF-36v2, beginning with person-specific considerations—age, reading level, language, and level of cooperation and understanding—for determining how appropriate it is for the respondent to complete the instrument. Considerations for selecting the appropriate form (standard vs. acute) are also addressed. Specific guidelines for administration are also provided, including suggested scripts for introducing and concluding administrations to respondents and groups of respondents. Common questions and concerns raised by administrators (e.g., *What should I do if the respondent does not answer all the items?*) and respondents (e.g., *What do my answers mean?*) are identified and addressed, and a tabular summary of the most important *Dos* and *Don'ts* of SF-36v2 administration is provided.

Following the provided administration instructions and recommendations is particularly important when the survey administrator administers the paper-and-pencil version of SF-36v2 in person to one or more respondents. The survey can also be administered via face-to-face or telephone interview, mail-out/mail-back paper form, or online. (Note that scripts for face-to-face or telephone administration are available from Quality-Metric Incorporated.) Specific considerations for each of these administration modes are provided here, as are summaries of studies that have investigated the effects of some of these data collection methods. Finally, matters pertaining to the administration environment are discussed, as is the inclusion of the SF-36v2 as part of a longer interview, survey, or other data collection effort.

The guidelines that follow assume that a trained administrator oversees the administration of the SF-36v2 and that the respondent meets the eligibility requirements for completing the survey. For in-person administrations, it is particularly important for the administrator to establish rapport with the respondent and encourage completion of the survey. The administrator

can emphasize to respondents the importance of their answers to the completion of a study or as an addition to their medical records. The administrator can also answer questions, address concerns about the SF-36v2, and ensure the surveys are correctly and completely filled out. Respondents are more likely to fill out a survey honestly and completely if they have a positive impression of or relationship with the administrator.

Determining Respondent Eligibility

Age

The SF-36v2 was normed for use with adults; thus, use of the norms in this manual should be limited to respondents aged 18 years and older. Items like those in the SF-36 have been successfully administered to respondents as young as 14 years using self-administration and interviewer administration over the telephone and in-person (Ware, Brook, et al., 1980), and SF-36 translations have been successfully administered to those as young as 15 years (Gandek & Ware, 1998a).

Reading Ability

In situations where participation requires completion of a self-administered survey, potential respondents should be excluded if they are unable to read the survey due to limited reading ability. Before giving a respondent a survey form, the examiner should determine if any information is available regarding the respondent's ability to read. Using the Microsoft® Word readability determination feature, the SF-36v2 standard form was found to have a Flesch-Kincaid Grade Level score of 6.9 and a Flesch Reading Ease score of 68 on a 100-point scale. Note that the closer a Flesch Reading Ease score is to 100, the easier the text is to read. In most cases, a Flesch Reading Ease score of 60 to 70 is desirable (Millhollon & Murray, 2001).

If a study is expected to have a large number of respondents who have visual impairments, a large-type version of the survey should be prepared. It should be noted that the printing of special forms does add to the cost and complexity of data collection and survey administration; however, when necessary, this is a good investment. Also note that any large-type version must maintain the instrument's standardized content and format.

If a respondent is unable to read the SF-36v2 form for any reason, do not offer him or her the survey form; rather, conduct the assessment using the appropriate (standard or acute form) interview script (available from QualityMetric Incorporated, as previously noted), and record that the survey was not self-administered due to reading ability. The interview script can also be used if the SF-36v2 is administered to a large group of respondents who are unable to read. In this case, printed survey forms and pencils would be provided to the respondents, the items would be read aloud, the numbers corresponding to the response options for each item would be read along with the responses, and the respondent would be asked to record his or her response using the item response numbers on the survey form as a guide.

It is important to note that the order of administration of Items 7 and 8 from the Bodily Pain health domain scale is reversed on the SF-36v2 standard and acute form interview scripts. Thus, the scores for these items obtained using an interview script must be reversed (i.e., the response to Item 7 from the interview script should be entered in the Item 8 response area on the paper form, and vice versa) before applying the BP scale scoring rules (see Chapter 5).

Non-English-Speaking Respondents

If a respondent does not speak English, first determine if information is available regarding the respondent's ability to read English. If it is believed that the respondent is able to read English at least at the sixth grade level, proceed with survey administration. If he or she is unable to read English at this level or prefers to complete a translation of the survey, provide the respondent with a version that is translated into his or her native language. Bilingual respondents should be given the choice of completing either the English or translated form, if the appropriate one is available. A list of translated versions of the SF-36v2 can be found at <http://www.qualitymetric.com>. If the respondent cannot read English but can understand and speak English, administer the survey using one of the standardized interview scripts. In lieu of the availability of either option, record that the SF-36v2 was not completed due to a language barrier.

Level of Respondent Cooperation and Understanding

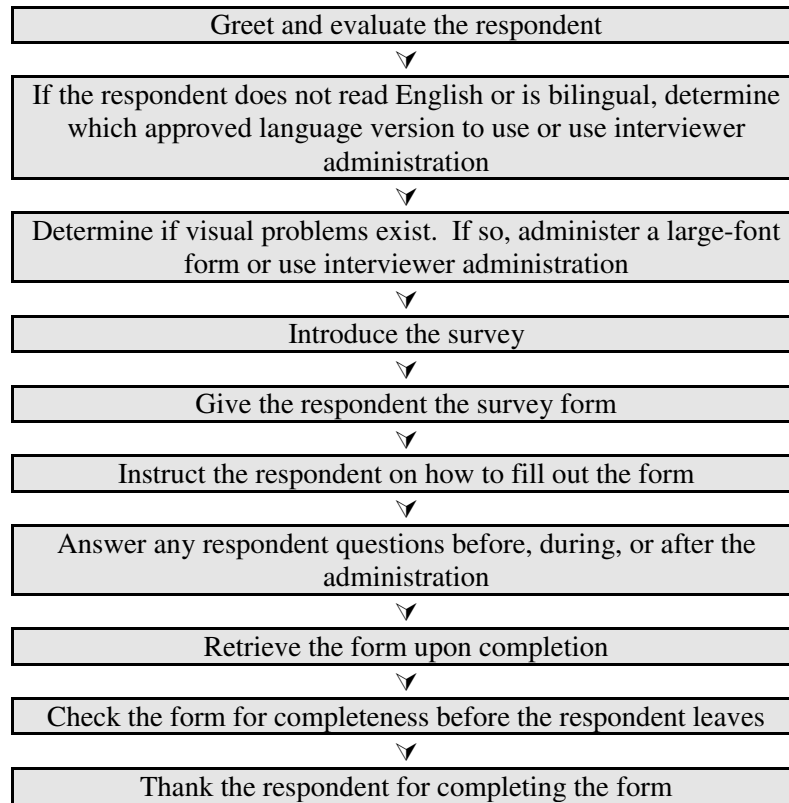
It is important for those completing the SF-36v2 to be willing to openly and honestly answer the survey questions. Generally, those administering the survey will find respondents to be interested in and cooperative when answering the survey questions. The SF-36v2 requires little in terms of respondent time (5–10 minutes on average) and its content is generally nonthreatening. However, there are times when administrators will encounter respondents who have difficulty or are resistant to completing all or part of the survey. Suggestions about how to handle these situations are presented later in this chapter.

There may also be times when the respondent's physical or mental condition precludes him or her from responding to items in a manner that accurately reflects his or her health status, despite his or her willingness to complete the survey. For example, a respondent experiencing a psychotic episode with poor reality testing may exaggerate or minimize his or her general health status due to an inability to comprehend the items or otherwise accurately assess his or her physical and mental health. A respondent experiencing acute and/or severe pain may display similar problems. In situations such as these, it is better to delay administration of the survey until the respondent's condition has stabilized.

Guidelines for Administration

The SF-36v2 should be administered in a standardized manner using the standardized administration formats. Any change to the physical format of the survey form or, in the case of interview administration, the interview script may affect the way respondents answer the questions, thus compromising the validity of results. This includes removing specific questions from the printed form or interview script. Maintaining standardization in administration helps to ensure the accuracy and correct interpretation of results. Those wishing to use an abbreviated version of the SF-36v2 should consider instead using the SF-12v2 or SF-8 (see Chapter 3).

Each SF-36v2 standardized paper form includes specific instructions, questions, and response choices presented in a standardized format. Using the standardized SF-36v2 standard (4-week recall) and acute (1-week recall) forms that are available from QualityMetric Incorporated or its authorized resellers helps to ensure standardization of administration and accuracy in the interpretation of survey results.

Figure 4.1 Recommended Steps for Administering the SF-36v2

The flow chart in Figure 4.1 summarizes recommended steps for in-person administration using either the standard or acute version of the paper form.

When to Administer the Survey

In a clinical setting, the SF-36v2 should be administered before the respondent sees a health care provider so that the interaction between the respondent and the provider does not influence the respondent's answers to the survey. Ideally, the survey should also be administered before the respondent is asked other health questions or about concurrent illnesses, again so that any such discussion of health problems does not influence the respondent's answers to the survey questions.

Introducing the SF-36v2 to the Respondent

The following script (or a variation appropriately reworded to sound more like the administrator's style of speech) is suggested for introducing the SF-36v2:

We would like to better understand how well you are able to do your usual activities and how you rate your own health. To help us better understand these things about you, please complete this questionnaire about your general health.

The questionnaire is simple to fill out. Be sure to read the instructions on the top of the first page [point to them]. Remember, this is not a test and there are no right or wrong answers. Choose the response that best represents the way you feel. I will quickly review the questionnaire when you are done to make sure that all the items have been completed.

Please fill out the questionnaire now. I will be nearby in case you want to ask me any questions. Return the questionnaire to me when it is complete.

[As appropriate, add:] *You should answer these questions by yourself. Spouses, other family members, or friends should not assist you in completing the questionnaire.*

Addressing Problems and Questions

It is not unusual for respondents to ask questions or display certain types of behaviors before, during, or after the administration of the survey. Several common questions and behaviors that experienced SF-36v2 administrators have encountered over the years and suggestions as to how to respond to them follow.

What should I do if the respondent refuses to fill out the SF-36v2? Respondents are not required to complete the survey. If the respondent is able to self-

administer the survey but refuses to participate, tell the respondent that completion of the survey is voluntary but that it would provide helpful health-related information. In clinical settings, point out that survey completion would help the physician better understand the respondent's health problems.

Emphasize that the data the survey provides are as important as any other type of medical information. Explain that the survey responses are essential in order to get a complete picture of the respondent's health, emphasizing that the survey is simple to complete. Suggest that it is possible that this survey is different from others the respondent has filled out in the past, and that he or she may even enjoy completing this survey. If the respondent still refuses, take back the survey form, record the reason for refusal, and thank the respondent.

What if the respondent does not answer all of the items? If noncompletion is a result of the respondent having trouble understanding particular items (i.e., the questions and/or their response choices), ask the respondent to explain why he or she had difficulty responding. Reread the items aloud for him or her verbatim, but do not rephrase the items in any way. If the respondent is still unable to complete the survey, accept the survey as incomplete, and indicate that the respondent was unable to complete the entire survey due to difficulty understanding the items.

If the respondent is unable to self-administer the survey, document the reason. If the reason is health-related, indicate the specific condition.

What should I do if the respondent asks for clarification of an item? While completing the survey, some respondents might ask for clarification of specific items so that they can better understand and respond to them. If this happens, assist the respondent by rereading the item aloud for him or her verbatim. If the respondent asks what something means, do not offer an explanation; rather, suggest to the respondent that he or she use his or her own interpretation of the item. All respondents should answer the items based on what they think each means.

Sometimes respondents may experience other types of difficulty with the response choices. They may answer, "I don't know," or something other than what is stated on the survey. In these circumstances, it is important to gently guide the respondent to indicate one of the response choices by saying something like:

I know that it may be hard for you to think this way, but which of these categories most closely expresses what you are thinking or feeling?

It is possible that respondents may ask if certain items, particularly the pain items, are limited to a specific

health problem. Explain to the respondent that these items are referring to their health in general.

If the respondent does not like an item or thinks it is unnecessary or inappropriate, emphasize that all items are in the survey for a reason that is very important to the clinician or researcher. Ask that they try to answer all of the items.

Differences in answers due to different wordings of survey items can bias results; thus, it is important to minimize these differences. If the respondent has repeated difficulties filling out the survey that the administrator cannot address using these suggestions, thank the respondent, take back survey form, and record the difficulty.

What should I do if the respondent wants to know what his or her answers mean? Sometimes a respondent may ask the survey administrator for an interpretation of his or her responses or for his or her scores. If the respondent's care provider is the person administering the survey, tell the respondent that you will discuss his or her responses after the survey is completed and scored. If administered by another person in a clinical setting (i.e., someone other than the care provider), tell the respondent that his or her provider will interpret the results for him or her. In research settings, tell the respondent that you are not trained to score or interpret the survey.

What should I do if the respondent is concerned someone will see his or her answers? Be honest with the respondent. If someone else might have access to his or her item responses or scored results and may identify them as belonging to him or her, tell the respondent who that might be and why they might be looking at the findings. Then address any concerns the respondent might have about this. Otherwise, emphasize that all respondents' responses to the SF-36v2 will be kept confidential. If an ID number is used to identify respondents, point out that their names do not appear anywhere on the survey, meaning their results will be linked with an ID number and not with their name. If the survey is administered as part of a clinical study, tell respondents that their survey answers will be pooled with other respondents' answers and analyzed as a group rather than on an individual basis.

What should I do if the respondent asks why the SF-36v2 must be completed more than once? If the SF-36v2 is to be readministered in the future, explain that respondents must fill out the same survey at a later time to see if their answers change, which will provide a more complete picture of each respondent's health over time.

What should I do if some of the questions do not pertain to the population that I am studying (e.g., having paraplegics answer the walking items in the

Physical Functioning scale)? While acknowledging that some items may not seem to apply to a given respondent, the respondent should be asked to answer *all* of the items, regardless of any permanent physical or mental limitations or impairments. There are three reasons for doing so. First, as previously indicated, all items must be administered in order to maintain the standardization of the instrument. Second, the items will accurately reflect the functional status of the respondent on the domain in question. For example, asking a paraplegic respondent if he or she can walk 100 yards is a legitimate question; if answered honestly, the item will accurately reflect his or her physical impairment, which is what the SF-36v2 was designed to do. Third, depending on the purpose of the assessment, the effects of known permanent impairments can be addressed or taken into consideration when group or individual respondent results are interpreted.

Can I administer an SF-36v2 health domain scale separately? The eight health domain scales cover content areas that can be scored and meaningfully interpreted separately. Administration of all health domain scales, however, allows one to compute the PCS and MCS measures, which yield even more information. However, there may be circumstances in which administration of only a subset of Short Form scales is desired. It is not uncommon for QualityMetric to grant permission to use one or more individual Short Form health domain scales apart from the others. A common example is the use of a single health domain scale in a randomized clinical trial. The validity of an extracted scale can be maintained, depending on the context in which it is administered. In some instances, however, the comparability and/or interpretation of a single scale administered apart from its source could become compromised. If one chooses to administer a single SF-36v2 scale, it is recommended that it be administered before any disease-specific instrument that may also need to be administered to the respondent. An exception, however, should be made with regard to the MH scale. MH scale items, which may be upsetting to some respondents experiencing emotional problems, are rarely administered first for that reason.

In other circumstances, users may wish to extract and use only specific SF-36v2 items. It is important to be aware that the administration of single items from a health domain scale may yield data with limited interpretability. If one wishes to use a briefer instrument, the SF-8 or SF-12v2 should be considered. In either case, be aware that single items usually provide coarser measures than multi-item scales or measures.

Can I use the SF-36v2 with another generic survey or a disease-specific survey? The SF-36v2 can be used with a disease-specific survey or with another generic

survey. The benefits of doing so, as well as advances in the assessment of disease impact, are addressed in Chapter 1. However, the survey items should maintain their order and format and should not be mixed with items from other instruments. When used with a disease-specific survey, the SF-36v2 should be administered before the other measure to avoid sensitizing the respondent to disease-specific health status issues that may then influence his or her responses to the SF-36v2 questions about general health status.

Concluding Survey Administration

When the respondent returns the survey form, check it for completeness. Note whether all of the survey questions have been answered. If the survey is not complete, ask the respondent whether he or she had any difficulty completing it, and record the reasons for noncompletion. Finally, thank the respondent using the following exit script (or a variation appropriately reworded to sound more like the administrator's style of speech):

Thank you for taking the time to complete this survey. It is possible you will be asked to complete the questionnaire again at a later date.

In some instances, the respondent may be providing other information during his or her visit. In such cases, a specific thank you for completing the survey may not be required or appropriate. Finally, the completed survey form should be stored in a safe and secure place to ensure confidentiality.

Specific *Dos* and *Don'ts* for SF-36v2 administration are summarized in Table 4.1.

Modes of Administration

QualityMetric offers a variety of ways respondents can complete the SF-36v2, which are described in the following sections.

Paper and Pencil

As previously described, the paper-and-pencil mode of administration allows respondents to complete a paper-based version of the SF-36v2. Administration via paper form can be done in-office or through mail-out/mail-back or fax-back procedures.

Interviewer Script

A standardized interviewer script is available for oral administration of the SF-36v2. This is ideal when respondents are unable to complete the survey on their own or when survey administration via the telephone is required.

Table 4.1*SF-36v2 Administration Dos and Don'ts*

DOs	DON'Ts
DO introduce the SF-36v2 and explain the reasons for completing it and the importance and advantages for the respondent of doing so.	DO NOT minimize the importance of the SF-36v2.
DO have respondents complete the survey before they fill out any other health data forms and before they see their healthcare provider.	DO NOT discuss respondents' health, health data, or emotions with them before they complete the survey.
DO be warm, friendly, and helpful.	DO NOT force or command respondents to complete the survey.
DO request and encourage respondents to complete the entire survey.	DO NOT accept incomplete survey forms without first encouraging respondents to respond to any unanswered items.
DO read and repeat a question and its response choices verbatim for respondents if they ask for clarification.	DO NOT change the wording of questions or response choices.
DO tell respondents to answer items based on what they think each item means.	DO NOT interpret or explain items.
DO have respondents complete the survey by themselves.	DO NOT allow spouses, family members, or friends to help respondents complete the survey. Ideally, caregivers should not be present during this assessment.
DO inform respondents if they will be asked to fill out the same survey again.	
DO thank respondents for completing the survey.	

Online

Online administration allows respondents to complete QualityMetric health surveys online from any location where Internet access is available. Two online options are available: standard versions via QualityMetric's <http://www.amihealthy.com> website and fully customized versions that act as an extension of a client's existing Internet presence. Once an online health survey is submitted, the data is captured directly into QualityMetric's Smart Measurement System (see Chapter 5) for scientifically valid scoring, interpretation, and reporting in real time, eliminating the need for time-consuming data entry.

Fax

The fax mode of administration allows respondents to complete a specialized, paper-based version of the health surveys. Once a survey is completed, it is faxed to QualityMetric's centralized server via a number provided to the user. The data are then loaded directly into QualityMetric's Smart Measurement System for scoring, interpretation, and reporting in real time, eliminating the need for time-consuming data entry and possible transcription errors. This mode is ideal for organizations with limited Web presence, Internet access, or technical infrastructure.

Smartphone

Smartphone administration is valuable for those users who are on the go and require the fast turnaround of scored data. It is well-suited to providers that have

embraced handheld devices as part of their everyday workflow and have a high degree of interaction with their patient population. Once a survey is submitted, the data are then transmitted via the Internet for scoring by QualityMetric's Smart Measurement System. Scores are immediately calculated, and a report is then sent to the user's device for review. In addition, full reports are available in real time via the Smart Measurement System platform.

Tablet or Kiosk

QualityMetric supports administration of its online versions of the SF-36v2 via tablet or kiosk, operating much like a laptop, provided that these devices are Internet-enabled. Single-item electronic patient-reported outcomes (ePRO) forms are provided to licensed customers for the programming of single-item presentation via tablet, kiosk, or other similar device. Once such forms are obtained, customers then contract directly with an ePRO vendor for software development.

Considerations for the Use of Interview, Mail, or Online Format

The instructions and recommendations provided up to this point apply when the SF-36v2 is administered to one or more respondents in person. Common modes of administration for clinical purposes include in-office *supervised self-administration*, just previously described, and *mail-back administration*, in which an established respondent is given the form to complete at home and then return by mail. Administration via other modes or

methods—*face-to-face or telephone interview, mail-out/mail-back, or online*—for either clinical or research purposes require additional considerations in order to elicit reliable and valid information.

Administration by interview. The SF-36v2 can be administered by interview, either face-to-face or over the telephone. In either case, it should be administered using a script available from QualityMetric Incorporated. The administrator should request that the respondent's caregiver (if present) leave the room during administration of the survey, unless circumstances indicate that it would be better for the caregiver to be present.

As with any health survey, administrators should be familiar with SF-36v2 administration guidelines in advance and should ensure that the assessment environment is conducive to its purpose. An introduction to the administration, such as the following, should be given prior to reading the first question:

We would like to better understand how you feel, how well you are able to do your usual activities, and how you rate your own health. To help us better understand these things about you, please answer some questions about your general health.

This is not a test, and there are no right or wrong answers. Choose the response that best represents the way you feel. Please answer every question. As we proceed, please feel free to ask me any questions you may have.

If the respondent is to indicate his or her answers on a paper form rather than by giving an oral response, the administrator should provide the respondent with a firm writing surface, such as a clipboard or table top, and a pen or (if using a scannable answer sheet) a #2 pencil.

As during a paper-and-pencil administration, the administrator should not attempt to interpret or explain any of the items; rather, he or she may repeat an item verbatim if asked. The administrator should request and encourage respondents to provide an answer for each question but should not force them to do so. Additional instructions specific to each section of the assessment are presented in the interview scripts available from QualityMetric.

When administered to respondents with mild cognitive impairment or early dementia, it is recommended that the administrator be suitably prepared and trained to properly administer the survey. Respondents with mild cognitive impairment may demonstrate some behaviors unlike other groups of respondents, and patience and redirection may be necessary to encourage survey completion. If possible in these situations, the same administrator should interview respondents for each subsequent survey readministration required.

Note that when using the interview script for oral administration of the survey, Items 7 and 8 from the Bodily Pain health domain scale are administered in reverse order from the way they appear on the printed SF-36v2 form. Reversing the order of the presentation of these two items facilitates the flow of their oral administration. Therefore, when conducting an oral administration of the survey, the administrator must inform respondents using an SF-36v2 paper form of the order discrepancy to ensure that the intended responses are marked in the appropriate response areas. That is, the response to Item 7 from the interview script should be entered in the Item 8 response area on the paper form, and vice versa. If the administrator is writing down the respondent's oral answers, he or she must be mindful of the reverse ordering of the items when entering and scoring the responses.

Administration by mail. Administration of the SF-36v2 using a mail-out/mail-back (MO/MB) system is a common and efficient means of conducting research that involves large numbers of subjects who are scattered over a large geographical area and/or multiple administrations of the instrument over long periods of time. This method can also be useful for clinical purposes. For example, it can provide a means of monitoring patients with chronic conditions during long intervals (e.g., 6 months) between scheduled visits. It can also be used to assess the enduring effects of treatment long after treatment has been terminated.

There are many issues to consider when deciding whether to use an MO/MB system. In addition to concerns about maintaining patient privacy, confidentiality, and standardization of administration, other practical considerations should be addressed, such as identifying the most effective MO/MB methodology for the population being assessed, the cost of implementing such a system, and the expected return on that investment. It is beyond the scope of this manual to adequately address these issues. Therefore, those employing an MO/MB methodology are referred to Dillman, Smyth, and Christian (2009) or other resources that specifically address these and other important issues to consider in conducting mail surveys.

Online administration. Use of online administration of the SF-36v2 generally has some of the same advantages and involves some of the same issues as the MO/MB methodology. In addition, it is imperative that the standardized format of the survey's items be maintained as much as possible for online screen presentation until alternative presentations have been empirically investigated. Online administration of the SF-36v2 through QualityMetric Incorporated is available at <http://www.amihealthy.com>. Note that Dillman et al. (2009) is an

excellent resource to consult when considering the Internet administration of surveys such as the SF-36v2.

Effects of Data Collection Method

Several studies involving the SF-36, SF-36v2, or other Short Form instruments have demonstrated that different methods of administration may have an effect on the results obtained. Because of the comparability of the SF-36 and SF-36v2 (see Chapter 13), the findings from a few of those studies are presented here and their methodologies and results are summarized in tabular form in Table 4.2.

To begin, studies have shown that responses to the SF-36 tend to be more favorable when data are collected by face-to-face or telephone interview (McHorney, Kosinski, & Ware, 1994; Ware, Kosinski, & Keller, 1994). In a randomized trial conducted during the norming of the SF-36, McHorney, Kosinski, & Ware (1994) found a lack of equivalence in some domains between responses to MO/MB surveys and those from personal interviews administered by phone. Average scores for the MCS measure were 2.43 points higher (± 0.3 , $p < .001$) for those interviewed by telephone than for those who self-administered the survey by the MO/MB method. This difference is nearly one-fourth of a standard deviation, a noteworthy amount. In other terms, the effect of data collection method on MCS scores is approximately one-fourth the impact of a depressive disorder. Underlying this difference in MCS scores were significant differences for seven of the eight health domain scales (all but GH). There was no effect on the PCS measure.

In another study, Ware, Kosinski, DeBrotta, Andrejasich, and Bradt (1995) sought to determine the effect of three SF-36 administration methods on patient acceptance, cost and quality of data, mean scale scores, test-retest and internal consistency reliabilities, and empirical validity using a randomized study with cross-over of half the patients at the time of retest administration. Respondents recruited at ambulatory care facilities and nonmedical business work sites ($N = 525$) were randomly assigned to complete their first SF-36 by personal interview over the telephone, self-administration through the mail, or by IVR technology. Two-weeks later, half of the respondents completed the survey again using the same administration method, while the other half were randomly assigned to another method, for a total of nine possible sequences of data collection methods. Preliminary results revealed no differences in data quality or tests of scaling assumptions across the three administration methods. Average PCS scores did not differ by method. However, average MCS

scores were more favorable (by 1.8 points, $p < .01$) for the personal telephone interview compared to both self-administered and IVR-administered surveys. The latter two methods did not differ.

Among the most important issues involved in the widespread use of patient-based health outcomes assessments are their cost and the comparability of results across data collection methods. Results from the Ware, Kosinski, DeBrotta, et al. (1995) study suggest that responses to SF-36 mental health scales administered by personal telephone interview should not be directly compared with those administered by other methods without adjustment for the effect of data collection method. Selected findings from this study are further discussed in Ware et al. (2007).

Because of the impact that data collection methods have demonstrated in previous studies of the SF-36 and the common practice of varying data collection methods within and between studies, investigations into the data collection methods used during the 2000 norming of the SF-8 were replicated and extended (Ware, Kosinski, Dewey, & Gandek, 2001). All studies included the SF-36v2 so as to replicate previous analyses. Also, the new studies were expanded to include online self-administrations of both the SF-8 and SF-36v2 ($N = 768$). In the study of online administration, responses were compared with those obtained from personal interviews administered by phone ($N = 750$) and MO/MB self-administered forms ($N = 907$).

Ware, Kosinski, Dewey, and Gandek (2001) found that for the SF-36v2 health domain scales and component summary measures, the pattern of differences in average scores across groups who were interviewed by phone versus MO/MB was not unlike the pattern observed in previous studies, although the obtained differences tended to be somewhat smaller. Differences were also apparent in both the PCS and MCS measures, as opposed to only the MCS measure. Across the health domain scales, five of eight scale differences (PF, BP, GH, VT, MH) were significant, with higher average scores for phone interviews (1.2–3.75 T -score points) in comparison with the MO/MB method. PCS and MCS scores were also significantly higher for those interviewed by telephone (1.68, $p < .001$, and 1.38, $p < .01$, respectively). In the 10 comparisons made between average SF-36v2 scores for online and MO/MB samples, no significant differences were found for the health domain scales or component summary measures. However, significant differences were found in other studies that compared the results of computer administration and paper-form administration of the SF-36 to the results from disease-specific health status surveys (see below).

It is important to note that the Ware, Kosinski, Dewey, and Gandek (2001) studies of data collection methods involved general population samples that were based on convenient samples in which study participants were not randomized to data collection methods. Further, because respondents differed substantially in their characteristics and response rates across methods, it was necessary to adjust for these differences using regression methods, as was done in previous studies. Despite attempts to thoroughly adjust for all measured differences in respondent characteristics, these regression-based estimates of the effects of data collection methods may be biased.

In another study, Saleh et al. (2002) mailed out a paper version of the SF-36 and the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) to 160 orthopedic (knee or hip pain) patients. Three weeks later, those who completed the MO/MB surveys were asked to complete the SF-36 again, either by paper form ($n = 45$) or on a “palmtop” computer ($n = 42$). Comparison of results for the two subsamples revealed no significant differences in mean squares, standard deviations (*SDs*), floor and ceiling percentages, or retest intraclass correlation coefficients (*ICCs*) based on the results of the second assessment. Significant differences were found, however, in the Cronbach’s alphas for both the PF and SF health domain scales for the two groups ($p < .03$), with greater internal consistency being noted for the paper-form administration.

Using a sample of 68 patients from an outpatient asthma clinic, Caro, Caro, Caro, Wouters, and Juniper (2001) compared results obtained from paper-form administration of both the SF-36 and the Asthma Quality of Life Questionnaire (AQLQ) to those obtained from electronic administration. Administration was counter-balanced so that half of the respondents were administered both instruments by paper form first; 2 hours later, both instruments were administered again, this time by “electronic diary.” The remaining respondents were administered the two instruments in the opposite order, using the same time interval. Concordance of responses to items across the two administration formats ranged from 59% to 91%, with almost half achieving a concordance rate of 80% or higher. The *ICCs* between the health domain scales from the two administration formats ranged from .83 for the MH scale to .97 for the BP scale, with no consistent variation being noted.

In a retrospective study, Hanscom, Lurie, Homa, and Weinstein (2002) compared the quality of SF-36 data obtained from 15,815 paper-form administrations of a survey that included both the SF-36 and the Oswestry Low Back Pain Disability Questionnaire to the results of 3,574 laptop touch-screen administrations of the same

surveys to patients being seen at member clinics of the National Spine Network. The computer survey sample was found to differ from the paper-form survey sample with regard to age, percentage of females, and percentage of high school graduates (all $ps < .000$); however, the computer survey sample was just as likely as the paper-form survey sample to be working or receiving worker’s compensation. At the same time, those completing the computer version of the survey were less likely to have completed high school.

Hanscom et al. (2002) found SF-36 data quality for the computer responders to be better than that for the paper-form responders from many perspectives, including missing value rates (the number of questions with missing responses divided by the total number of questions) for the survey overall (1.66 vs. 3.34, $p < .001$) and for the individual health domain scales and component summary measures ($p < .001$); percentage of surveys completely filled out (85% vs. 68%); percentage of health domain scales that could not be calculated due to missing responses (1% vs. 2–3%, $p < .001$); percentage of component summary measures that could not be calculated due to missing responses (3% vs. 8%, $p < .001$); and Response Consistency Index (RCI; see Chapter 5) scores (.12 vs. .16, $p < .001$). The reported statistical significance of RCI scores, as well those for age and gender, may be attributable to the large sample size. The investigators found that adjustments for the differences in age and education between the two samples actually enhanced the relationships between the method of administration and both the missing response rate and response consistency. Adjustments for gender had only a small effect on the findings.

Perkins and Sanson-Fisher’s (1998) study conducted in Australia revealed, in brief, that data collection costs were lower for the telephone mode of administration, contrary to what has been shown in other studies. A significantly higher consent rate was achieved with the telephone mode, with younger respondents being more likely to refuse participation via mail mode and older respondents more likely to refuse administration via telephone. The rate of missing responses was higher with the mail mode, and health ratings were generally more favorable with telephone administration. In addition, Cronbach’s alpha coefficients for the RP, VT, SF, and RE scales were found to differ significantly by administration method, with higher coefficients being obtained for the RP, SF, and RE scales with telephone administration.

In another study, Gwaltney, Shields, and Shiffman (2008) conducted a meta-analysis of 65 published studies that investigated the equivalence of paper and

electronic versions of a variety of HRQOL measures, including seven SF-36 studies. In some studies, a PDA was employed as the electronic mode, whereas the other studies employed a PC or laptop computer. Mean HRQOL scores for the paper and electronic versions were not significantly different (average mean difference was 0.2% of the scale range). Thirty of the 32 studies reporting correlations between paper form and computerized assessments had average correlations greater than .75 and the weighted summary correlation between modes was .90. In the four studies reporting paper-paper test-retest reliability and paper-computer concordance, the average correlations (.88 and .91, respectively) did not differ significantly, nor did the average PDA-paper correlation (.91) differ significantly from the average PC-paper correlation (.90). Although age was found to be negatively related to the paper-electronic correlations, the trend was very small and the correlations for the oldest age groups were greater than .75. Overall, Gwaltney et al. concluded that the two modes of administration produce equivalent HRQOL scores.

In light of the findings of these studies (summarized in Table 4.2), one should be aware that the method by which SF-36v2 data are collected may impact the obtained results. Consequently, the means of data collection should be considered in all studies involving the SF-36v2. For example, the National Committee for Quality Assurance (NCQA) subtracted 1.9 *T*-score points from the PCS score and 4.5 points from the MCS score derived from HOS results obtained from SF-36 telephone-interview surveys based on the findings of a Veterans Administration HOS subsample that also completed a VA survey (NCQA, 2004). Ideally, data collection should always be limited to one method if the data are to be aggregated or when an individual respondent's results are to be compared to results from his or her own survey, from another respondent, or from a group of respondents. When data collection methods do vary within a sample or when results are compared across samples assessed using different methods, the

effects of the methods used should be evaluated and the results interpreted with due caution.

For studies of elderly individuals being treated under Medicare, one should consider the recommendations published by the NCQA (2004) for correcting PCS and MCS *T* scores obtained from telephone administration of the SF-36. However, general population findings that included the Medicare population (McHorney, Kosinski, & Ware, 1994) support recommendations for correcting MCS scores but not PCS scores. Further studies are needed to determine whether different adjustments are warranted for PCS or MCS scores and to determine if adjustments are warranted for general population scores as well.

Additional Considerations

In addition to the guidelines previously provided, other considerations should be taken into account when administering the SF-36v2 as part of a clinical routine or a research protocol.

Environmental conditions. In all cases, the administrator should ensure that the environment is suitable for the purposes of assessment by controlling for unnecessary distractions such as noise, extremes in temperature, crowding, and interruptions. When the survey is administered via interview, the administrator should be warm and friendly towards the respondent; however, communications between the administrator and respondent should focus on SF-36v2 instructions and the interviewer script, in accordance with the administration guidelines previously set forth.

Order effects. In some cases, the SF-36v2 will be a component of an assessment battery that the respondent will undergo more than once. In these cases, its place in the order of the initial administration in a battery of assessment instruments and/or procedures should be maintained during follow-up assessments. A clear and concise instruction set should precede the administration of the SF-36v2, regardless of its placement in the assessment battery.

Table 4.2

Effects of Method of Data Collection on Short Form Results: Findings From Selected Studies

Investigation	Sample	N	Short Form		Methods of Data Collection	Key Findings
			Survey	Form		
Bliven, Kaufman, & Spertus (2001)	Cardiology clinic outpatients	66	RAND-36		Paper form, touch-screen online	Both interclass correlation coefficients (ICCs) and Pearson correlations for matching scales from the two modes were significant, suggesting no systematic variation by mode.
Buskirk & Stein (2008)	Cancer survivors	140 & 155	SF-36		Telephone interview, mail-out/mail-back (MO/MB)	With each sample using only one or the other administration mode, all scales from both modes achieved a Cronbach's alpha > .70, with alphas for the MO/MB scales generally higher by as much as .11. Overall, mean unadjusted scale scores were found to be higher for the phone mode, and the multivariate effect of mode was seen in higher RP, VT, and MH scores for the phone mode.
Caro, Caro, Caro, Wouters, & Juniper (2001)	Asthma outpatients	68	SF-36		Paper form, electronic diary	Concordance for identical responses to items across the two administration formats ranged from 59% to 91%, with almost half achieving 80% or higher; health domain scale ICCs ranged from .83 to .97, with no consistent variation being noted.
Hanscom, Lurie, Homa, & Weinstein (2002)	Low back-pain patients	15,815 & 3,574	SF-36		Paper form, laptop computer	Data quality for the computer responders was found to be better than that for the paper-form responders from several perspectives, including missing value rates for health domain scales, component summary measures, and the overall survey; percentage of surveys completely filled out; percentage of health domain scales that could not be calculated due to missing responses; percentage of component summary measures that could not be calculated due to missing responses; and Response Consistency Index (RCI) scores.
Jorgarden, Wettergen, & von Essen (2006)	Swedish citizens covered by civil registration, aged 13–24 years	585	SF-36		MO/MB, telephone interview	Cronbach's alphas for telephone interviews ranged from .62 (BP) to .86 (PF), while ranging from .77 (RP) to .91 (PF) for the MO/MB administration. Males scored significantly higher than females on 5 of the 10 SF-36 measures from the telephone interview and on 7 of the 10 measures on the MO/MB form.
Lungenhausen, Lange, Maier, Schaub, Trampisch, et al. (2007)	Randomly sampled patients taking part in the German Acupuncture Trials	823	SF-12		MO/MB, telephone interview	While the 16–19-year-old subsample did not differ significantly from the 20–23-year-old subsample on any of the 10 measures from either mode of administration, the 13–15-year-old sample scored significantly higher than the 16–19-year-olds on the telephone RE and the 20–23-year-olds on the telephone VT, SF, and MCS.
Lyons, Wareham, Lucas, Price, Williams, & Hutchings (1999)	British general medicine, urology, endocrinology, and gastroenterology outpatients	210	SF-36		MO/MB, face-to-face interview	Mean MO/MB MCS scores were significantly lower than telephone interview scores (mean difference = 3.5 T-score points) when compared to the mean MO/MB PCS score difference of 1.8 (considered to be within the range of equivalence for this study); administration order effects were also noted for MCS. MO/MB scores were generally lower than interview scores, regardless of order of administration, with the difference being significant ($p < .05$) for the GH, MH, PF, RE, and VT scales.

(Continues overleaf)

Table 4.2 (continued)
Effects of Method of Data Collection on Short Form Results: Findings From Selected Studies

Investigation	Sample	N	Short Form Survey	Methods of Data Collection	Key Findings
McHorney, Kosinski, & Ware (1994)	U.S. general population	1,682 & 782	SF-36	MO/MB, face-to-face interview, telephone interview	Results revealed a lack of equivalence between responses to MO/MB surveys and those from personal interviews administered by phone. Average scores for the SF-36 MCS measure were 2.43 <i>T</i> -score points higher for those interviewed by telephone versus by MO/MB. Underlying this difference in MCS scores were significant differences for all health domain scales except GH.
Perkins & Sanson-Fisher (1998)	Australian random community sample	418 & 421	SF-36	Telephone interview, MO/MB	Data collection costs were lower for the telephone mode. Significantly higher overall consent rate was achieved with telephone mode, with younger respondents being more likely to refuse participation via MO/MB mode and older respondents more likely to refuse administration via telephone. The rate of missing responses was higher with the MO/MB mode; health ratings were generally more favorable using the telephone administration. Cronbach's alpha coefficients for the RP, VT, SF, and RE scales were found to differ significantly by administration method, with the higher coefficients being obtained from the telephone administration, except for VT.
Ravens-Seiberer, Erhart, Wetzell, Krugel, & Brambosch (2008)	Two groups of randomly sampled German adults with children aged 8–17 years	899 & 791	SF-8	Telephone interview, MO/MB	With each sample using only one or the other administration mode, the phone sample achieved significantly higher scores on VT, MH, and MCS, whereas the MO/MB group was significantly higher on PCS. Interscale correlations were slightly higher for the MO/MB sample. On the phone survey, women were significantly higher on MH and MCS, whereas men were significantly higher on BP for MO/MB. Overall, the phone survey data showed greater variation than those for MO/MB.
Ryan, Corry, Attewell, & Smithson (2002)	Healthy high school and university students and staff, senior citizen club members, and chronic pain patients in Australia	101	SF-36	Paper form, electronic (unspecified)	Scale score differences (electronic minus paper) were less than 4% and ranged from -2.8 (SF) to 3.9 (RE); the paper SF score was significantly higher ($p < .05$) than the electronic SF score; order effects were noted for VT and MH (with lower scores obtained on the version administered first). For item-level comparisons, quadratic kappa coefficients ranged from .64 to .93; exact agreement of item responses ranged from 49% to 93%; global agreement of responses (within one response category of the other response) ranged from 87% to 100%.
Saleh, Radosevich, Kassim, Mousa, Dykes, et al. (2002)	Orthopedic (knee/hip) patients	87	SF-36	MO/MB, palmtop computer	No significant differences in mean squares, <i>SDs</i> , floor or ceiling percentages, or ICC retest correlations were found (coefficients were based on the results of the second assessment). Cronbach's alphas were significantly different for both the PF and SF scales for the two groups, with greater internal consistency with the paper form.

Table 4.2 (continued)

Effects of Method of Data Collection on Short Form Results: Findings From Selected Studies

Investigation	Sample	N	Short Form Survey	Methods of Data Collection	Key Findings
Suris, Borman, Lind, & Kashner (2007)	VA medical and psychiatric patients	97	SF-36	Paper-and-pencil, computer	No mode of administration bias was found for any of the three study groups (paper-paper [PP], computer-computer [CC], paper-computer [PC]). The authors reported comparability of PP with CC and PP with PC, based on mean domain scale and Health Transition ICCs (2-week) for PP, CC, and PC (.57, .61, .52, respectively), with no clear pattern of differences. Results were noted to suggest equivalency between the two administration modes for this population.
Ware, Kosinski, DeBrotta, Andrejasich, & Bradt (1995)	Patients recruited at ambulatory care facilities and nonmedical business work sites	525	SF-36	MO/MB, telephone interview, IVR	Preliminary results revealed no differences in data quality or tests of scaling assumptions across the three administration methods. Average PCS scores did not differ by method. However, average MCS scores were more favorable (by 1.8 T-score points, $p < .01$) with the telephone interview compared to both self-administered and IVR-administered surveys. The latter two methods did not differ.
Ware, Kosinski, Dewey, & Gandek (2001)	U.S. general population	750, 768, & 907	SF-36v2, SF-8	MO/MB, Internet, telephone interview	Five health domain scale differences (PF, BP, GH, VT, and MH) were significant, with higher average scores for phone interviews (1.2–3.75 T-score points) in comparison with the MO/MB method. PCS and MCS scores were also significantly higher for those interviewed by telephone. Comparisons for the Internet and MO/MB samples found no significant differences for the health domain scales, PCS, or MCS. However, significant differences were found in other studies that compared the results of computer administration and paper-form administration of the SF-36 along with disease-specific health status surveys.
Weinberger, Oddone, Samsa, & Landsman (1996)	Outpatient veterans	172	SF-36	MO/MB, face-to-face interview, telephone interview	With one exception, all eight scales achieved an acceptable level of internal consistency (.70); average scale test-retest correlations for the 3 modes ranged from .79 to .83; and scale score differences between modes were not significant. With a few exceptions, scale ceiling and floor percentages were approximately the same. Average scale correlations between modes of administrations ranged from .74 to .81.
Wilson et al. (2002)	British rheumatology outpatients	51	SF-36	Paper-and-pencil, desktop computer	Cronbach's alphas for all but the BP and SF scales ranged from .83 to .94 for the paper version and from .84 to .95 for the computer version, while Spearman's ρ for BP and SF ranged from .68 to .97 (all $ps < .01$). Scale score correlations for the two modes ranged from .80 (SF) to .96 (PF); 67% of the respondents preferred the computerized version, with significantly more of the younger group (< 47 years old) than the older group having this preference.
Wood & McLauchlan (2006)	Elderly British patients who underwent total hip arthroplasty	90	SF-36	MO/MB, face-to-face interview	The interview sample's RP and RE scores were significantly higher ($ps < .01$) than those for the MO/MB sample.



5

Scoring Procedures

Originally, SF-36 health domain scale raw scores were transformed to scores ranging from 0 to 100. Using this metric, 0 represented the lowest possible score (worst health state) and 100 represented the highest possible score (best health state), with scores in between representing the percentages of the total possible score achieved by respondents on a given scale. The PCS and MCS measures, however, have been scored using norm-based *T* scores since their publication in 1994 (Ware & Kosinski, 2001b). Subsequently, the healthcare research field has evolved and comparisons between health domain scales and component summary measures have become important. Because the two different scoring systems did not facilitate direct comparisons, procedures for scoring all health domain scales and component summary measures using the *T*-score metric were developed as an alternative to the 0–100 scoring metric (Ware & Kosinski, 2001b). Thus, with the development of the SF-36v2 came the development of *T* scores for all the health domain scales and the component summary measures, as well as the ability to make direct comparisons between the two. (Please see Chapter 13 for further explanation of the *T*-score scoring method.)

This chapter provides an overview of the scoring instructions for the SF-36v2's eight health domain scales, PCS and MCS measures, and Self-Evaluated Transition (SET) item. First, the importance of maintaining standardization in survey content and scoring is discussed. This is followed by general scoring information for the health domain scales and steps for data entry and scoring that are common to all items. Next, a description of procedures for item aggregation and transformation of health domain scale raw scores to a 0–100 metric is presented, followed by a description of how 0–100 scores are transformed to *T* scores. An overview of the scoring procedures for the PCS and MCS measures is then presented, followed by information regarding the optional Response Consistency Index (RCI). Finally, this

chapter concludes with a brief description of the Short Form scoring software and services that are available from QualityMetric Incorporated and its authorized resellers, including the scoring of the survey's measures and scales, either with or without the application of Missing Score Estimation (MSE) procedures, and scoring of data quality evaluation (DQE) indicators.

Note that guidelines for evaluating the quality of SF-36v2 data and verifying the accuracy of scoring are presented in Chapter 6 of this manual. Furthermore, issues discussed in this chapter that are related to the scoring of the PCS and MCS measures (e.g., use of oblique vs. orthogonal solutions in defining the components) and the health domain scales (e.g., recalibrations and/or dependencies for the BP and GH scale items), as well as issues concerning deviations from the standardized scoring steps, are addressed in Chapter 13.

Importance of Standardization

Standardization of content and scoring is what makes possible the valid and reliable interpretation of SF-36v2 health domain scales and component summary measures. The survey's content and the scoring algorithms used were selected and standardized following careful study of many options. The algorithms selected and described in this chapter were designed to be as simple as possible, to satisfy the assumptions of the methods used to construct SF-36v2 health domain scales and component summary measures, and to maximize comparability between SF-36v2 and SF-36 scores throughout their in-common range, in order to preserve the original interpretations of the scales and measures.

The SF-36v2 utilizes norm-based scoring involving a linear *T*-score transformation method so that scores for each of the health domain scales and component summary measures have a mean of 50 and a standard

deviation of 10, based on the 2009 U.S. general population. Thus, scores above and below 50 are above and below the average, respectively, in the 2009 U.S. general population. Also, because the standard deviation is 10, each 1-point difference or change in scores has a direct interpretation; that is, 1 point is one-tenth of a standard deviation, or an effect size of .10. (See Chapter 13 for further discussion of the advantages of the *T*-score metric over the 0–100 scoring metric.)

There are two important reasons to adhere to the content and scoring standards described in this manual. First, doing so is most likely to produce scores with the same reliability and validity as those previously reported for SF-36v2 health domain scales and component summary measures. Be aware that making changes to the survey's content or scoring methods may compromise the reliability, validity, and interpretation of obtained scores. Second, deviating from the content and scoring standards will likely produce scores sufficiently biased as to invalidate normative comparisons and to prevent comparisons of results across studies. In short, standardization allows differences in scores to have the same interpretations across studies.

It is important to note that mean scores obtained from the 2009 norms vary from those based on the 1998 normative data for most of the SF-36v2 health domain scales (see Chapter 14). Many of these differences are statistically significant but not very meaningful; regardless, this underscores the importance of using the most up-to-date SF-36v2 norms that were collected in 2009.

The SF-36v2 uses the same factor score coefficients

as the SF-36 to score the PCS and MCS measures. Because the original “recipe” for aggregating the health domain scales has been preserved, the PCS and MCS scores of the two SF-36 versions are highly comparable. For both the health domain scales and component summary measures, *T* scores based on 2009 norms simply shift the score distribution to better reflect the health of the U.S. population in 2009. Otherwise, 2009 scores have the same interpretations as 1998 scores.

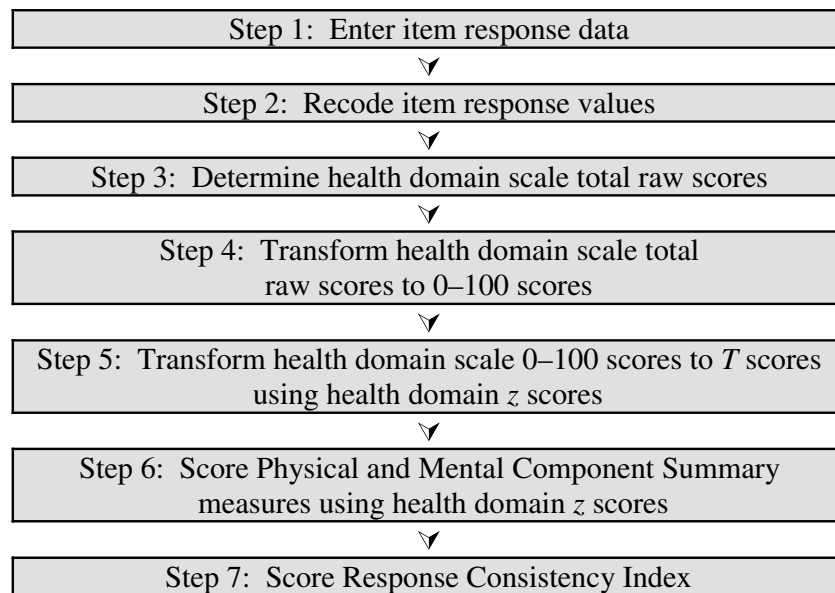
Prior to applying the scoring rules, it is essential to verify that the questionnaires being scored—including the questions asked (item stems), response choices, and values assigned to response choices at the time of data entry—have been exactly reproduced. The scoring rules described in this chapter apply to the questions, response choices, and number values assigned to said response choices on the SF-36v2 standardized paper forms, regardless of the application for which the instrument is being used. Modifying forms, such as by changing item wording or by omitting items or response categories, can result in scores that are invalid.

SF-36v2 users with questions about using nonstandard item wording or scoring procedures should contact QualityMetric Incorporated..

Scoring the SF-36v2

This section presents an overview of the process for obtaining norm-based *T* scores for the SF-36v2 health domain scales and component summary measures. All

Figure 5.1 Process for Scoring SF-36v2 Health Domain Scales and Component Summary Measures



items, scales, and summary measures are scored so that a higher score indicates a better health state. The scoring process is summarized in Figure 5.1.

Step 1: Entering Data

Scoring the SF-36v2 begins with ensuring that the survey form is complete and the respondent's answers are unambiguous. It is not uncommon for those scoring the SF-36v2 to encounter completed test forms that have scoring problems. Such problems can be avoided by quickly scanning the survey form and assuring that the respondent's intended answers are clear before he or she leaves the room. Unfortunately, this is not always possible when conducting group administrations of the survey, when an individual other than the administrator is responsible for scoring the survey, or when the survey is completed by mail-out/mail-back (MO/MB) administration. Some of these problems are alleviated when the instrument is administered via desktop, online, or a hand-held device that prevents respondents from making errors and automatically submits entered responses for scoring. When the SF-36v2 is administered via paper form but scored by software, however, it is important that the item responses are valid and that they are entered into the scoring program as intended and as coded on the survey form.

The following are three common problems that the administrator should be aware of before submitting an SF-36v2 response set for scoring.

Items with out-of-range response values. In instances where item values are entered into an electronic data file, all 36 items should be checked for out-of-range response values prior to assigning the final item values. Out-of-range values are those that are lower than an item's precoded minimum value or higher than an item's precoded maximum value. Usually caused by data-entry errors, out-of-range values should be changed to the correct value through verification with the original survey. If the survey is not available, any out-of-range values should be treated as missing data.

Missing item responses. Sometimes respondents do not answer one or more SF-36v2 items, albeit a generally infrequent occurrence (1–2% of the time or less). One important advantage of multi-item scales is that a scale score can be estimated even when responses to some of its items are missing. By using a scoring algorithm that estimates missing values, it is usually possible to derive scores across the eight SF-36v2 health domain scales for nearly all survey respondents. Historically, the instrument's developers have advocated for the *Half-Scale Rule*, which states that a score can be calculated if the respondent answers at least

50% of the items in a multi-item scale. In such cases, the recommended algorithm substitutes an estimate for the missing item data that is based on the respondent's answers to the other items.

While the Half-Scale Rule has served the field well when dealing with missing data, the progress that has been made in understanding missing data has led to better methods for and more confidence in handling missing data. For group-level analyses, a review of studies using SF-36 data has indicated that a health domain scale score can be estimated even when only *one* item in said scale is answered. Using the *Full Missing Score Estimation (Full MSE)* method, the missing item responses in a given scale are assumed to be the same as the response to the scale's answered item and the final item response values are then assigned accordingly. Note that this approach should not be used to estimate item responses on the PF scale due to the hierarchical nature of its items. When necessary, the PF scale score can be estimated using item response theory (IRT), which is utilized by the QualityMetric Health Outcomes™ Scoring Software 5.0 (Saris-Baglana et al., 2011) which is discussed at the end of this chapter.

Regardless of the method employed when dealing with missing data, be aware that respondents who don't answer all the survey's items are often individuals who are in poor health and that correlates of scores are developed without the contribution of data from this group. Therefore, be mindful that conclusions drawn from estimated scores may be based on correlates derived from the responses of individuals who differ from the respondent in important ways.

Single items with multiple responses. Sometimes a respondent is careless or cannot decide among the response choices for a given question. If the SF-36v2 is administered via paper form, a respondent may indicate two or more responses in an effort to convey what he or she considers to be an accurate answer. When a respondent provides multiple responses to a single item, apply the following rules to each item with multiple responses before submitting the survey for scoring:

1. If a respondent marks two responses that are adjacent to each other, randomly pick one, and enter that number.
2. If a respondent marks two responses that are not adjacent to each other, consider that item missing.
3. If a respondent marks three or more responses, consider that item missing.

Alternatively, one can opt to treat all items with more than one response as missing.

Step 2: Recoding Item Response Values

The next step after data entry is to recode the item response choices, a process that derives the final item response values, or scores, to be used when calculating the raw scale score for each health domain. Several steps are included in this process, including (a) changing out-of-range values to missing, (b) recoding values for 10 items, and (c) substituting person-specific estimates for missing items.

Table 5.1 presents response value recoding information for one of the SF-36v2 items (Item 8, from the Bodily Pain scale), including the response choices for the item, the precoded response values printed on the survey form, and the final response values that are used for scoring the item (i.e., the recoded values). Note that for all 36 items, the precoded response values for each item correspond to both the standard and acute forms. As demonstrated in Table 5.1, the precoded value associated with a given response choice may not match its recoded response value. When entering data, it is important to enter the precoded response value for each survey item. QualityMetric Incorporated's Health Outcomes Scoring Software 5.0 and online scoring services automatically assign final (i.e., recoded) response values after the administrator enters the precoded response values.

Note that there are scoring differences amongst the survey's 36 items. First, 10 of the items are reverse scored, a method that is used to ensure that higher item

Table 5.1

Bodily Pain Item 8 Response Choices and Scoring Information

Response Choice	Precoded Response Value	Final Response Value
Not at all	1	5
A little bit	2	4
Moderately	3	3
Quite a bit	4	2
Extremely	5	1

values indicate better health on all the items, health domain scales, and component summary measures. Therefore, SF-36v2 items that require reverse scoring are those that are worded such that a higher precoded item value indicates a poorer health state. Second, the procedure used to determine final item values vary depending on the item. For 34 of the items, research to date supports the assumption of a linear relationship between item scores and the underlying health construct defined by their scales. However, as discussed in Chapter 13 of this

manual, empirical work has shown that two items (one each from the GH and BP scales) require recalibration to satisfy this important scaling assumption.

The Self-Evaluated Transition (SET) item does not require recoding of its response values because it is not scored as part of any SF-36v2 scale or measure. Responses to this item are treated as ordinal level data that can be used to analyze the percentage of respondents who select each response choice or to estimate the measured change (observed changes in health domain scale scores) reported for each response category.

Step 3: Determining Health Domain Scale Total Raw Scores

After item recoding, which includes resolving items with missing data, a *total raw score* is then computed for each health domain scale. The total raw score is the simple algebraic sum of the final response values for all the items in a given scale. For example, the total raw score for the RP scale is the sum of the final response values (i.e., recoded response values or, when applicable, imputed values) for items 4a, 4b, 4c, and 4d. This simple scoring method is possible because all the items in a given scale have roughly equivalent relationships to the underlying health construct being measured and because no item is used on more than one scale. As a result, it is not necessary to standardize or weight items. Note that these assumptions have been extensively tested and verified for both the SF-36 and SF-36v2 (McHorney, Ware, & Raczek, 1993; Ware, Kosinski, & Keller, 1994; Ware et al., 2007; see also Chapter 13).

Step 4: Transforming Health Domain Scale Total Raw Scores to 0–100 Scores

The next step when scoring the health domain scales involves transforming each total raw scale score to a 0–100 scale score using the following formula:

$$\text{Transformed scale score} = \frac{(\text{Actual raw score} - \text{Lowest possible raw score})}{\text{Possible raw score range}} \times 100$$

For example, a Physical Functioning total raw score of 21 would be transformed as follows:

$$\frac{(21 - 10)}{20} \times 100 = 55$$

As shown, the lowest possible PF score equals 10 and the possible PF total raw score range equals 20. This transformation converts the lowest and highest possible raw scores to 0 and 100, respectively. Scores between

these values represent the percentage of the total possible score achieved.

Note that raw and transformed scale scores are not calculated for the SET item. As previously indicated, responses to this item should be treated as ordinal-level data. The SET item can also be used as an interval-level scale or as a categorical variable (descriptor).

Step 5: Transforming Health Domain Scale 0–100 Scores to *T* Scores

This step involves transforming each 0–100 scale score to a *T* score using the standard score formulas. As previously mentioned, the advantages of standardizing the health domain scales and converting 0–100 scores to norm-based scores using a *T*-score transformation (see Anastasi, 1988) are that health domain scale results can be meaningfully compared with each other and that these scale scores have a direct interpretation in relation to the distribution of scores in the 2009 U.S. general population. For more information regarding the advantages of using the *T*-score metric for the SF-36v2 health domain scales and component summary measures, please see Chapter 13 of this manual.

Transforming 0–100 scores to *z* scores. The first step in calculating *T* scores consists of standardizing each SF-36v2 health domain scale using a *z*-score transformation. A linear *z*-score transformation is used so that each health domain scale has a mean of 0 and a standard deviation of 1 in the 2009 U.S. general population. A *z* score is computed by subtracting each health domain scale's 2009 U.S. general population mean from the 0–100 score for that scale, and then dividing the difference by the given scale's standard deviation. Thus, using 1998 scoring algorithms, the formula for computing the *z* score for the standard form Physical Functioning scale, is as follows:

$$PF\ z = (PF - 83.29094) \div 23.75883$$

Note that, in the above formula, PF represents the 0–100 score for that scale.

Transforming *z* scores to *T* scores. This step transforms each *z* score to a *T* score (mean = 50, *SD* = 10). To do so, multiply each *z* score by 10, and then add 50 to the resulting product. The formula for computing the *T* score for each health domain scale, once again illustrated using the PF scale, is:

$$PF\ T\ score = 50 + (PF\ z \times 10)$$

Step 6: Scoring the Physical and Mental Component Summary Measures

After scoring the eight health domain scales using the SF-36v2 *z*-score formulas presented in Step 5, the Physical Component Summary (PCS) and Mental Com-

ponent Summary (MCS) measures are then scored using a three-step procedure, regardless of whether a standard or acute form was administered. First, the eight health domain scales are standardized using means and standard deviations from the 2009 U.S. general population. Second, these standardized scores are aggregated using weights (factor score coefficients) from the 1990 U.S. general population (see Chapter 13 for more information regarding the use of 1990 factor score coefficients). These are the same weights as those used to score the SF-36 PCS and MCS measures (Ware et al., 1994; see also Chapter 13) and as those used to score the SF-36v2 with 1998 norms (see Ware et al., 2007). Third, aggregate PCS and MCS scores are standardized using a linear *T*-score transformation with a mean of 50 and a standard deviation of 10.

Note that the same factor score coefficients are used to score the PCS and MCS measures for both standard and acute forms. The process of deriving *T* scores from the SF-36v2 standard (4-week recall) and acute (1-week recall) forms is presented in the following sections.

Aggregation of scale scores. The first step in computing PCS and MCS scores involves computing aggregate scores using the physical and mental factor score coefficients from the 1990 U.S. general population and the *z* scores previously computed for each of the eight health domain scales (see Step 5). Computation of an aggregate physical component score consists of multiplying each SF-36v2 health domain scale *z* score by its respective physical factor score coefficient and then summing the eight products. Similarly, an aggregate mental component score is obtained by multiplying each SF-36v2 health domain scale *z*-score by its respective mental factor score coefficient and summing the eight products. To illustrate, a portion of the formula for aggregating scales when estimating a standard form aggregate mental component score is as follows:

$$\text{Aggregate mental component score} = \\ (PF\ z \times -.22999) \dots + (MH\ z \times .48581)$$

Transforming summary scores to *T* scores. The second step involves transforming each aggregate component score to a *T* score. This is accomplished by multiplying each aggregate component scale score by 10, and then adding 50 to the resulting product. The formulas for computing the norm-based *T* score for each component summary measure are:

$$PCS\ T\ score = \\ 50 + (\text{Aggregate physical component score} \times 10)$$

$$MCS\ T\ score = \\ 50 + (\text{Aggregate mental component score} \times 10)$$

PCS and MCS Missing Score Estimation. Using Full MSE, a PCS score can be computed if at least seven scales have been scored, one of which being the PF scale; similarly, an MCS score can be computed if at least seven scales have been scored, one of which being the MH scale. Users wishing to take advantage of the MSE procedures that are available for the SF-36v2 can do so by scoring their data using the scoring software or services offered by QualityMetric Incorporated and its authorized resellers.

Step 7: Scoring the Response Consistency Index

One of the many SF-36v2 data quality indicators available is the *Response Consistency Index (RCI)*. Scoring the RCI is optional; however, doing so is a simple and easy way to evaluate the consistency of responses to individual survey items. The RCI comprises 15 pairs of items and assesses each pair for consistency. If a pair of responses is consistent, then the RCI score for that pair would be 0. Conversely, a pair of inconsistent responses would earn a score of 1. For example, if a respondent indicates that he or she can “walk more than a mile” but, at the same time, cannot “walk 100 yards,” then this item pair would be considered inconsistent and would earn 1 RCI point. For a given respondent, the final RCI score is the sum of the scores earned on the 15 consistency checks. Thus, the best (i.e., most consistent) RCI score is 0 and the worst (i.e., least consistent) score is 15. Note that it is not necessary for a respondent to have complete data for all 15 pairs to compute the RCI (pairs with missing or out-of-range data are not used in the final calculation). However, if all 15 pairs have missing data for one or both items, then the RCI for that respondent cannot be scored. For additional information regarding the RCI, please see Chapter 6 of this manual.

Scoring Software and Services

QualityMetric Incorporated offers a variety of ways to score the SF-36v2. The following sections briefly discuss these options.

Smart Measurement System

The Smart Measurement™ System is a convenient, all-in-one, Internet-based, health survey data collection service that uses the latest technologies to capture, benchmark against general and disease-specific norms, and interpret survey data. This information technology platform is ideal for individuals and organizations that want to quickly and confidentially measure functional

health and well-being, all while obtaining results in real time.

The Smart Measurement System features include

- automatic scoring of surveys, with real-time reports;
- reporting that tracks changes in health over time and makes comparisons between treatments, programs, respondents, and populations;
- access via confidential login at any time, from any location where Internet access is available;
- multi-user capability that allows several respondents/administrators to log into the system and complete tasks simultaneously;
- multiple administration modes, including paper-and-pencil, online, Smartphone, and more;
- an optional automated respondent reminder system that uses e-mail and postal mail to increase survey completion compliance;
- administration management tools for sponsors, groups, sites, and individuals;
- data warehousing for storage and recall of completed surveys;
- data import/export capabilities with customer sites using secure FTP connections;
- compliance with FDA 21CFR Part 11, HIPAA (U.S.), and PIPA (Canada) privacy and security regulations for electronic data capture of ePROs.

In addition, the Smart Measurement System can be used via an interface that makes it appear to respondents that they never leave the host website. When using this feature, a link is created on the host site that connects to QualityMetric Incorporated’s Smart Measurement System and respondents are provided with a “single sign-on” to take the survey.

For more information about the Smart Measurement System, please visit <http://www.qualitymetric.com>.

Health Outcomes Scoring Software 5.0

QualityMetric’s Health Outcomes™ Scoring Software 5.0 is available to score the SF-36v2 and some of its associated health outcomes instruments. This software is designed to provide standardized scoring methods via an easy-to-use system centered around projects, giving users confidence that their SF-36v2 data have been scored in accordance with the standards set by the developers of the survey. In addition, the scoring software evaluates data quality, applies missing score recovery methods, and has other optional features. The system provides several options for importing raw data (e.g., CSV, Fast Data Grid, Form Entry). Once captured, the raw data are scored and securely saved for later use.

Health Outcomes Scoring Software 5.0 sample reports for individual patient data are presented in Appendix A

of this manual. Sample reports for group-level data are presented in Appendix B.



PART III: INTERPRETATION



6

Data Quality Evaluation

The importance of routinely evaluating the quality of data obtained from administrations of the SF-36v2, or any psychometric measure, cannot be overemphasized. With a complete evaluation of data quality, users can more readily identify the sources of and correct any problems, or at least take them into account when conducting analyses. For the SF-36v2, there are several *quantitative* checks that can be performed to determine the quality of the obtained data. These include: (a) completeness of data, (b) responses within range, (c) consistent responses, (d) percentage of estimable scale scores, (e) item internal consistency, (f) item discriminant validity, (g) scale reliability, and (h) confirmation of the two-component structure. All of these data quality checks are discussed in this chapter, and all but the last are performed by the QualityMetric Health Outcomes Scoring Software 5.0 (Saris-Baglama et al., 2011; see also Chapter 5). This chapter also addresses the following *qualitative* checks: (a) responses inconsistent with respondent presentation, (b) unusually quick or long completion time, and (c) patterned responses. The purpose of this chapter is to discuss each of the quantitative and qualitative data quality indicators, what each may indicate, and how associated problems, when identified, can be resolved.

Considerations for Analyzing Data From Groups of Respondents or Multiple Administrations to the Same Respondent

Before applying the general quality assurance procedures described in latter sections of this chapter, other considerations should be taken into account when entering SF-36v2 data for groups of respondents or data from multiple assessments of a single respondent.

Combining and Analyzing Data From Standard and Acute Forms

Caution should be taken when combining and interpreting data gathered from the SF-36v2 standard and acute forms. Generally, the results from administrations of the two forms substantially agree. However, users may sometimes find that results from the acute form will differ from those obtained from the standard form. Keller et al. (1997), for example, found that the effect of the form approached significance ($p = .08$) with two small samples of asthma patients participating in a controlled study that used the SF-36 to examine the effects of inhaled corticosteroid on HRQOL. In addition, univariate analyses revealed higher scores on the 0–100 scoring metric from the SF-36 acute form, with RE scores averaging nearly 7 points higher ($p = .05$), RP scores averaging nearly 5 points higher, and SF scores averaging nearly 3 points higher. It is important to note, however, that these findings were not replicated in a U.S. general population sample during the 1998 norming of the SF-36v2, an effort that reported cross-sectional health domain scale scores from the standard and acute forms were very similar. Results from the Keller et al. study are probably more relevant in the context of a randomized clinical trial in which changes in health status can occur relatively quickly; therefore, the cautionary note previously stated should be kept in mind for other acutely ill patient samples as well.

Combining and Analyzing Data From Different Data Collection Methods

Data collection should always be limited to one method of administration (e.g., online, paper form) if SF-36v2 data from groups of respondents are to be aggregated. When data collection methods do vary within a sample or when results are compared across samples assessed using different methods, the effects

of the methods used should be evaluated and the results interpreted with due caution (see Chapter 4).

Combining and Analyzing Data From Different Translated Forms

It has been well documented that data from translated forms of the SF-36 can be aggregated and successfully analyzed in clinical trials. The most comprehensive and thorough tests of the equivalence of such translations, as well as formal tests of the psychometric assumptions underlying their scoring and interpretation in such combined analyses, were published in a special issue of the *Journal of Clinical Epidemiology* that documented dozens of empirical evaluations of SF-36 translations that were performed during the International Quality of Life Assessment (IQOLA) Project (Gandek & Ware, 1998b).

As of 2008, 28 peer-reviewed publications reporting results from clinical trials that used one or more SF-36 translations had been identified and more than two-thirds of new and ongoing clinical trial protocols included SF-36 forms in two or more languages. At this rate, it appears that this approach will continue for years to come. In general, the authors of this manual know of no evidence that language-related differences in SF-36/SF-36v2 results are any larger than differences found between study sites within the same country using same-language forms. Consequently, the authors recommend that any such differences found using different-language forms be handled in the same manner as those obtained using same-language forms. However, to ensure that data entry has been properly performed and that data

quality satisfies minimum standards, it is recommended that the indicators discussed in this chapter be evaluated separately for English- and non-English-language forms whenever possible.

Quantitative Evaluation of Data Quality

When analyzing SF-36v2 aggregated group-level data, there are eight quantitative checks (summarized in Table 6.1) that can be performed to determine the quality of those data. These quantitative checks are: (a) completeness of data, (b) responses within range, (c) consistent responses, (d) percentage of estimable scale scores, (e) item internal consistency, (f) item discriminant validity, (g) scale reliability, and (h) confirmation of the two-component structure. Note that the quantitative data quality indicators discussed in this section should be used only when evaluating the quality of SF-36v2 data for groups of at least 30 respondents.

Those evaluating the quality of SF-36v2 data should be aware that the quantitative checks discussed in this chapter were developed for use with group data; however, some are also appropriate for use with individual respondent data. Conversely, the qualitative checks are more appropriately and easily applied to individual respondent data. It should also be noted that when evaluating the quality of group data, analyses should not be limited to only the results of the combined total sample as other units of analysis that may reveal data quality problems should be considered. A logical unit of analysis would be anything

Table 6.1
SF-36v2 Quantitative Data Quality Indicators

Data Quality Indicator	Description	Minimum Satisfactory Value
Completeness of data	Percentage of the total number of items with valid item responses	90%
Responses within range	Percentage of the total number of completed items that have responses within the acceptable range for all completed SF-36v2 forms	100%
Consistent responses	Percentage of respondents with a Response Consistency Index (RCI) score of 0	90%
Percentage of estimable scale scores	Percentage of health domain scale and component summary measure scores that are computable using either of two approaches (Full Missing Score Estimation or Complete Data)	90%
Item internal consistency	Percentage of correlations between items and their hypothesized scales that are .40 or greater	90%
Item discriminant validity	Percentage of hypothesized item-scale correlations that are higher than the alternative item-scale correlations	80%
Scale reliability	Percentage of the health domain scales that have Cronbach's alphas of .70 or greater	100%
Confirmation of the two-component structure	Degree to which correlations between the health domain scales and component summary measures confirm: (a) the two-component structure of the SF-36v2 in a manner that is generally consistent with what has been found in the U.S. general population and other developed countries and (b) that each health domain scale has its intended interpretation as a measure of physical or mental health status	Informed judgment of the clinician/researcher

that could cause or contribute to such problems. Examples of other ways to evaluate the quality of SF-36v2 data are by mode of administration, language of the respondents, site of survey administration, baseline and follow-up administrations, and sociodemographic subgroups.

Completeness of Data

The first data quality indicator is *completeness of data*. To evaluate this data quality indicator:

1. Determine the number of items that have valid responses for all completed SF-36v2 forms.
2. Divide the total number of items with valid responses (Step 1) by the total number of possible survey responses for the group (36 x number of respondents).
3. Multiply the result (Step 2) by 100 to determine the percentage of items completed.

Data quality is considered satisfactory for this indicator when the result is at least 90%. For example, if 10 respondents complete the SF-36v2 and each of four respondents has four responses that are missing or out-of-range, then a total of 16 items are considered incomplete. Accordingly, the total number of possible responses is 360 (36 items per form x 10 respondents) and the number of items with valid responses is 344 (360 total items – 16 incomplete items). Thus, the completeness of data result for this example is: $(344/360) \times 100 = 95.6$, or 95.6%. This would be classified as a satisfactory result.

When collecting completed surveys, administrators should closely review each item that is missing data, particularly when a significant number of items have missing data, when the items with missing data tend to be those that are presented at particular points in the survey (e.g., before or after a page break in the form), or when the items with missing data are from a particular health domain scale. Causes of missing data vary and should be investigated to ensure the integrity of SF-36v2 data. For example, individual items with substantial missing data may indicate that a group of respondents as a whole had difficulty understanding them. Alternatively, a data entry problem or a formatting problem may have caused the problem. Finally, regardless of the ability to apply MSE scoring corrections, data missing from a significant proportion of items from a single health domain scale may suggest problems or concerns regarding functioning in that domain.

Responses Within Range

The second data quality indicator is *responses within range*. To determine the percentage of responses within the allowable limits:

1. Count the number of items that have responses within the acceptable range for all completed SF-36v2 forms. (Note that items with missing responses should *not* be included.)
2. Divide the total number of items with in-range responses (Step 1) by the total number of possible survey responses for the group (36 x number of respondents).
3. Multiply the result (Step 2) by 100 to determine the percentage of responses within range.

This data quality indicator is considered satisfactory only when the result is 100%. For example, if 10 respondents complete the SF-36v2 and the group has a total of nine out-of-range responses, then the total number of possible responses is 360 (36 items per form x 10 respondents) and the number of items scored within range is 351 (360 total items – 9 out-of-range item responses). Thus, the percentage of responses within range is: $(351/360) \times 100 = 97.5$, or 97.5%. This would not be considered a satisfactory result.

Each item with an out-of-range value should be closely reviewed to determine the cause and, when possible, correct the error. Whether random or systematic in nature, likely causes include data entry or data formatting errors made by users and data recording errors made by respondents. For example, an isolated data entry error could cause one item's responses to be submitted for another item for a small portion of the sample. Or, a systematic formatting error could cause data for a given variable to be entered into the wrong column, thereby shifting by one column the data for all subsequent variables for the entire sample. Another cause of out-of-range responses may be the use of incorrect scoring algorithms, which would result in misscored data. When this is suspected, the data file should be rechecked to ensure that the algorithms used are consistent with those described in Chapter 5 of this manual.

When out-of-range responses occur, administrators should, when possible, obtain the correct values from the original surveys and correct the out-of-range values. If this is not possible, convert the out-of-range values to missing so that the incorrect data is not scored.

Consistent Responses

The third data quality indicator is *consistent responses*, which is objectively measured using the Response Consistency Index (RCI; see Chapter 5). To evaluate the RCI for group data:

1. Determine the number of respondents who have an individual RCI score of 0 (i.e., has consistent responses for all 15 item pairs; see Chapter 5).

2. Divide the number of respondents with RCI scores of 0 (Step 1) by the total number of respondents in the data set.
3. Multiply the result (Step 2) by 100 to determine the percentage of respondents with an RCI score of 0.

This data quality indicator is considered satisfactory when the result (i.e., the percentage of respondents with an RCI score of 0) is at least 90%. For example, if eight respondents complete the SF-36v2 and four respondents earn an RCI of zero, then the group's RCI score is: $(4/8) \times 100 = 50$, or 50%. This would not be considered a satisfactory result.

Using the RCI to evaluate the consistency of a group's responses is helpful because it offers a quick glance into potential sources of information about the group's respondents and/or data entry or scoring problems. For example, a small percentage of consistent responses may indicate that items were already reversed scored or were mislabeled, which would warrant a rechecking of the data set to determine if data entry problems occurred. Or, if 20% of a given sample has inconsistent responses, administrators would be wise to identify the offending item pairs to determine whether they reflect errors in the testing process or insights into the well-being of the respondents.

Tables 6.2 and 6.3 present the RCI score frequency distributions for the SF-36v2 standard (4-week) and acute (1-week) forms, respectively, based on the 2009 U.S. general population normative sample. Note that the higher the RCI score, the more inconsistent the respondent was in his or her responses to survey items. For each of the two forms, approximately 94% of the

U.S. general population sample responded consistently to all 15 item pairs.

While it may be acceptable to include those surveys that contain one or two inconsistent responses, users may want to consider excluding respondents whose surveys containing multiple inconsistencies before scoring the data. It is possible that these particular respondents did not understand the items or did not carefully read and respond to the items. Also, if data were collected and entered at different sites, users should determine whether any inconsistencies are contained within a particular subset of the data and, if so, recheck that subset for data entry problems. An example of RCI use can be found in Hanscom, Lurie, Homa, and Weinstein's (2002) examination of differences in missing-response rates and response consistency between computerized and paper-and-pencil versions of the SF-36 (see Chapter 4).

Percentage of Estimable Scale Scores

The fourth data quality indicator is *percentage of estimable scale scores*. Calculating this indicator can be achieved using either of two approaches: *Complete Data* or *Full Missing Score Estimation (Full MSE)*. The Complete Data approach utilizes health domain scale and component summary measure scores that are computed using only the respondent's available scores (i.e., none of the item values have been estimated). In contrast, the Full MSE approach utilizes a combination of the respondent's available health domain scale and component summary measure scores and scores computed using estimated response values (see Chapter 5).

This data quality indicator reports the percentage of SF-36v2 scales and measures that can be scored, regardless of how the scores were calculated (i.e.,

Table 6.2

SF-36v2 Standard (4-Week Recall) Form Response Consistency Index (RCI) Frequencies, 2009 U.S. General Population (N = 4,024)

RCI	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
0	3,750	93.19	3,750	93.19
1	162	4.03	3,912	97.22
2	37	0.92	3,949	98.14
3	25	0.62	3,974	98.76
4	29	0.72	4,003	99.48
5	6	0.15	4,009	99.63
6	8	0.20	4,017	99.83
7	1	0.02	4,018	99.85
8	6	0.15	4,024	100.00

Note. Includes cases with a PCS or MCS score.

Table 6.3

SF-36v2 Acute (1-Week Recall) Form Response Consistency Index (RCI) Frequencies in the 2009 U.S. General Population (N = 2,056)

RCI	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
0	1,946	94.65	1,946	94.65
1	66	3.21	2,012	97.86
2	16	0.78	2,028	98.64
3	7	0.34	2,035	98.98
4	16	0.78	2,051	99.76
5	2	0.10	2,053	99.85
6	2	0.10	2,055	99.95
7	0	0.00	2,055	99.95
8	1	0.05	2,056	100.00

Note. Includes cases with a PCS or MCS score.

available and/or estimated scores). Therefore, to evaluate the percentage of estimable scale scores using either approach:

1. Count the number of available and estimated health domain scale and component summary measure scores for all completed SF-36v2 forms.
2. Divide the total number of available and estimated scores (Step 1) by the total number of possible scale and measure scores (10 x number of respondents).
3. Multiply the result (Step 2) by 100 to determine the percentage of estimable scale scores.

Table 6.4 presents the number of completed items required for each health domain scale score when using each of these methods. For the Complete Data method, both the PCS and MCS measures require scores for all eight health domain scales. Meanwhile, using the Full MSE approach to estimate the PCS measure requires scores for seven scales, one of which must be the PF scale. Similarly, estimating the MCS score also requires scores for seven scales, one of which must be the MH scale.

Table 6.4

Number of Completed Items for Each SF-36v2 Health Domain Scale Required for Each Score Estimation Method

Estimation Method	PF	RP	BP	GH	VT	SF	RE	MH
Complete Data	10	4	2	5	4	2	3	5
Full MSE	1	1	1	1	1	1	1	1

This data quality indicator is considered satisfactory when the result is at least 90%. To illustrate, Table 6.5 presents the health domain scale and component summary measure *T* scores for three respondents. Note that two PF scores and two PCS scores could not be calculated (indicated by -1 in this data set). Using the data found in the Table 6.5 and the steps previously outlined, the total number of possible scales/measures is 30 (10 scales/measures x 3 respondents) and the number of actual scoreable scales/measures is 26 (30 possible scale/measure scores - 4 unscorable scales/measures). Thus, the percentage

of estimable scale scores is $(26/30) \times 100 = 86.7$, or 86.7%. This would not be considered a satisfactory result. However, if the Full MSE method of estimating scores were applied to this example, then a significantly higher percentage of health domain scale and component summary measure scores may be computable.

As previously mentioned, it is important to be aware that particular types of respondents are more likely to have missing SF-36v2 data, such as elderly or less educated respondents. For example, Kosinski, Bayliss, Bjorner, and Ware (2000) reported that in the U.S. general population, 5.77% of non-elderly respondents and 20.68% of elderly respondents had one or more missing items. The percentages were higher for patients in the Medical Outcomes Study (MOS), with 28.66% of non-elderly and 44.11% of elderly respondents having one or more missing items. Kosinski et al. also reported that almost one in four of the Medicare Health Outcomes Study (HOS) respondents in the 1998 cohort had one or more missing items. Because one would not want to bias the sample by excluding these respondents, users may decide that it is important to use missing score estimation to ensure a more representative sample. However, note that if the scores are significantly below the norm, then a large amount of data may be missing and a check of the data set for problems would be warranted.

Item Internal Consistency

The fifth SF-36v2 data quality indicator is *item internal consistency*. When combined with item discriminant validity (see following section), item internal consistency becomes a measure of item convergent validity. Tests of item internal consistency are performed to determine whether the items in a scale are linearly related to the underlying construct. For example, because Item 3a is in the PF scale, then Item 3a should be related to the overall PF scale score, even when the contribution of Item 3a to the scale score is removed.

To evaluate item internal consistency:

1. Examine the correlation between each item and its hypothesized health domain scale score, corrected for overlap (i.e., the item being tested is removed from the scale score before

Table 6.5

SF-36v2 Health Domain and Component Summary Measure Sample Data Set

PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
23.55	46.45	22.32	18.45	28.40	42.59	33.99	51.03	21.61	49.58
-1	52.02	-1	27.99	54.22	45.87	60.93	56.40	21.61	57.89
-1	47.13	-1	25.60	49.22	34.13	51.95	51.03	21.61	49.58

the correlation is computed), for all completed SF-36v2 surveys.

2. Compute the percentage of items that correlate .40 or greater with their hypothesized scales.

This data quality indicator is considered satisfactory when at least 90% of the hypothesized item-scale correlations are .40 or greater. Note that Items 3a and 3j often correlate less than .40 because they define the ceiling and floor of the PF scale, resulting in correlations that are weaker due to the skewed distribution. Any items that fail tests of internal consistency (i.e., items that correlate less than .40 with their hypothesized scales) should be evaluated to determine where potential problems might have occurred in the survey data. For example, if MH Items 9d and 9h failed to correlate .40 or greater with the MH scale score, these items might have been reversed scored before the data set was submitted for scoring. Alternatively, the items could have been incorrectly labeled. Whatever the cause, the data file should be checked and any problems corrected before resubmitting the data for scoring.

Item Discriminant Validity

The sixth data quality indicator is *item discriminant validity*. Tests of item discriminant validity are conducted to evaluate the validity of the hypothesized item groupings. When assessing data quality, it is not sufficient to demonstrate just that an item appears to measure the construct it was intended to measure (as evidenced by tests of item internal consistency; see previous section). It is also important to determine whether an item measures any other constructs that it was not intended to measure. For example, because Item 3a is in the PF scale and not the MH scale, then Item 3a should be more strongly related to the overall PF scale score than to the overall MH scale score.

To evaluate item discriminant validity:

1. Examine the correlation between each item and its hypothesized health domain scale score, corrected for overlap, for all completed SF-36v2 surveys.
2. Examine the correlations between each item and the remaining seven health domain scale scores (i.e., those scales the item does not belong to).
3. Determine if the correlation between an item and its hypothesized scale is greater than the correlations between said item and each alternative scale.
4. Compute the percentage of items that correlate higher with their hypothesized scales than with alternative scales.

This data quality indicator is considered satisfactory when at least 80% of the hypothesized item-scale correlations are higher than the alternative item-scale correlations. Even when the overall result is satisfactory (i.e., the 80% criterion is met), each item that correlated more strongly with an alternative scale than with its hypothesized scale should be examined to determine the sources of potential problems in the survey data. For example, items that failed tests of item discriminant validity could have been incorrectly labeled or might have been reversed scored prior to final scoring. Whatever the cause, the data file should be checked and any problems corrected before resubmitting the data for scoring.

Scale Reliability

The seventh data quality indicator is *scale reliability*. Measurement reliability refers to the extent to which the measured variance in a given scale score reflects the true score, rather than random error. A common approach used to evaluate scale score reliability uses an estimate of internal consistency reliability based on the number of items in a scale and item homogeneity (similarity) called Cronbach's alpha coefficient. When Cronbach's alpha coefficient is greater than or equal to .70, then scale reliability is generally considered to meet minimum standards of acceptability (Nunnally & Bernstein, 1994).

To evaluate scale reliability:

1. Determine the Cronbach's alpha coefficient for each health domain scale for all completed SF-36v2 surveys.
2. Compute the percentage of scales that have coefficients of .70 or greater.

This data quality indicator is considered satisfactory only when 100% of the scales have Cronbach's alpha coefficients of .70 or greater. When this criterion is not met, each item in each scale with a coefficient of less than .70 should be examined to determine the sources of potential problems in the survey data. Whatever the cause, the data file should be checked and any problems corrected before resubmitting the data for scoring. For a detailed discussion of scale reliability issues, please see Chapter 15 of this manual.

Confirmation of the Two-Component Structure

The eighth data quality indicator is *confirmation of the two-component structure*. Applying this quality check allows users to establish and appraise the relationship of each SF-36v2 scale with the PCS and MCS measures. To evaluate this data quality indicator:

1. Examine the pattern of correlations between the health domain scales and component summary measures for all completed SF-36v2 surveys.
2. Determine if each scale and measure has its intended interpretation as a measure of physical or mental health status.
3. Confirm whether the survey's two-component structure is generally consistent with what has been found in the U.S. general population (see Tables 16.1 and 16.2).

Unlike the other quantitative indicators discussed in this chapter, no specific criterion or cutoff score to confirm the two-component structure is offered here. Instead, this data quality indicator is considered satisfactory when the informed judgment of the user has been satisfied. When the two-component structure cannot be confirmed (i.e., the factor structure of a group's scales and measures is not consistently replicated), caution is warranted when interpreting the scores of said group.

Qualitative Evaluation of Data Quality

When analyzing individual respondent data, there are three additional data quality checks that can be performed to determine the quality of the SF-36v2 data. These qualitative checks are: (a) responses inconsistent with respondent presentation, (b) unusually quick or long completion time, and (c) patterned responses. Although more subjective than the quantitative indicators previously discussed, these qualitative indicators can provide additional insight into respondents' scores, which may prompt administrators to more closely scrutinize survey results and may help to determine the validity of respondents' item answers and overall survey scores.

Results Inconsistent With Respondent Presentation

At times, administrators may notice a discrepancy between how respondents answer items and how they present themselves during the testing session. For example, the validity of results should be questioned when a respondent has indicated on the SF-36v2 form that he is limited *a lot* in walking more than a mile and feels worn out *all of the time*, yet during an informal conversation with the administrator has indicated that he runs 3 miles every day to stay in shape.

Unusually Quick or Long Completion Time

The SF-36v2 is a relatively brief measure of health status that can be completed by most respondents within

5 to 10 minutes. Completion of the survey in significantly less time (e.g., less than 2 minutes) suggests that the respondent might have answered the items randomly or without much consideration of the items' content or the accuracy of his or her responses. Completion of the survey in a significantly greater amount of time than usual (e.g., 20 minutes) may indicate poor motivation, the presence of reading problems, or difficulty understanding item content. In such cases, the administrator should ask the respondent about his or her motivation to complete the survey honestly, his or her understanding of the survey items, or other questions appropriate to the situation. Depending on how the respondent reacts, the administrator may want to ask the him or her to complete the survey at a different time and/or using a different mode of administration (e.g., interview format).

Patterned Responses

This qualitative check is conducted by visually inspecting a completed paper form or a printed listing of item response numbers generated from an automated (e.g., online) administration. Generally speaking, one should be suspicious of results that demonstrate any of the following characteristics:

- The same response choice (e.g., the first, the last, the middle) is selected for all items.
- The response choice indicating the worst level or the best level of functioning is always selected.
- The response pattern is sequential from one item to the next within a given scale (e.g., the 10 PF items are answered 1, 2, 3, 1, 2, 3, 1, 2, 3, 1) or across the entire survey (e.g., 1, 2, 1, 2, 1, 2, 1, 2, 1, 2 . . .).

While possible, it's highly unlikely that these and other types of patterned answers truly reflect honest and valid responses to survey items. When such patterns appear, determine the accuracy of survey results by asking the respondent to explain his or her item responses.

Data Quality Evaluation of Individual Health Domain Scales

Thus far, the recommendations made in this chapter have been discussed in terms of SF-36v2 results *as a whole*. However, most of these same data quality checks can be applied to the data on a *scale-by-scale* basis. When evaluating the data quality of each individual scale, users should follow the same guidelines and apply the same criteria that are used for determining the quality

of the data as a whole (i.e., all the scales and measures considered together). For example, the 90% criterion should still be used when evaluating the PF scale's completeness of data. Similarly, the item discriminant validity of the MH scale can be evaluated by calculating

the percentage of MH items that have greater correlations with the MH scale itself (corrected for overlap) than with the other seven scales. As with group-level data, this quality check would be considered satisfactory when the result is 80%.

7

General Strategies for Interpreting the SF-36v2 Profile

Once users are confident that their SF-36v2 results satisfy data quality standards, interpretation of those results can begin. As discussed in this and the next five chapters, SF-36v2 results can serve as a rich source of information for understanding the health status of individual respondents or groups of respondents when different approaches to examining the data are taken. The purpose of this chapter is to provide users with a basic, general approach to and rules for guiding the interpretation of results from SF-36v2 administrations.

The general interpretive approach described in this chapter employs a systematic examination of the SF-36v2 profile, first from a broad perspective and then conducting a more detailed analysis of the data. This approach involves determining if the *T* scores for the PCS and MCS measures deviate from what is considered the average range for the U.S. general population. This is followed by examining the health domain scale scores to make a similar determination. Each of these decisions is based on separate, empirically based individual respondent- and group-level guidelines. The guidelines for interpreting high and low scores on the PCS and MCS measures and on each health domain scale are presented in tabular format. Overall, this examination serves as the context in which the content-based and criterion-based approaches to the interpretation of results (described in Chapters 8 and 9, respectively) should take place.

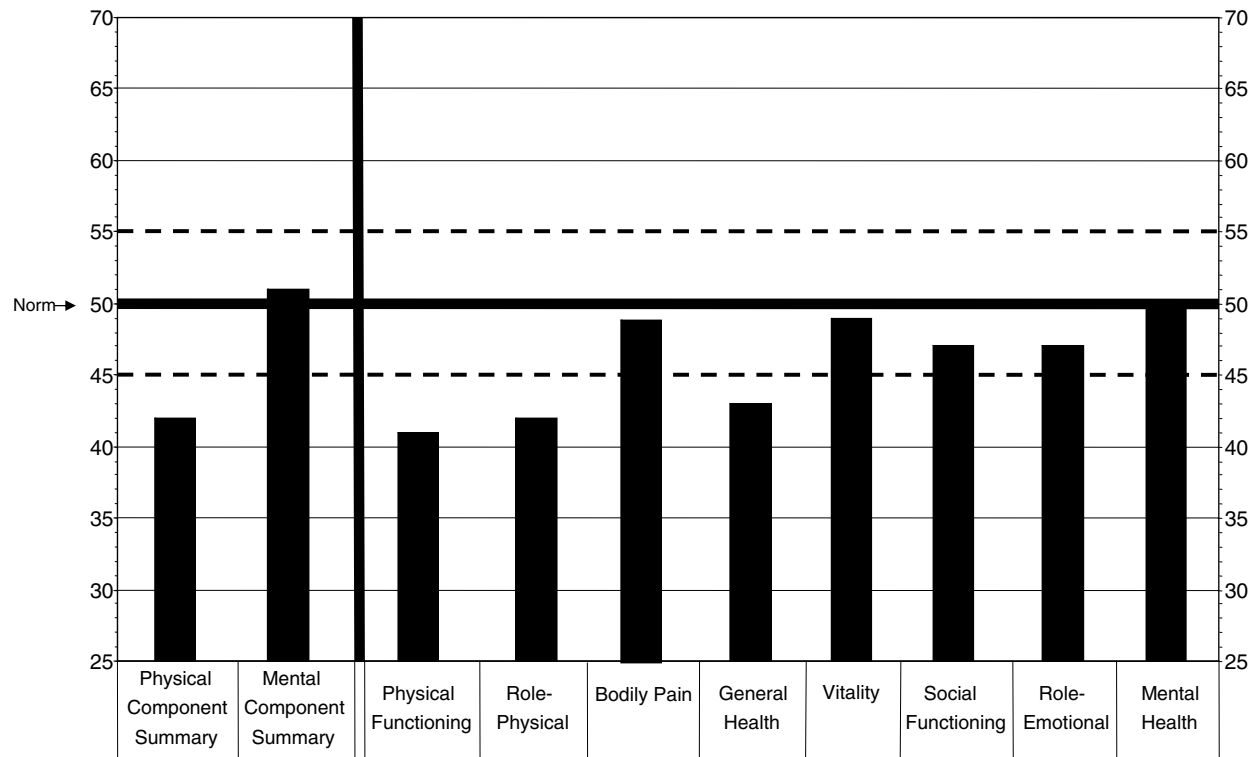
Also discussed in this chapter are considerations regarding the interpretation of SF-36v2 results in light of results from other psychometric perspectives. The application of measure- or scale-specific standard errors of measurement (*SEMs*) allows users to determine, within specific levels of confidence, intervals in which a respondent's true score falls on each measure and scale. The availability of gender- and age-based norms, as well as benchmark data for 40 disease groups, provide additional channels for better understanding the meaning of the observed scores.

General Considerations for Norm-Based Interpretation

Interpretation of the results should begin with a visual examination of the SF-36v2 profile of scores. The profile, which may represent the observed scores for an individual survey respondent or the mean scores for a group of respondents, provides a broad overview of health status. The scores presented first in the profile are the PCS and MCS scores. Placement of these measures at the beginning (left side) of the profile emphasizes the importance of first considering individual or group results with regard to overall functioning in the physical and mental health dimensions (see Figure 7.1). Thus, one can quickly determine upon visual examination of the profile whether deviations from the norm are more or less apparent in the general physical or mental health status for an individual or a group.

To obtain a clearer picture of a respondent's health status, a close examination of the norm-based *T* scores for the health domain scales is recommended. Note that the ordering of the health domain scales facilitates interpretation of the profile, with differences on the left side of the health domain profile (PF, RP, BP, and GH) generally reflecting physical health status and differences on the right side (VT, SF, RE, and MH) generally reflecting mental health status (see Figure 7.1; see also Chapters 13 and 16 for empirical evidence supporting the ordering of the health domain scales).

In reviewing Figure 7.1, users can quickly determine that the health burden in this example is primarily observed among measures of physical health status. For example, the PCS score is well below the general population norm score of 50, whereas the MCS score is slightly above the norm score. Likewise, three of the four health domain scales at the left of the profile show deficits in comparison to the norm, whereas scores for three of the four health domain scales at the right of the profile are

Figure 7.1 Sample SF-36v2 Profile of Scores

Note. The dashed lines (---) indicate the upper (55) and lower (45) bounds of *T* scores considered to be in the average range of functioning for individual respondents.

at or near the norm. As Figure 7.1 demonstrates, in most instances the four health domain scales at the left of the profile will usually correspond to what is observed with the PCS measure, and the four health domain scales at the right of the profile will usually correspond to what is observed with the MCS measure.

In the guidelines that follow, note that the recommendations for interpreting differences in individual respondent scores differ slightly from the recommendations given for interpreting group-level mean scores. These differences reflect the fact that group-level mean scores contain less measurement error than individual respondent-level scores. One can therefore have greater confidence in the interpretation of group mean scores than in the interpretation of individual respondent scores. Consequently, smaller differences in group mean scores can be meaningfully interpreted. Because individual respondent scores likely contain more measurement error, there is less confidence that the obtained score represents the respondent's true score. Thus, interpreting individual respondent scores requires less stringent or conservative guidelines that take into consideration the likelihood of measurement error.

As illustrated in more detail in Chapter 13, the interpretation of SF-36v2 results has been made easier

with the use of the *T*-score metric, based on 2009 U.S. general population normative data, for scoring the health domain scales and component summary measures. Specifically, *T* scores have proven to be very useful when interpreting differences across the eight health domain scales and for purposes of comparing those domains with the two component summary measures. With *T* scores, each scale is scored using the same mean (50) and the same standard deviation (10 points) found in the 2009 U.S. general population. Thus, each *T*-score point is one-tenth of a standard deviation (*SD*). With this method, one can determine the status of the health dimension (physical or mental) or domain represented by the measure or scale, relative to the average, without referring to tables of norms.

As a general rule, when considering individual respondent data, it is recommended that scores within 0.5 *SD*, or 5 *T*-score points, of the mean be considered within the "average" or "normal" range for the U.S. general population. Thus, an individual respondent's score on any health domain scale or component summary measure that falls outside the *T*-score range of 45 to 55 (i.e., more than 0.5 *SD* below or above the mean norm-based score of 50) should be considered outside the average range for the U.S. general population.

The further a score is from the mean, the greater the likelihood that the respondent is above-average or below-average in a given area of functioning or well-being. Generally, when considering *individual respondent* results, one can be confident that health domain scale or component summary measure scores falling more than 1 *SD* (10 *T*-score points) below the population mean (i.e., mean minus 10 *T*-score points) are indicative of significantly impaired functioning. Thus, scores less than 40 indicate impairment in that health domain or dimension. Scores in the 40-to-44 range fall within a “gray” area of interpretation and require further investigation to determine the presence of impaired functioning (further discussed later in this chapter). Finally, *T* scores of 45 or greater should be considered average or above average for individual respondents, as compared to the 2009 normative sample.

As a general rule, when considering group-level data, it is recommended that scores within 0.3 SD, or 3 T-score points, of the mean be considered within the “average” or “normal” range for the U.S. general population. Any health domain scale or component summary measure score falling outside the *T*-score range of 47 to 53 (i.e., more than 0.3 *SD* below or above the mean norm-based score of 50) should be considered outside the average range for the U.S. general population for group data. Thus, when considering *group-level* results, a score on a health domain scale or component summary measure that is less than 47 should be considered indicative of impaired functioning within that health domain or dimension. This more stringent cutoff for group-level results reflects the greater confidence that one can have in the obtained group mean scores, as previously discussed. Similar to individual respondent data, group mean scores 47 or greater should be considered average or above average as compared to the 2009 normative sample.

In analyzing SF-36v2 group-level results, it is also important to consider the percentage of the sample that scored above the average range for the *individual respondent classification* (i.e., 56 or higher) and the percentage that scored below the average range for the *individual respondent classification* (i.e., 44 or lower) on each component summary measure and health domain scale. These data provide information beyond what can be conveyed by group-level summary scores on these variables and can serve as a means of determining what percentage of the sample scored within or outside of the average range of scores observed in the general population. Such data can also assist in evaluating the effectiveness of an intervention in clinical trials or treatment programs or in comparing the outcomes of two or more types of intervention.

Thus far, a very basic strategy for interpreting SF-36v2 results has been presented. The discussion will now begin to move toward more specific interpretation strategies. These strategies will only be introduced here, with a much more detailed discussion provided in Chapters 8, 9, and 10 of this manual.

Interpretation of the Component Summary Measures

The two component summary measures—PCS and MCS—provide reliable and valid summaries of a respondent’s or group’s physical and mental health status. As previously noted, *T* scores in the 45-to-55 range should be considered average for individual respondents. That is, *T* scores of 45 or greater indicate at least average overall functioning in the general health dimension—physical or mental—assessed by its associated measure. *Individual respondent T* scores that are less than 40 and *group mean scores* that are less than 47 indicate the presence of impaired functioning in the associated dimension. Meanwhile, individual respondent scores in the 40-to-44 range require further investigation, including consideration of the confidence interval around the score and the choice of age-, gender-, and/or disease-based norms, to determine whether the score is more indicative of impaired or unimpaired functioning in the respective health dimension.

In Table 7.1, each SF-36v2 component summary measure and health domain scale is described in terms of: item composition, number of score levels, lowest and highest possible *T* scores for the standard and acute forms, and the health states associated with the lowest and highest observable scores. These descriptions are based on the general content of the health domain scales and component summary measures and/or the pattern of responses necessary to achieve these extreme scores. This information can be used to summarize what each component summary measure and health domain scale assesses and can serve as a basis for broad-level interpretation of SF-36v2 results. Approaches to understanding the meaning of health domain scale and component summary measure scores falling between the extreme scores are provided in Chapters 8 and 9.

Because the PCS and MCS measures are composites that reflect a combination of physical and mental *functioning* and *well-being*, the extent of social and role *disability*, and personal evaluation of *health status*, the meanings of scores on these measures are not as straightforward as they are for the more homogeneous health domain scales. In other words, there are more ways to obtain each possible score for each component summary

Table 7.1

Composition and Interpretation of Lowest and Highest T Scores for SF-36v2 Component Summary Measures and Health Domain Scales

Scale/Measure	Composition	Number of Score Levels	Standard Form		Range of Possible T Scores		Acute Form	Lowest Possible Score	Highest Possible Score	Description
			Lowest*	Highest*	Lowest*	Highest*				
Physical Component Summary	All scales	486	7.32	70.14	10.80	75.51				No physical limitations, disabilities, or decrements in well-being; high energy level; health rated <i>excellent</i>
Mental Component Summary	All scales	494	5.79	69.91	5.62	69.65				Frequent positive affect; absence of psychological distress or limitations in usual social/role activities due to emotional problems; health rated <i>excellent</i>
Physical Functioning	Items 3a–3j	21	19.26	57.54	19.03	57.60				Performs moderate types of physical activities without limitations due to health; is not limited in climbing several flights of stairs
Role-Physical	Items 4a–4d	17	21.23	57.16	21.89	57.12				<i>Never</i> experiences problems in accomplishing as much work or other daily activities as one would like, including not being limited in the kind of work or activities performed, as a result of physical health
Bodily Pain	Items 7, 8	11	21.68	62.00	21.39	60.87				No interference with normal work due to pain
General Health	Items 1, 11a–11d	21	18.95	66.50	21.29	65.40				Evaluates one's health as <i>poor</i>
Vitality	Items 9a, 9e, 9g, 9i	17	22.89	70.42	25.60	69.15				Has a lot of energy <i>all of the time</i>
Social Functioning	Item 6, 10	9	17.23	57.34	17.20	56.74				No interference with social activities due to physical and emotional problems <i>all of the time</i>
Role-Emotional	Items 5a–5c	13	14.39	56.17	9.84	55.64				<i>Never</i> experiences problems in accomplishing as much work or other daily activities as one would like, including not being limited in the kind of work or activities performed, as a result of emotional problems
Mental Health	Items 9b, 9c, 9d, 9f, 9h	21	11.63	63.95	13.12	62.67				Feels downhearted and depressed <i>all of the time</i> ; <i>never</i> feels calm and peaceful
Self-Evaluated Transition (SET)	Item 2	5	N/A	N/A	N/A	N/A				Health much worse than 1 year/week ago

*Highest and lowest observed T scores in 2009 U.S. general population normative sample.

measure, in comparison to the number of ways to obtain each possible score for each health domain scale. However, very high PCS and MCS scores (i.e., at or around the highest possible *T* score) indicate the best possible physical and mental performance and capacity, respectively, while very low scores (i.e., at or around the lowest possible *T* score) indicate the worst possible physical and mental performance and capacity, respectively. Specifically, very high scores on the PCS measure are indicative of no measured physical limitations, disabilities, or decrements in well-being; a high level of energy; and a self-rating of health as *excellent*. Conversely, very low PCS scores are indicative of substantial limitations in self-care, physical, social, and role activities; severe bodily pain; frequent tiredness; and health rated as *poor*. For the MCS measure, very high *T* scores indicate frequent positive affect, absence of psychological distress and limitations in usual social or role activities due to emotional problems, and health rated as *excellent*. In contrast, very low MCS scores indicate frequent psychological distress, substantial social and role disability due to emotional problems, and health rated as *poor*.

Note that it is important to recognize two key aspects of the PCS and MCS measures. First, although the operational definitions are similar for some of the physical and mental health items, they are conceptually different. The PCS measure reflects *physical morbidity and etiology*, whereas the MCS measure reflects *psychological or mental morbidity and etiology*. Second, a very high PCS score requires more than just freedom from physical limitations and social and role disability; it requires an evaluation of current health as *excellent*. Likewise, the most favorable personal evaluation of health as *excellent* is not enough for a very high score because PCS scores decrease with limitations or disabilities in the physical spectrum, reflecting the consequences of such limitations and disabilities in physical health. This same logic is reflected in the scoring of the MCS measure. Both PCS and MCS place considerable weights on both the personal and the social implications of different health states. For these reasons, PCS and MCS are unique in their comprehensiveness as summary measures of health. A more detailed discussion of content-based interpretation of PCS and MCS scores is presented in Chapter 8.

Interpretation of the Health Domain Scales

Scanning the remainder of the SF-36v2 profile (see Figure 7.1) allows the user to obtain a general picture of the respondent with regard to impairment in specific

physical and mental health domains, which in turn helps to better understand what is driving the obtained PCS and MCS scores. Using the previously presented general guidelines for interpretation, respondent profiles with *T* scores that are 45 or greater, or group profiles with mean *T* scores that are 47 or greater, on all eight health domain scales are indicative of a respondent/group whose health status is either near or above average on all assessed health domains, as compared to the 2009 U.S. general population.

In comparison with the PCS and MCS measures, the interpretation of very high and very low health domain scale scores is more straightforward. For example, the lowest possible score on the PF scale means that a respondent reports being limited a lot in performing all assessed physical activities, including vigorous activities (running, lifting heavy objects, or strenuous sports), moderate activities (pushing a vacuum cleaner, bowling, or golf), walking, climbing stairs, lifting and carrying groceries, and bathing or dressing. The highest possible score on the PF scale means that a respondent reports being capable of performing all these physical activities without any limitation. (Chapters 8 and 9 provide guidelines for interpreting health domain scale scores that are between the extreme scores.)

Careful examination of the health domain scales is particularly useful in cases where the score of either or both of the PCS and MCS measures are in the impaired range. Recall that the first four health domain scales on the sample profile—PF, RP, BP, and GH (found on the left side of the health domain scale section of the profile)—have the greatest physical factor content among the health domains. The last four scales—VT, SF, RE, and MH (found on the right side of the health domain scale section of the profile)—have the greatest mental factor content. Furthermore, to facilitate interpretation, the domain scales are ordered from left to right according to their physical and mental health factor content: from the best physical health measure (PF) on the left to the best mental health measure (MH) on the right. Thus, a low score on the PCS measure will most often be reflected in a low score on one or more of the scales located on the left side of the health domain portion of the SF-36v2 profile. Similarly, a low score on the MCS measure will most often be reflected in a low score on one or more of the scales located on the right side of the health domain portion of the profile. Low scores across the profile are indicative of impairment in both of the physical and mental health components.

Once identified through an evaluation of the health domain scale profile, individual scores falling into the impaired range can provide general information about

the respondent's functional and/or emotional limitations in the associated health domains (see Table 7.1). Limitations found at other score levels can be ascertained through an analysis of responses to individual items on the scale(s) in question or through the data presented in Chapters 8 and 9.

At times, users will find that the health domain scale scores will not be what are expected given the observed PCS or MCS scores. For example, one may find that a PCS *T* score for an individual respondent that falls below the average range (44 or lower) is accompanied by physical health domain (PF, RP, BP, or GH) *T* scores that are in the average or above-average range (45 or greater). Similarly, mental health domain scale scores may not be consistent with what is expected given a low MCS score. Such apparent discrepancies can be attributed to the way in which the PCS and MCS *T* scores are calculated, with all eight health domain scores contributing to both the PCS and MCS scores. When such findings are observed for individual respondents, the user should examine the responses to each of the SF-36v2 items. This can help provide a clearer picture of the degree of impaired functioning than is possible from the component summary measure *T* scores alone.

Additional Considerations for Interpreting SF-36v2 Findings

Thus far, this chapter has provided a very broad, general approach to interpreting the SF-36v2 profile of scores. Use of other available data can provide further information that may help the user better interpret the SF-36v2 profile and arrive at a more refined picture of an individual respondent's health status.

The Standard Error of Measurement (*SEM*) and Confidence Intervals (*CI*s)

The recommended procedure for interpreting SF-36v2 results is to evaluate the PCS and MCS *T* scores as well as the health domain scale *T* scores. In doing so, it is important to be mindful of the random measurement error that is contained in each measure and scale score and, as necessary, use it to temper the interpretation of the obtained results. To do this, one needs to consider the *standard error of measurement (SEM)* for each measure or scale.

Confidence intervals (*CI*s) provide valuable information about the amount of fluctuation that can be expected in a single score due to measurement error. A *CI* around an individual score is a function of the *SEM*, which is inversely related to the sample size and reliabil-

ity of the score (Nunnally & Bernstein, 1994). A scale or measure with a relatively small *SD* and high reliability has a small *SEM* and small *CI*s around individual scores. With smaller *CI*s, fluctuations in an individual respondent's scores due to chance are less likely, facilitating their use in monitoring individual patients in clinical practice. The *SEM* is further discussed in Chapter 15 of this manual.

Tables 7.2 and 7.3 provide estimates for constructing *CI*s for the eight SF-36v2 health domain scales and two component summary measures for both the standard and acute forms, respectively. Values for constructing score intervals for four levels of confidence are presented: 68% (± 1 *SEM*), 80% (± 1.28 *SEMs*), 90% (± 1.64 *SEMs*), and 95% (± 1.96 *SEMs*). The *CI* for each score is constructed by first adding and then subtracting the appropriate *T*-score value for the desired confidence level (see Tables 7.2 and 7.3) to and from the observed *T* score to establish the *CI* upper and lower limits, respectively. According to Table 7.2, individual respondent scores on the standard form PCS and MCS measures would be expected to fall within ± 2.0 and ± 2.7 *T*-score points of the observed score, respectively, about 68% of the time. For greater confidence in a respondent's PCS and MCS scores, one may choose to use the 90% or 95% *CI* for the component summary measures (i.e., ± 3.3 or ± 3.9 points, respectively, for the PCS, and ± 4.4 or ± 5.3 points, respectively, for the MCS). Note that the upper boundary of a *CI* will be limited for observed scores that approach the maximum score (ceiling) for a given measure or scale because a *CI* cannot extend beyond the maximum possible score for any scale or measure. Similarly, the lower boundary of a *CI* will be limited for observed scores that approach the minimum score (floor) for a given measure or scale.

The *SEM* depends on the precision of the measurement instrument, which in classical psychometrics is generally assumed to be invariant across populations (in contrast to the reliability coefficient, which may differ depending on the sample distribution of health outcomes). Mosteller, Ware, and Levine (1989) argued for maintaining standardization across populations, noting that moving away from standardization results in less information about variations in health status and what the variations mean, as well as lost opportunities for important comparisons between different health and demographic populations. SF-36v2 users are therefore encouraged to use the scoring services offered by QualityMetric Incorporated or its authorized resellers, along with the interpretive data published in this manual. However, if the *SEM* can be proven to be substantially different in particular samples, some (e.g., Nunnally

Table 7.2

Values for Constructing Confidence Intervals Around Individual Respondent SF-36v2 Standard (4-Week Recall) Form T Scores Based on the 2009 U.S. General Population Data (N = 4,024–4,036)

Scale/Measure	68% ^a	80% ^b	90% ^c	95% ^d
Physical Component Summary	2.0	2.5	3.3	3.9
Mental Component Summary	2.7	3.4	4.4	5.3
Physical Functioning	2.5	3.1	4.0	4.8
Role-Physical	2.0	2.6	3.3	3.9
Bodily Pain	3.6	4.6	5.9	7.1
General Health	4.2	5.4	7.0	8.3
Vitality	3.6	4.6	5.9	7.1
Social Functioning	4.0	5.1	6.6	7.8
Role-Emotional	2.6	3.4	4.3	5.2
Mental Health	3.6	4.6	5.9	7.1

Note. Estimates are based on reliability estimates and standard deviations for the eight health domain scales and the PCS and MCS measures in the 2009 U.S. general population.

^a68% CI = observed *T* score ± 1 SEM.

^b80% CI = observed *T* score ± 1.28 SEMs.

^c90% CI = observed *T* score ± 1.64 SEMs.

^d95% CI = observed *T* score ± 1.96 SEMs.

Table 7.3

Values for Constructing Confidence Intervals Around Individual Respondent SF-36v2 Acute (1-Week Recall) Form T Scores Based on the 2009 U.S. General Population Data (N = 2,056–2,061)

Scale/Measure	68% ^a	80% ^b	90% ^c	95% ^d
Physical Component Summary	1.8	2.3	2.9	3.5
Mental Component Summary	2.8	3.5	4.5	5.4
Physical Functioning	2.2	2.9	3.7	4.4
Role-Physical	2.0	2.6	3.3	3.9
Bodily Pain	3.5	4.4	5.7	6.8
General Health	3.9	5.0	6.4	7.6
Vitality	3.6	4.6	5.9	7.1
Social Functioning	4.4	5.6	7.1	8.5
Role-Emotional	2.4	3.1	4.0	4.8
Mental Health	3.5	4.4	5.7	6.8

Note. Estimates are based on reliability estimates and standard deviations for the eight health domain scales and the PCS and MCS measures in the 2009 U.S. general population.

^a68% CI = observed *T* score ± 1 SEM.

^b80% CI = observed *T* score ± 1.28 SEMs.

^c90% CI = observed *T* score ± 1.64 SEMs.

^d95% CI = observed *T* score ± 1.96 SEMs.

& Bernstein, 1994) recommend that the CIs should be re-estimated using published formulas (see Thissen & Wainer [2001] for a discussion on measurement error within classical psychometrics; see also Thissen & Orlando [2001] for a discussion on this topic from the perspective of item response theory). For example, inclusion and exclusion criteria in clinical trials produce homogeneous samples whose SF-36v2 scores vary less than general population scores. Consequently, SEM es-

timates will tend to be much smaller for these samples than the estimates observed in the general population.

Individual *T* scores on the eight health domain scales and two component summary measures can be compared to U.S. general population norms or to norms for specific demographic or diagnostic groups (see following sections) by using the CI values presented in Tables 7.2 and 7.3. For example, suppose that a clinician wants to know whether a PCS score of 43, obtained from the administration of the SF-36v2 standard form to a 70-year-old male, is below average compared to the U.S. general population. By applying ± 1.96 SEMs (3.9 *T*-score points) to the observed score, the clinician can be 95% confident that the respondent's true PCS score falls within the *T*-score range of 39.1 to 46.9.

SEMs can also be used in analyzing group-level data. Most group-level outcomes are presented in terms of average change scores, which can mask underlying variability in those outcomes. SEMs can be used to classify individual change scores as *better*, *same*, or *worse*, and the proportion of better:same:worse can then be compared between groups.

Supplemental Norms for Age, Gender, and Gender-by-Age Groups

As part of the 2009 norming study, separate sets of SF-36v2 age, gender, and gender-by-age norms for the health domain scales, component summary measures, and SF-6D were developed for both the standard and acute forms. The age groupings were selected (a) to be large enough to satisfy minimum standards for precision, (b) to correspond with standard practices for defining age-specific groups, and (c) to correspond with age groupings used by others when reporting norms for the SF-36 (Brazier et al., 1992; Jenkinson, Coulter, & Wright, 1993; Ware et al., 2007; Ware, Snow, Kosinski, & Gandek, 1993; Ware, Kosinski, & Keller, 1994). These supplemental norms are useful for determining whether a score for a male or a female is above or below the average score for males or females in a particular age group in the U.S. general population. To illustrate, using the previous example and the 2009 SF-36v2 standard form gender-by-age norms for males aged 65 through 74, it is apparent that a PCS score of 43 is only 0.36 SDs below the *T*-score norm of 46.63 for a 70 year-old male ($46.63 - 43 = 3.63$ *T*-score points = 0.36 SD). Thus, the clinician should feel confident that the respondent's score of 43 is within the norm for males of a similar age in the 2009 U.S. general population.

With regard to the 2009 SF-36v2 age-based norms for both the standard and acute forms, three points are worth noting. First, as with the general population

norms (see Tables 14.8 and 14.9) and with a few noted exceptions, the medians (50th percentile scores) for each health domain scale and component summary measure are higher than their mean scores. This reflects the skewness of the score distributions in the 2009 U.S. general population sample. Consequently, in the general population, one can expect a greater proportion of respondents to score above the mean. Second, comparing results across age groups clearly shows that health status, in particular physical health, is related to age. Generally, the mean scores for all physical health scales and component summary measures decline with age. For example, whereas the mean PF *T* score for the total normative sample is 50.00, the mean for the 18-to-24-year-old group is higher (54.20) and the mean for the 75-and-older group is lower (41.00). Third, beginning with the 45-to-54-year-old age group, mean PCS scores begin to decline noticeably while the opposite is true of the MCS scores.

Supplemental Benchmarks for Disease-Specific Populations

The usefulness of the Short Form family of instruments in describing the burden of disease is documented in publications describing more than 150 diseases and conditions, with at least 16 conditions *each* being addressed in more than 100 publications representing more than 850 controlled clinical trials studying the impact of treatment (see Chapter 2). As part of the SF-36v2 2009 U.S. general population normative data gathering effort, participants were asked to indicate whether they were suffering from one or more of 40 diseases or physically or mentally impairing conditions. This information enabled the development of specific sets of benchmarks for each of the listed conditions and disease states, which are listed in Tables 14.12 and 14.13.

When it is known that the respondent belongs to a specific disease or chronic condition population for which SF-36v2 benchmark data are available, or if the data being analyzed is for a group of patients with one or more of those conditions, it is useful to compare their scores to the disease- and condition-specific data. For example, comparison of an individual respondent's or a group's SF-36v2 profile to the profile of the mean health domain scale and component summary measure *T* scores for the relevant disease group can provide an indication of how similar they are to the prototypical or "average" member of that disease group, in terms of health status and the extent of the associated limitations. This allows users to generate hypotheses regarding the severity of impairment, which may in turn have implications for the description, treatment, and prognosis of the condition.

Additional information regarding the utility of the disease-specific benchmarks can be found in the percentage of respondents in each disease group who scored the highest possible score (i.e., the ceiling) and the percentage of respondents who scored the lowest possible score (i.e., the floor) for each scale. The percentage of each of the 40 disease groups scoring at the floor and at the ceiling of the standard and acute form health domain scales can be found in Tables 14.12 and 14.13, respectively.

SF-36v2 reports incorporating 2009 age, gender, and disease-specific normative and benchmark information are available through scoring services offered by QualityMetric and its authorized resellers.

Use of Information From Other Instruments

As with other psychometric instruments, SF-36v2 results can be more clearly understood when interpreted within the context of other information known about a respondent or a group of respondents. Common sources of additional information include lab tests, face-to-face interviews, chart reviews, and other self-report instruments such as disease-specific measures of HRQOL. Data from all of these sources may provide insight into the nature and extent of any health status problems revealed by SF-36v2 component summary measure and health domain scale scores. When used primarily for research, data from other sources can also serve several purposes, such as cross-validating SF-36v2 findings, identifying important covariates, and determining the generalizability of the results. For individual respondents in clinical settings, use of multiple assessment instruments can assist in arriving at more accurate and comprehensive diagnoses and lead to the development of more effective treatment plans. In most clinical settings, data from many of these sources can be obtained at the time the respondent completes the SF-36v2.

When evaluating an individual respondent or a group of respondents with the SF-36v2 and another psychometric instrument that purportedly measures one or more of the same constructs, users may obtain results that appear contradictory. This can occur for any of several reasons, including (a) one instrument was developed to broadly sample a given domain, whereas the other was developed to provide a comprehensive or focused assessment of a specific aspect of that same domain; (b) the normative sample for one instrument differs from that of the other instrument; (c) one instrument is more appropriate than the other for assessing the person or population in question; (d) one instrument is more valid than the other for the purpose for which it is being used; or (e) errors occurred in scoring one or more of the administered instruments. Assuming that both instruments are valid and

were administered at the same time, findings that appear contradictory should lead the administrator to determine if the differences can be attributed to the content, norms, or intended purposes of the two instruments. Resolution

of any contradictory findings obtained from the SF-36v2 and another instrument (or from any two instruments) often leads to a better understanding of the respondent or group of respondents being assessed.



8

Content-Based Interpretation

General norm-based strategies for interpreting the scores of each health domain scale and component summary measure based on the 2009 SF-36v2 normative data are provided in Chapter 7 of this manual. *Content-based interpretation*, an approach based on analyses of the content of and responses to *individual items*, is another strategy that can be used to interpret differences in health domain scale and component summary measure *T* scores across the range of possible scores. This is accomplished by plotting specific responses to SF-36v2 items across score levels of the health domain scales and component summary measures. For example, it would be useful to know that more than 87% of the 2009 U.S. general population sample that earned a *T* score of lower than 30 on the PF scale were limited in walking 100 yards.

The purpose of this chapter is to present empirical, item-level SF-36v2 data from the 2009 norms study, and an approach to analyzing those data that can be used to understand the meanings of scores that fall between the extreme scores (highest and lowest scores) of each health domain scale and component summary measure.

Interpretation of Scales and Measures Across All Score Ranges

Content-based interpretation guidelines for each component summary measure and health domain scale, across all score ranges, were developed in several steps. First, responses to each of the SF-36v2 items collected during the 2009 norms study were dichotomized in a meaningful way that was thought to be capable of revealing differences across levels of the scale or measure in the score ranges of interest. Generally, the two or three responses selected to serve as the basis for analysis for

each item were those thought to be indicative of notable problems in the construct or behavior being assessed by said item. These same responses to each item scored for all the health domain scales most highly correlated with each component summary measure—PF, RP, BP, and GH for the PCS measure, and VT, SF, RE, and MH for the MCS measure—served as the bases for the content-based interpretation of the PCS and MCS measures.

Second, the percentage of the 2009 normative sample who responded to the selected item responses at each respective health domain scale and component summary measure *T*-score level being interpreted was determined. Generally, the score levels, which can range from 7 to 9 depending on the scale or measure and the form (standard or acute), represent 5-point *T*-score intervals throughout the range of scores observed in the 2009 U.S. general population for each summary measure and scale. Often, however, the highest and lowest score levels are combined to encompass a larger, more meaningful range of scores.

Third, the percentage of the 2009 normative sample who endorsed the previously selected item responses at each of the *T*-score levels for the parent scale was evaluated. All items were found to provide useful interpretations across the entire continuum or at particular levels of component summary measure and health domain scale *T* scores and were retained as recommended sources of content interpretation for the SF-36v2. Note that the 2009 SF-36v2 standard (4-week recall) form percentages are presented in Tables 8.1 through 8.18, and the 2009 acute (1-week recall) form percentages are presented in Tables 8.19 through 8.36.

To facilitate the interpretation of the results presented here, the same format is used for all tables. As such, the range of *T* scores, the mean *T* score, and the sample size for each level are presented in the left-most columns.

Content-Based Interpretation of the Standard Form Component Summary Measures

Content-based interpretation of the SF-36v2 standard (4-week recall) form PCS measure is facilitated through an examination of the percentage of respondents from the 2009 normative sample at each of 9 levels of PCS *T* scores whose responses to items from those health domain scales most closely associated with the physical health dimension—Physical Functioning, Role-Physical, Bodily Pain, and General Health—were indicative of problems or limitations imposed by the respondents' physical health status. Similarly, content-based interpretation of the MCS measure is facilitated through an examination of the percentages of respondents from the 2009 normative sample at each of 9 levels of MCS *T* scores whose responses to items from those health domain scales most closely associated with the mental health dimension—Vitality, Social Functioning, Role-Emotional, and Mental Health—were indicative of problems or limitations imposed by the respondents' mental health status.

Physical Component Summary (PCS)

Tables 8.1 through 8.5 provide data for the content-based interpretations of SF-36v2 standard form PCS *T* scores relative to limitations in physical and role-functioning activities, pain severity and interference, and ratings of general health.

Physical functioning and PCS. As shown in the column labeled *1* in Table 8.1, more than 90% of the general population reported limitations in performing vigorous activities at each of the lower six PCS score levels (Levels 4–9). The percentage reporting these limitations in the top score levels declined from 12.9% at the highest score level (Level 1) to 69.2% at Level 3. More than twice as many of respondents at Level 2 (26.8%) than at Level 1 indicated similar problems. Overall, this item is most useful in explaining score differences at the highest PCS score levels.

Limitations in moderate activities (Column 2) were more directly related to PCS score level, with 1.4% of the general population reporting such limitations at the highest level (Level 1) and 100% reporting limitations at the lowest level (Level 9). A noticeable percentage (10.4%) of respondents in the upper half of the “average” score range (Level 3, *T*-score range = 50.0–54.9) began reporting problems in performing moderate activities. This percentage more than tripled at the next lower score level (Level 4) and continued to linearly increase through

Level 9. A similar yet more gradual pattern of increasing problems with decreasing PCS scores was seen in lifting and carrying groceries (Column 3). Overall, the moderate activities item is most useful in interpreting score differences in the middle PCS score levels, while the lifting/carrying groceries item's utility is found at both the middle and highest levels.

Difficulties climbing one flight or multiple flights of stairs proved to be a useful indicator of overall physical health status across various PCS score levels. As shown in Table 8.1, problems in climbing multiple flights of stairs (Column 4) started to become apparent even at the higher PCS score levels (3.5% at Level 1, *T*-score range = 60+) and rose quickly through the middle levels, plateauing at 100% at Level 8. Climbing one flight of stairs also was useful in interpreting score differences, beginning in the middle score levels (16.2% at Level 4) and extending into the lower score levels (99.1% at Level 9).

Limitations in bending, kneeling, or stooping (Table 8.2, Column 6) were useful in interpreting score differences across the highest and middle PCS score levels, with 1.7% reporting difficulties in these activities at the highest level (Level 1, *T*-score range = 60+), increasing to 9.0% at Level 2, and finally plateauing at 98% at Level 8 (*T*-score range = 25.0–29.9).

A pattern of increasing percentages of reports of limitations similar to that seen for bending/kneeling/stooping was seen in walking more than a mile (Table 8.2, Column 7) through the highest and middle PCS score levels. The reports of limitations in walking several hundred yards (Column 8) and 100 yards (Column 9) through the PCS levels also were similar, with meaningful percentages beginning to appear at score Level 3 (5.1% for walking several hundred yards, 3.8% for walking 100 yards) and reaching 98.2% and 96.5%, respectively, at Level 9. Overall, both types of limitations are useful in interpreting score differences across the 9 score levels.

Limitations in bathing oneself (Table 8.2, Column 10) became apparent beginning with PCS scores that fall at the low end of the average range and steadily increased with scores below that range. At Level 5 (*T*-score range = 40.0–44.9), 12.8% reported such limitations, increasing by more than threefold (39.3%) at Level 7, and then topping out at 75.4% at the lowest score level (Level 9). Thus, this item is useful in interpreting PCS scores throughout both the middle and lower levels.

Figures 8.1 and 8.2 present graphs of the percentage of the sample scoring at each PCS score level that reported each limitation or characteristic defined or evaluated in Tables 8.1 and 8.2, respectively.

Role functioning and PCS. As shown in Table 8.3, each of the four RP items appears to be most useful for interpreting differences in PCS scores at the middle and lower levels. In general, respondents with PCS *T* scores above the population mean did not report that their physical health had led them to cutting down on time spent at work or other activities (Column 1), accomplishing less than they would like (Column 2), being limited in the kind of work or other activities (Column 3), or having difficulty performing work or other activities (Column 4) either *all* or *most of the time*.

Figure 8.3 presents a graph of the percentage of the sample scoring at each PCS score level that reported each limitation or characteristic defined or evaluated in Table 8.3.

Bodily pain and PCS. Table 8.4 examines the relationship between PCS score levels and (a) the perceived severity of bodily pain (Columns 1 and 2) and (b) the degree to which pain interfered with one's ability to work inside or outside of the home (Columns 3 and 4). The point at which the report of *severe* or *very severe* pain appeared to be most useful for interpreting PCS score differences was Level 4, the low end of the average range of scores for the general population. At Level 4, 4.0% of the general population reported *severe* or *very severe* pain (Column 1). The point at which the report of pain interfering with work *quite a lot* or *extremely* appeared to be most useful for interpreting PCS score differences was also Level 4 (Column 3). Here, 4.2% of the general population reported such a level of pain interference. For both pain-related problems, the percentages increased with each score level change towards the bottom of the PCS score distribution, indicating that the usefulness of these item responses extends across the middle and lower PCS score level ranges. As expected, the percentage of those reporting *little* or *no* problems with either pain or its interference with work (Columns 2 and 4, respectively) decreased from the highest to the lowest score levels, indicating the usefulness of positive responses to the pain severity item across all PCS score levels and the pain interference item in the middle and lowest score ranges.

Figure 8.4 presents a graph of the percentage of the sample scoring at each PCS score level that reported each limitation or characteristic defined or evaluated in Table 8.4.

General health and PCS. As Table 8.5 shows, the responses to each of the five GH items that were most indicative of general health problems (Columns 1–5) appeared to be useful across all PCS score levels, with the percentages of those giving such responses increasing from the highest to the lowest levels. One interest-

ing aspect of the percentage distributions in Table 8.5 is that at even the lowest PCS score level (Level 9), not everyone reported *fair* or *poor* health (79.8%, Column 1), getting sick easier as *mostly* or *definitely true* (32.7%, Column 2), or health expected to get worse as *mostly* or *definitely true* (54.4%, Column 4).

Figure 8.5 presents a graph of the percentage of the sample scoring at each PCS score level that reported each limitation or characteristic defined or evaluated in Table 8.5.

Mental Component Summary (MCS)

Tables 8.6 through 8.9 provide data for the content-based interpretations of SF-36v2 standard form MCS *T* scores relative to reported limitations in vitality, social and role functioning, and mental health.

Vitality and MCS. Table 8.6 shows that the responses to each of the four VT items that are most indicative of fatigue and problems with energy level (Columns 1–4) appeared to be useful across all MCS score levels, with the percentages of those giving such responses increasing from the higher to the lower score levels. Figure 8.6 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.6.

Social functioning and MCS. As shown in Table 8.7, the percentages of those indicating their physical or emotional health resulted in significant (Column 1) and frequent (Column 3) interference in their social activities increased in a linear manner with decreasing MCS *T*-score levels. The opposite was true for those who reported that such disruption was slight or nonexistent (Column 2) or occurred either never or infrequently (Column 4). Overall, the two SF items are useful in explaining differences across all MCS score levels.

Figure 8.7 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.7.

Role functioning and MCS. Table 8.8 shows the percentages of respondents cutting down on the amount of time spent on work or other activities (Column 1), accomplishing less than they would like (Column 2), and performing work or other activities less carefully (Column 3) either *most* or *all of the time*. With regard to each of these three areas of functioning, there was a progressive increase in the percentages reporting limitations from the highest to lowest MCS score levels beginning at about Level 2, which includes the above-average MCS score for the U.S. general population. Each of the three RE items is useful in interpreting MCS score differences across the middle and lower levels.

Table 8.1

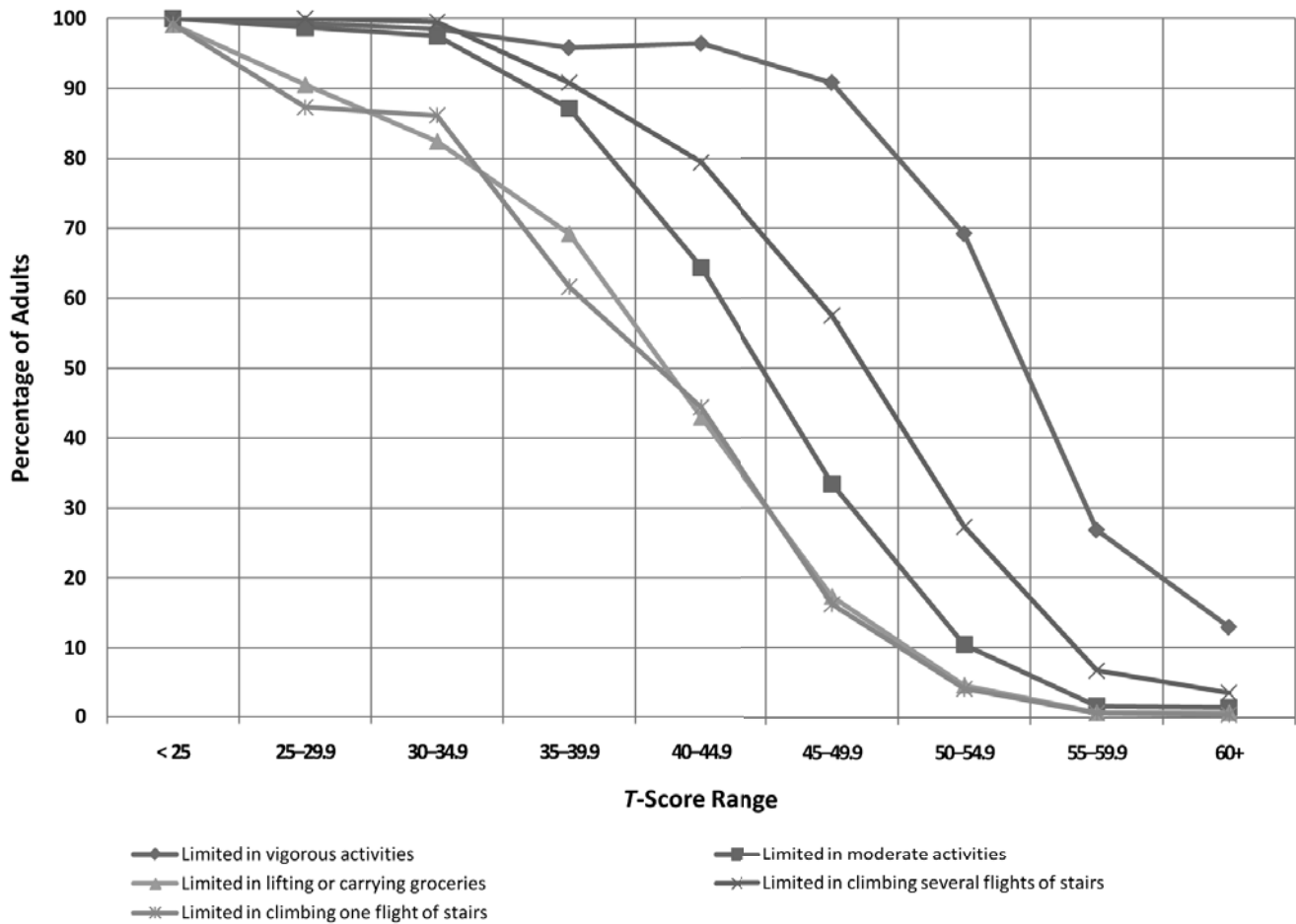
Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

PCS T-Score Level	T Scores		n	Limited in vigorous activities ^a	Limited in moderate activities ^b	Limited in lifting or carrying groceries ^c	Limited in climbing several flights of stairs ^d	Limited in climbing one flight of stairs ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	60+	61.58	289	12.9	1.4	0.7	3.5	0.4
2	55-59.9	57.49	1,233	26.8	1.6	0.7	6.7	0.6
3	50-54.9	52.67	912	69.2	10.4	4.6	27.3	4.1
4	45-49.9	47.71	502	90.8	33.4	17.3	57.5	16.2
5	40-44.9	42.42	358	96.4	64.4	43.0	79.4	44.4
6	35-39.9	37.54	265	95.8	87.1	69.2	90.8	61.6
7	30-34.9	32.79	201	98.5	97.5	82.4	99.5	86.1
8	25-29.9	27.65	150	99.3	98.7	90.5	100.0	87.3
9	< 25	20.64	114	100.0	100.0	99.1	100.0	99.1

^a% reporting any limitations in vigorous activities (Item 3a).
^b% reporting any limitations in moderate activities (Item 3b).
^c% reporting any limitations in lifting or carrying groceries (Item 3c).
^d% reporting any limitations in climbing several flights of stairs (Item 3d).
^e% reporting any limitations in climbing one flight of stairs (Item 3e).

(continued)

Figure 8.1 Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)



(continued)

Table 8.2

Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024) (continued)

PCS T-Score Level	T Scores		n	Limited in bending, kneeling, or stooping ^f	Limited in walking more than a mile ^g	Limited in walking several hundred yards ^h	Limited in walking 100 yards ⁱ	Limited in bathing yourself ^j
	Range	Mean		(6) %	(7) %	(8) %	(9) %	(10) %
1	60+	61.58	289	1.7	2.4	0.0	0.4	0.0
2	55-59.9	57.49	1,233	9.0	4.1	0.7	0.3	0.1
3	50-54.9	52.67	912	33.7	21.2	5.1	3.8	1.1
4	45-49.9	47.71	502	60.4	49.6	20.1	13.8	3.2
5	40-44.9	42.42	358	82.3	79.4	46.8	33.2	12.8
6	35-39.9	37.54	265	87.1	92.0	73.2	56.2	21.4
7	30-34.9	32.79	201	96.5	98.0	90.6	75.6	39.3
8	25-29.9	27.65	150	98.0	100.0	97.3	86.7	51.0
9	< 25	20.64	114	98.3	100.0	98.2	96.5	75.4

^f% reporting any limitations in bending, kneeling, or stooping (Item 3f).

^g% reporting any limitations in walking more than a mile (Item 3g).

^h% reporting any limitations in walking several hundred yards (Item 3h).

ⁱ% reporting any limitations in walking 100 yards (Item 3i).

^j% reporting any limitations in bathing or dressing oneself (Item 3j).

Figure 8.2 Percentage of Adults Reporting Limitations in Physical Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024) (continued)

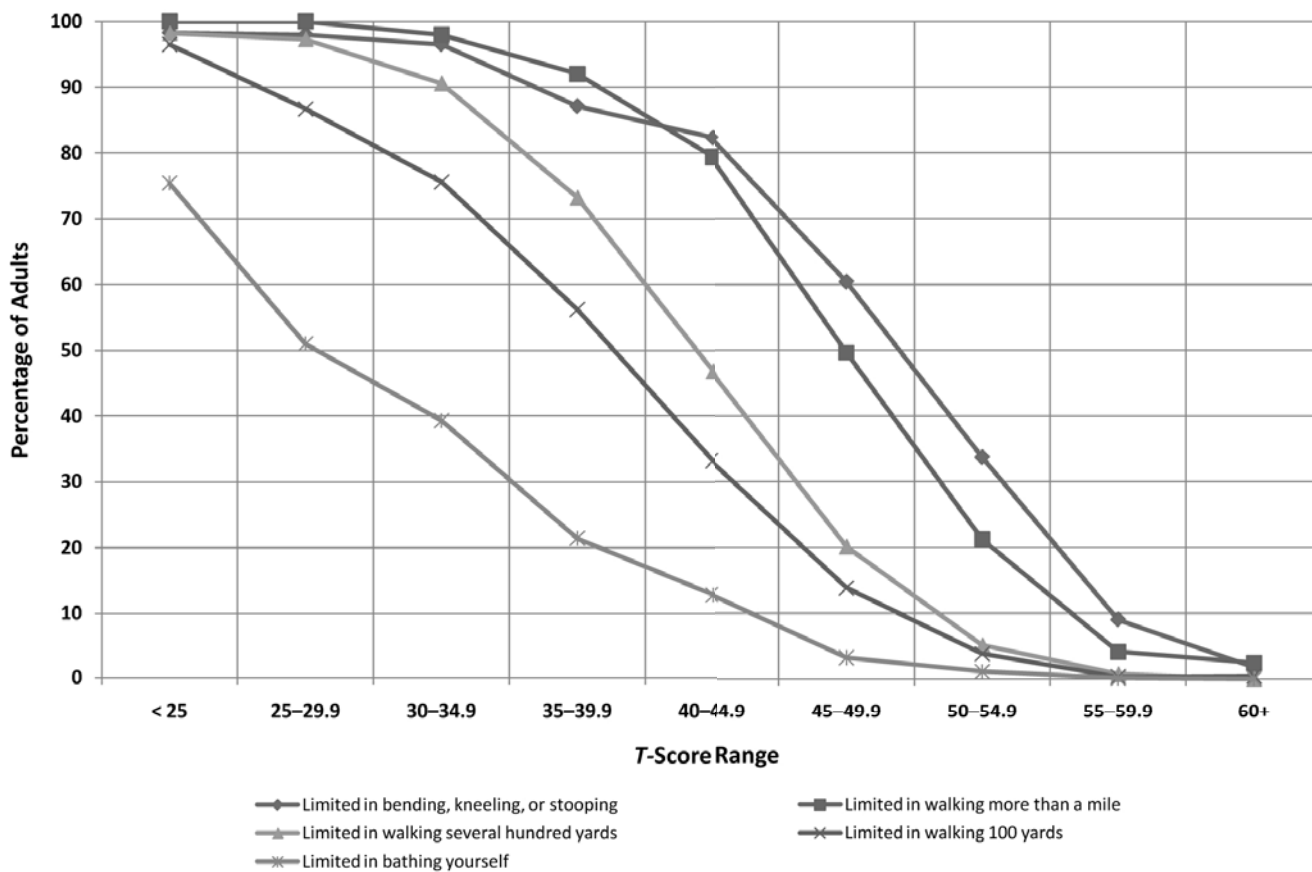


Table 8.3

Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

PCS T-Score Level	T Scores		n	Cut down time at work most or all of the time ^a	Accomplished less most or all of the time ^b	Limited in the kind of work or other activities most or all of the time ^c	Difficulty at work most or all of the time ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	61.58	289	0.4	0.0	0.0	0.0
2	55–59.9	57.49	1,233	0.2	0.6	0.0	0.1
3	50–54.9	52.67	912	0.2	1.8	0.3	0.2
4	45–49.9	47.71	502	3.6	5.6	2.8	2.8
5	40–44.9	42.42	358	7.5	15.2	12.6	15.4
6	35–39.9	37.54	265	22.4	33.6	33.8	32.3
7	30–34.9	32.79	201	38.2	62.0	60.7	54.5
8	25–29.9	27.65	150	65.3	78.0	88.0	86.7
9	< 25	20.64	114	93.0	99.1	98.3	99.1

^a% reporting having cut down amount of time spent on work or other activities *most or all of the time* (Item 4a).

^b% reporting having accomplished less than they would like *most or all of the time* (Item 4b).

^c% reporting being limited in the kind of work or other activities *most or all of the time* (Item 4c).

^d% reporting having difficulty performing work or other activities *most or all of the time* due to physical health (Item 4d).

Figure 8.3 Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

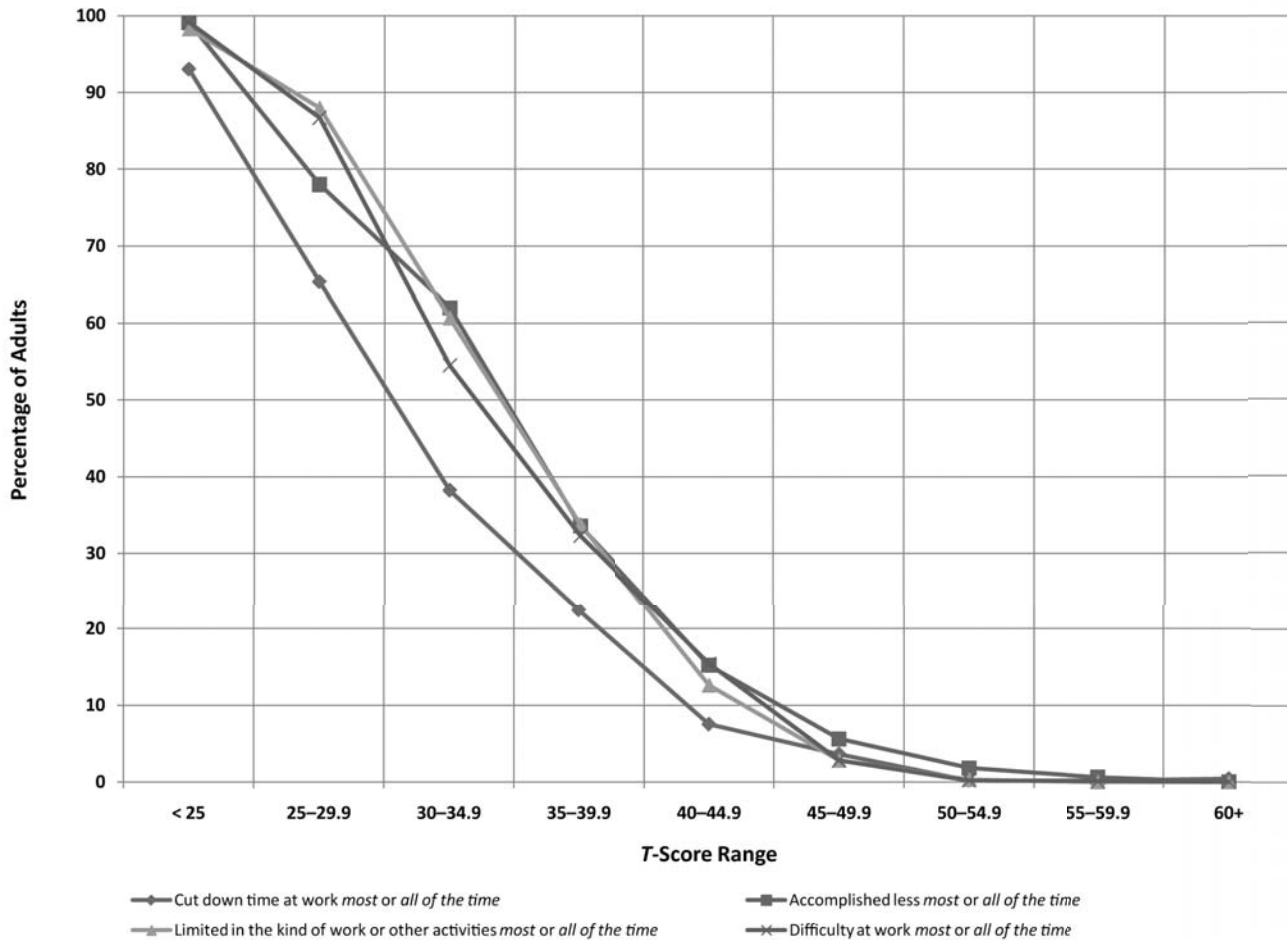


Table 8.4

Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

PCS T-Score Level	T Scores		n	Severe or very severe pain ^a	No or very mild pain ^b	Quite a lot or extreme interference with normal work ^c	Little or no interference with work ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	61.58	289	0.0	98.6	0.0	99.7
2	55–59.9	57.49	1,233	0.0	88.6	0.0	99.7
3	50–54.9	52.67	912	0.8	56.4	0.1	95.1
4	45–49.9	47.71	502	4.0	33.3	4.2	84.6
5	40–44.9	42.42	358	10.6	16.3	9.5	57.3
6	35–39.9	37.54	265	18.6	12.1	20.8	37.1
7	30–34.9	32.79	201	31.3	7.0	42.5	21.5
8	25–29.9	27.65	150	49.7	1.3	60.8	6.8
9	< 25	20.64	114	74.6	0.9	84.1	8.0

^a% reporting very severe or severe bodily pain (Item 7).

^b% reporting no or very mild pain (Item 7).

^c% reporting that pain interferes with normal work (inside and outside the home) quite a lot or extremely (Item 8).

^d% reporting that pain interferes with normal work (inside and outside the home) a little bit or not at all (Item 8).

Figure 8.4 Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

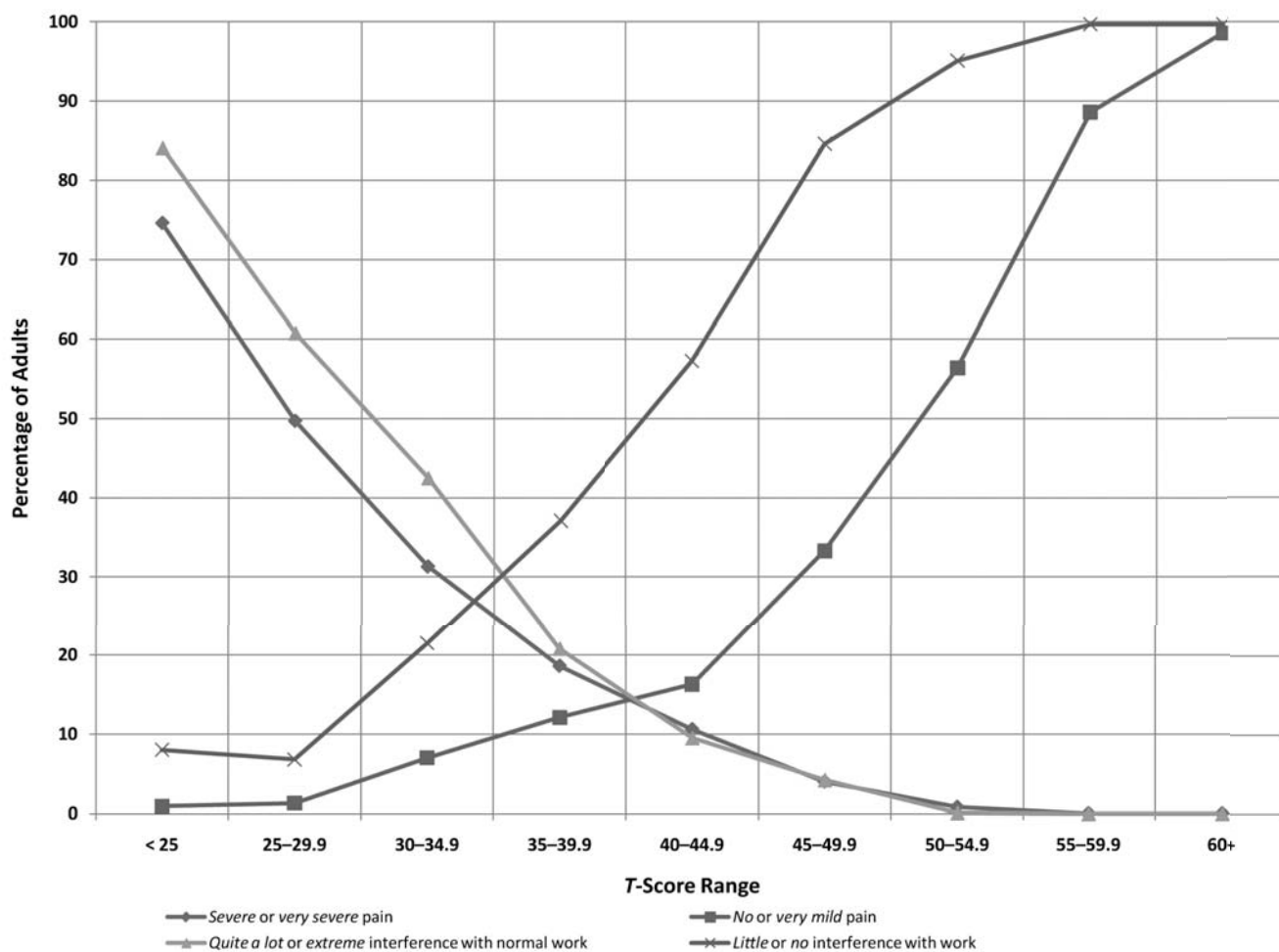


Table 8.5

Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

PCS T-Score Level	T Scores		n	Fair or poor health ^a (1) %	Getting sick easier mostly or definitely true ^b (2) %	As healthy as anybody mostly or definitely false ^c (3) %	Health expected to get worse mostly or definitely true ^d (4) %	Health is excellent mostly or definitely false ^e (5) %
	Range	Mean						
1	60+	61.58	289	0.4	0.7	1.0	3.1	3.8
2	55-59.9	57.49	1,233	1.2	2.8	4.9	5.8	5.8
3	50-54.9	52.67	912	5.9	7.4	12.4	18.5	18.9
4	45-49.9	47.71	502	13.9	11.4	21.6	24.3	41.2
5	40-44.9	42.42	358	22.7	15.2	27.0	26.7	49.9
6	35-39.9	37.54	265	37.5	17.1	38.6	30.9	62.9
7	30-34.9	32.79	201	56.2	17.4	54.2	37.3	77.0
8	25-29.9	27.65	150	68.7	24.0	67.1	39.3	79.2
9	< 25	20.64	114	79.8	32.7	78.1	54.4	93.0

^a% reporting fair or poor health (Item 1).

^b% reporting getting sick easier as mostly true or definitely true (Item 11a).

^c% reporting being as healthy as anybody they know as mostly false or definitely false (Item 11b).

^d% reporting health expected to get worse as mostly true or definitely true (Item 11c).

^e% reporting health is excellent as mostly false or definitely false (Item 11d).

Figure 8.5 Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

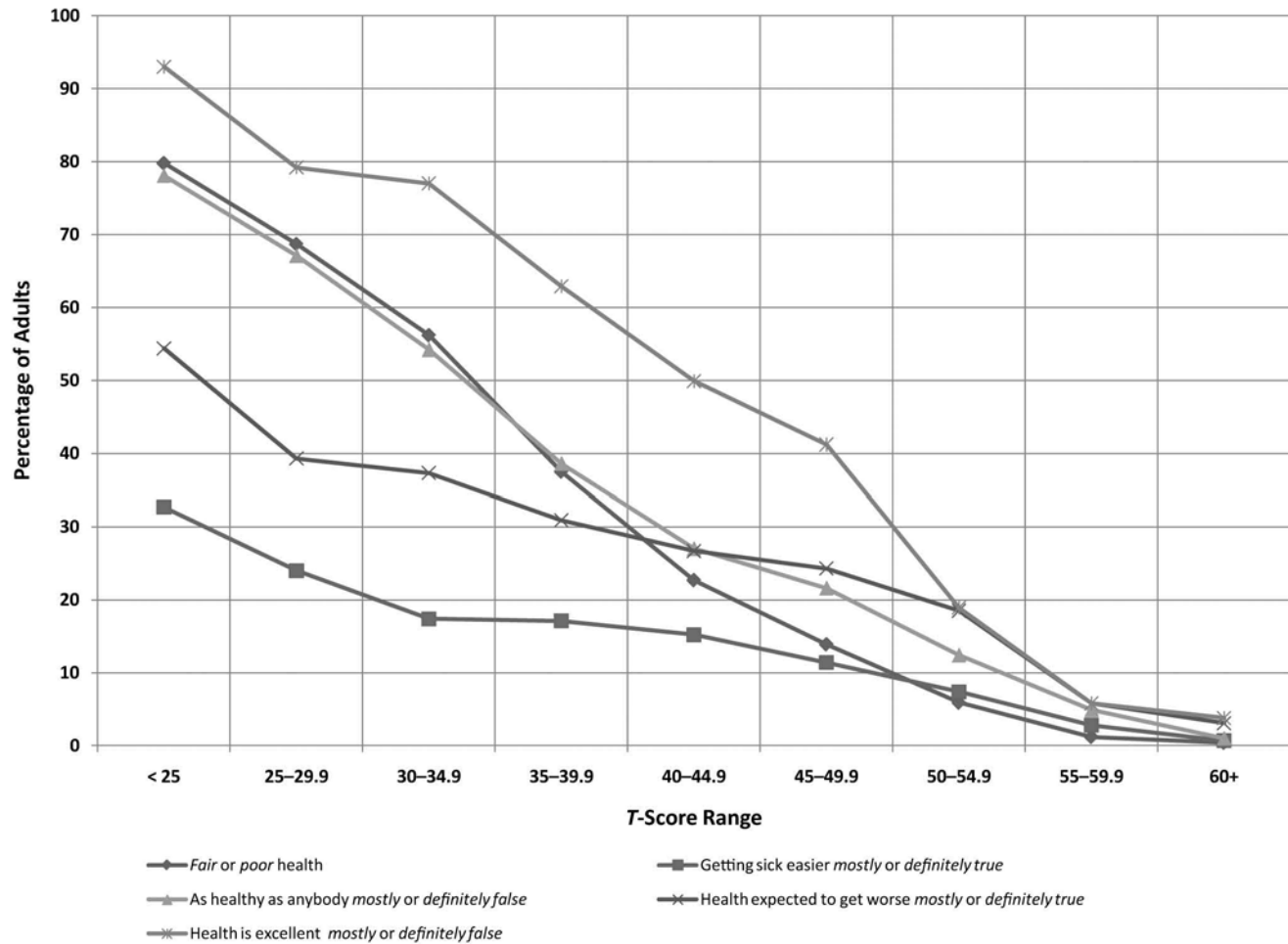


Table 8.6

Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

MCS T-Score Level	T Scores		n	Feeling full of life <i>little or none of the time</i> ^a	Having a lot of energy <i>little or none of the time</i> ^b	Feeling worn out <i>most or all of the time</i> ^c	Feeling tired <i>most or all of the time</i> ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	62.24	492	4.5	7.0	1.8	2.7
2	55–59.9	57.39	1,251	3.8	9.0	2.8	5.4
3	50–54.9	52.82	886	10.7	18.7	10.1	17.3
4	45–49.9	47.79	509	23.4	33.8	16.7	32.8
5	40–44.9	42.56	338	34.5	49.3	24.3	36.7
6	35–39.9	37.69	219	46.1	58.1	45.9	55.5
7	30–34.9	32.65	129	59.7	71.3	65.9	78.0
8	25–29.9	27.53	103	85.3	78.4	74.8	76.7
9	< 25	19.42	97	93.8	89.7	87.6	91.8

^a% reporting feeling full of life *little or none of the time* (Item 9a).

^b% reporting having a lot of energy *little or none of the time* (Item 9e).

^c% reporting feeling worn out *most or all of the time* (Item 9g).

^d% reporting feeling tired *most or all of the time* (Item 9i).

Figure 8.6 Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

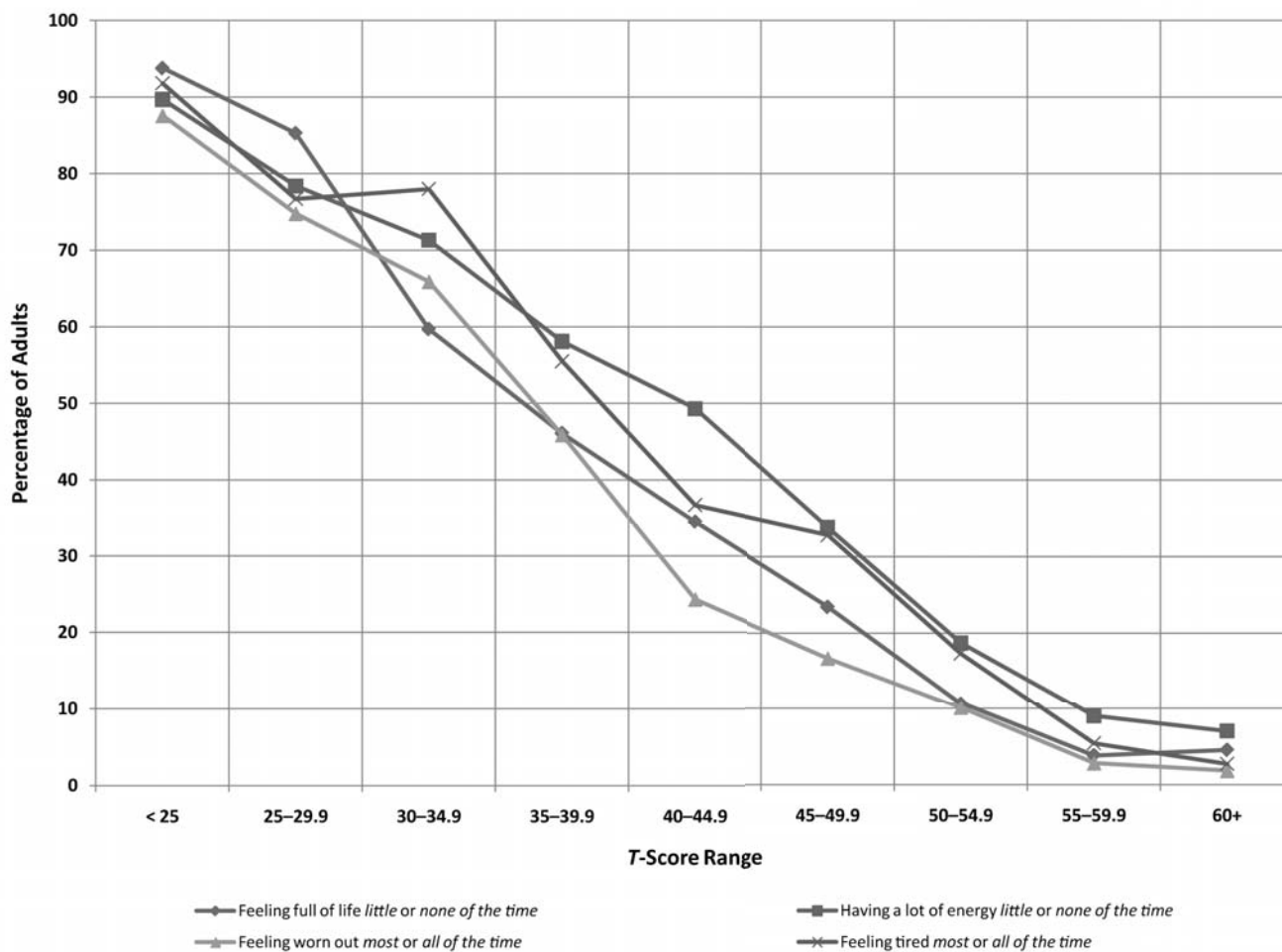


Table 8.7

Percentage of Adults Reporting Limitations in Social Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

MCS T-Score Level	T Scores		n	Health interfered with social activities <i>moderately, quite a bit, or extremely</i> ^a	Health interfered with social activities <i>slightly or not at all</i> ^b	Health interfered with social activities <i>most or all of the time</i> ^c	Health interfered with social activities <i>little or none of the time</i> ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	62.24	492	3.1	97.0	0.2	97.2
2	55–59.9	57.39	1,251	3.6	96.4	1.8	95.4
3	50–54.9	52.82	886	9.4	90.6	2.9	89.5
4	45–49.9	47.79	509	15.2	84.8	5.7	77.8
5	40–44.9	42.56	338	33.6	66.4	11.6	59.2
6	35–39.9	37.69	219	47.3	52.8	18.9	37.8
7	30–34.9	32.65	129	69.0	31.0	37.2	16.3
8	25–29.9	27.53	103	83.5	16.5	61.2	15.5
9	< 25	19.42	97	94.9	5.2	86.6	2.1

^a% reporting physical or emotional problems interfering with social activities *moderately, quite a bit, or extremely* (Item 6).

^b% reporting physical or emotional problems interfering with social activities *slightly or not at all* (Item 6).

^c% reporting physical or emotional problems interfering with social activities *most or all of the time* (Item 10).

^d% reporting physical or emotional problems interfering with social activities *little or none of the time* (Item 10).

Figure 8.7 Percentage of Adults Reporting Limitations in Social Functioning at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

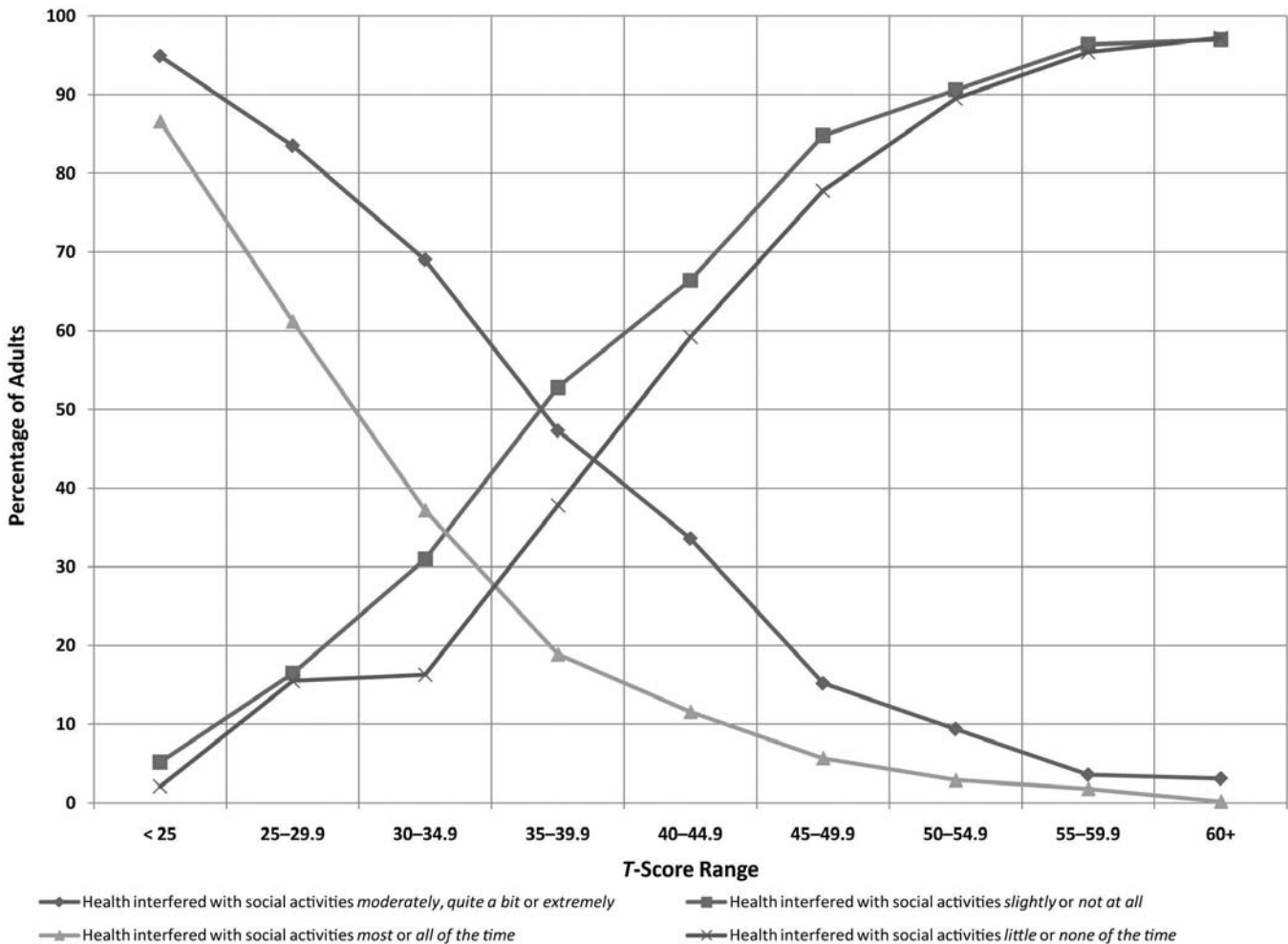


Figure 8.8 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.8.

Mental health and MCS. As would be expected, responses to each of the five MH scale items had a fairly clear relationship with the MCS measure, with progressively more respondents generally reporting emotional problems from the middle score levels to the lower levels (see Table 8.9). Very few (if any) respondents reported emotional problems at the higher MCS score levels, while a relatively high percentage reported these same types of problems at the lower score levels. At the same time, unique patterns of progressively increasing percentages emerged for each of the MH items. Respondents generally began reporting being very nervous (Column 1), down in the dumps (Column 2), or downhearted and depressed (Column 4) *most* or *all of the time* at MCS score Level 4 or 5, while reports of being calm (Column 3) or happy (Column 5) *little* or *none of the time* began at higher score levels (Level 2 or 3). In addition, at the lower score levels, none of the reported percentages reached 100%. Regardless, it is accurate to characterize the usefulness of the MH items in interpreting MCS scores as occurring generally at the middle and lower score levels.

Figure 8.9 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.9.

Content-Based Interpretation of the Standard Form Health Domain Scales

Content-based interpretations of the SF-36v2 standard form health domain scales are facilitated through an examination of the percentage of respondents from the 2009 U.S. general population normative sample whose responses to each item from each health domain scale were indicative of problems or limitations imposed by the respondents' health status, at each score level of a given health domain scale.

Physical Functioning (PF)

Tables 8.10 and 8.11 present the percentages of respondents reporting limitations at the 8 PF scale score levels. With the exception of bathing oneself (Table 8.11, Column 10), at least some limitations in performing the wide range of physical activities measured by the PF scale began to be reported by significant percentages

of the SF-36v2 respondents at the higher score levels (Levels 1–3). For example, nearly one quarter (24.3%) of those scoring in the highest PF *T*-score range (Level 1) reported some limitations in vigorous activities (Table 8.10, Column 1). The percentage nearly quadrupled (91.4%) at Level 2, which included the mean PF mean *T* score of 50.

For the most part, a linear pattern of increasing percentages of reported limitations appeared across the range of score levels, from the highest to the lowest levels. Overall, examination of the patterns of percentages for the PF items in Tables 8.10 and 8.11 indicates that limitations in moderate activities (Table 8.10, Column 2), lifting/carrying groceries (Table 8.10, Column 3), climbing several flights of stairs (Table 8.10, Column 4), climbing one flight of stairs (Table 8.10, Column 5), bending/kneeling/stooping (Table 8.11, Column 6), and walking 100 yards (Table 8.11, Column 9) are all useful in interpreting differences in PF scores across all score levels. In contrast, limitations in vigorous activities (Table 8.10, Column 1) are useful at the highest *T*-score levels, while walking more than a mile (Table 8.11, Column 7) and walking several hundred yards (Table 8.11, Column 8) are useful at both the highest and middle *T*-score levels. Finally, limitations in bathing oneself (Table 8.11, Column 10) are useful for interpreting score differences at both the middle and lowest score levels.

Figures 8.10 and 8.11 present graphs of the percentage of the sample scoring at each PF score level that reported each limitation or characteristic defined or evaluated in Tables 8.10 and 8.11, respectively.

Role-Physical (RP)

Table 8.12 provides data for content-based interpretation of the RP scale, with responses of *most* or *all of the time* to any of this scale's four items indicating limitations in role functioning due to the respondent's physical health. Overall, cutting down time at work or other activities (Column 1) is most useful in interpreting score differences at the lowest RP score levels, while limitations in the kind of work or other activities (Column 3) is most useful at the middle score levels. In contrast, responses to the items having to do with accomplishing less (Column 2) and experiencing difficulty at work (Column 4) are useful in interpreting differences at both the middle and lowest levels of RP scores.

Figure 8.12 presents a graph of the percentage of the sample scoring at each RP score level that reported each limitation or characteristic defined or evaluated in Table 8.12.

Table 8.8

Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

MCS T-Score Level	T Scores		n	Cut down time at work most or all of the time ^a	Accomplished less most or all of the time ^b	Did work less carefully most or all of the time ^c
	Range	Mean		(1) %	(2) %	(3) %
1	60+	62.24	492	0.0	0.0	0.0
2	55–59.9	57.39	1,251	0.1	0.2	0.2
3	50–54.9	52.82	886	0.7	1.0	0.2
4	45–49.9	47.79	509	3.0	4.2	2.4
5	40–44.9	42.56	338	6.6	7.4	3.9
6	35–39.9	37.69	219	16.4	18.7	10.6
7	30–34.9	32.65	129	27.9	40.3	21.7
8	25–29.9	27.53	103	53.4	63.1	34.0
9	< 25	19.42	97	76.0	90.7	61.9

^a% reporting cutting down amount of time spent on work or other activities *most* or *all of the time* (Item 5a).

^b% reporting accomplished less than they would like *most* or *all of the time* (Item 5b).

^c% reporting did work or other activities less carefully *most* or *all of the time* (Item 5c).

Figure 8.8 Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

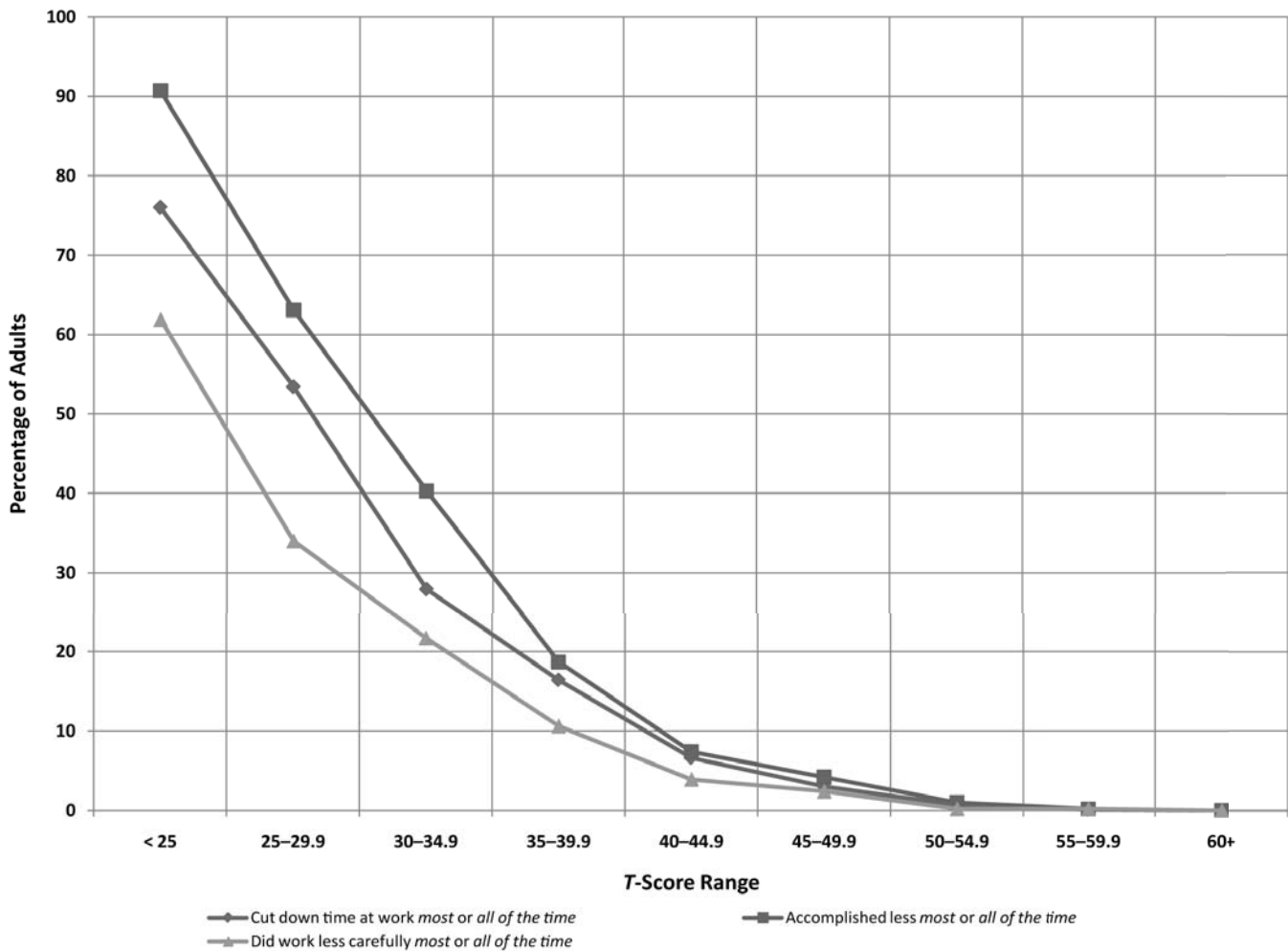


Table 8.9

Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

MCS T-Score Level	T Scores		n	Been very nervous most or all of the time ^a	Down in dumps most or all of the time ^b	Calm little or none of the time ^c	Downhearted and depressed most or all of the time ^d	Happy little or none of the time ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	60+	62.24	492	0.0	0.0	0.8	0.0	0.2
2	55-59.9	57.39	1,251	0.0	0.2	1.7	0.0	0.7
3	50-54.9	52.82	886	0.8	0.5	6.6	0.1	3.3
4	45-49.9	47.79	509	5.5	1.4	23.7	0.8	14.2
5	40-44.9	42.56	338	6.6	3.0	38.4	3.6	20.4
6	35-39.9	37.69	219	18.0	7.3	46.3	11.5	30.6
7	30-34.9	32.65	129	35.7	18.8	63.6	34.4	46.1
8	25-29.9	27.53	103	43.7	45.6	83.5	68.6	65.1
9	< 25	19.42	97	63.9	67.0	93.8	86.5	95.9

^a% reporting being very nervous *most or all of the time* (Item 9b).

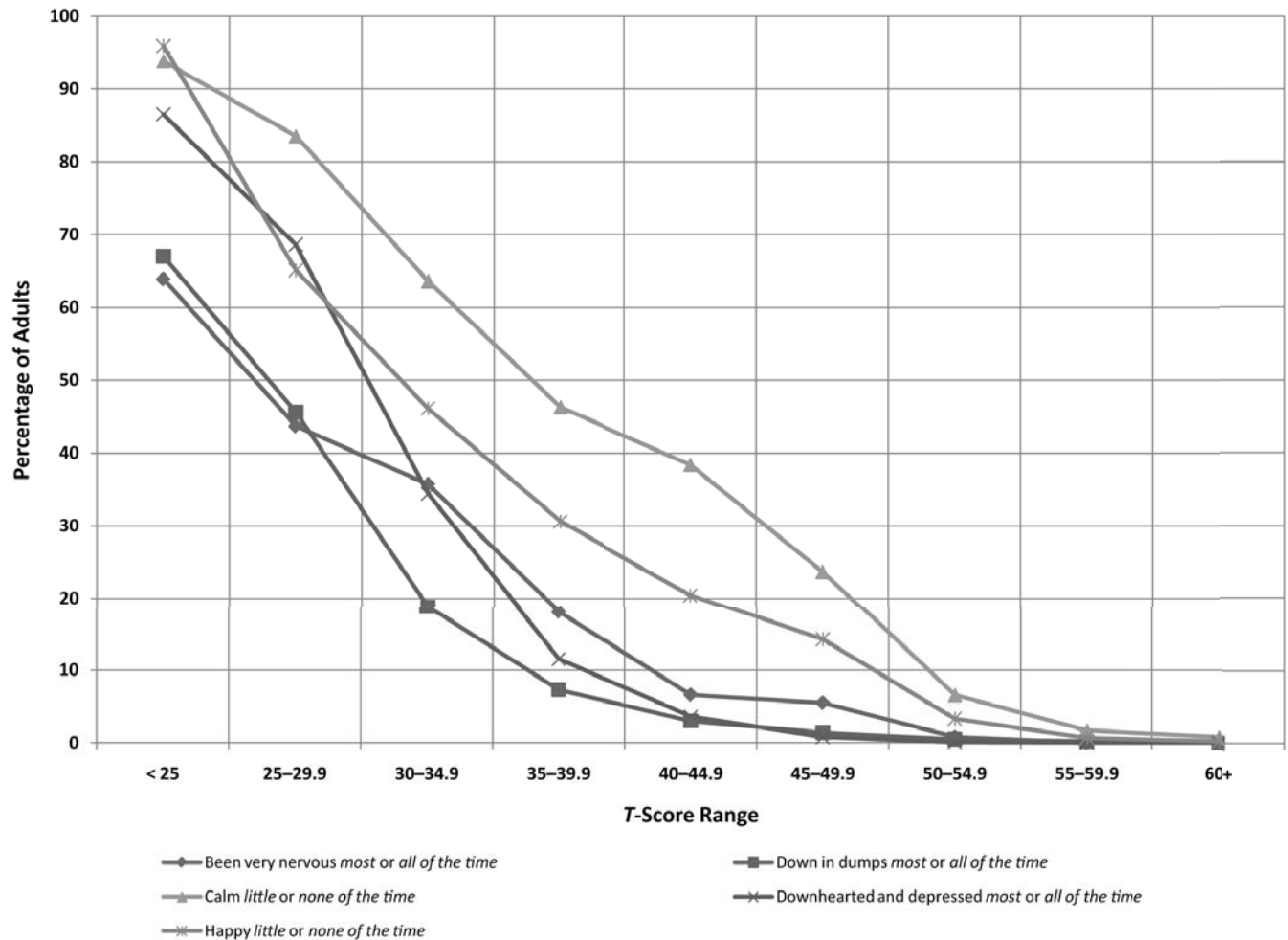
^b% reporting being so down in the dumps that nothing could cheer them up *most or all of the time* (Item 9c).

^c% reporting being calm and peaceful *little or none of the time* (Item 9d).

^d% reporting being downhearted and depressed *most or all of the time* (Item 9f).

^e% reporting being happy *little or none of the time* (Item 9h).

Figure 8.9 Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)



Bodily Pain (BP)

As indicated in Table 8.13, reports of *severe* or *very severe* pain (Column 1) did not begin to occur until BP score Level 5 (1.8%, *T*-score range = 40.0–44.9), but then quickly increased and reached 100% at the highest level (Level 8). Similarly, *quite a lot* and *extreme* interference with normal work (Column 3) were not reported until Level 6 (0.9%, *T*-score range = 35.0–39.9) but became quite prevalent at the next score level (85.4%, Level 7). Overall, reports of *severe* or *very severe* pain are useful in interpreting score differences at the middle and lowest score levels, while reports of *quite a lot* and *extreme* interference with normal work are most useful for this purpose at the lowest score levels.

As expected, reports of *no* or *mild* pain (Column 2) and *little* or *no* interference with work (Column 4) were more common at the highest BP score levels but decreased as scores decreased. Also, notable and rapid drops in reports of *no* or *mild* pain occurred from Level 2 to Level 3 (100% to 60.7%), and then through Levels 4 and 5 (3.3% to 0%). Overall, both seem to be most useful in interpreting scores in the middle score ranges.

Figure 8.13 presents a graph of the percentage of the sample scoring at each BP score level that reported each limitation or characteristic defined or evaluated in Table 8.13.

General Health (GH)

Table 8.14 shows that the percentages of respondents reporting generally negative perceptions of various aspects of their health status began at about GH score Level 3 (*T*-score range = 55.0–59.9). A slow but steady increase occurred in the percentages reporting either getting sick easier as *mostly* or *definitely true* (Column 2) or expecting their health to get worse as *mostly* or *definitely true* (Column 4). In contrast, the percentages of those reporting *fair* or *poor* health (Column 1), being as healthy as anybody as *mostly* or *definitely false* (Column 3), or excellent health as *mostly* or *definitely false* (Column 5) rose quickly from GH score Level 4 or 5 to Level 6, and then to Level 7. Overall, the five GH items appear to be most useful in interpreting score differences at the middle to lower score levels.

Figure 8.14 presents a graph of the percentage of the sample scoring at each GH score level that reported each limitation or characteristic defined or evaluated in Table 8.14.

Vitality (VT)

Inspection of Table 8.15 demonstrates that, like the GH items, the four VT items are most useful in interpreting differences in scores at the middle and lowest levels.

For the most part, indicators of low vitality—feeling full of life (Column 1) or having a lot of energy (Column 2) *little* or *none of the time*, or feeling worn out (Column 3) or tired (Column 4) *most* or *all of the time*—began at either Level 4 (*T* = 50.0–54.9) or Level 5 (*T* = 45.0–49.9). The reports of vitality-related problems then quickly became more common as VT scores progressed to the lower score levels.

Figure 8.15 presents a graph of the percentage of the sample scoring at each VT score level that reported each limitation or characteristic defined or evaluated in Table 8.15.

Social Functioning (SF)

Table 8.16 indicates that the degree (Column 1) and frequency (Column 3) at which physical or emotional health interferes with social activities begins at score levels that fall below the mean SF score (*T* = 50). The table reveals that 6.6% of those scoring at SF score Level 3 (*T* = 45.0–49.9) indicate that health interferes with their social activities either *moderately*, *quite a bit*, or *extremely*. A five-fold increase in the report of interference (33.1%) occurs at the next lower score level (Level 4), and then more than doubles (78.9%) at the next level down (Level 5). The opposite trend is noted when the degree at which physical or emotional health interferes with social activities (Column 2) is reported to be *slightly* or *not at all*. In this case, 100% of all of those scoring in the two highest SF score levels (Levels 1 and 2) and 93.4% of those scoring at Level 3 report slight or no interference of health with their social activities. Similar trends in percentages are noted when one considers those indicating that health interferes with their social activities *most* or *all of the time* (Column 3) or *little* or *none of the time* (Column 4). Overall, reports of no or little problems with regard to the degree (Column 2) and frequency (Column 4) at which physical or emotional health interferes with social activities are useful in interpreting score differences at the middle SF score levels. Responses indicating more significant problems with regard to the degree (Column 1) and frequency (Column 3) are also useful in interpreting differences in the middle score levels.

Figure 8.16 presents a graph of the percentage of the sample scoring at each SF score level that reported each limitation or characteristic defined or evaluated in Table 8.16.

Role-Emotional (RE)

The percentages of respondents reporting problems in role functioning due to emotional problems are presented in Table 8.17. The percentage of the general

population reporting emotional problems resulting in cutting down on time at work (0.2%, Column 1), accomplishing less than one would like (0.7%, Column 2), and doing work or other activities less carefully (0.2%, Column 3) *most or all of the time* began at RE score Level 3 (T -score range = 45.0–49.9), representing the lower half of the average range of scores. The percentages then quickly increased in a linear fashion through the lower score levels, each plateauing at 100% at Level 9 (T -score < 20). Thus, limitations due to emotional problems in the three aspects of work reported in Table 8.17 are most useful for interpreting RE score differences at the middle and lowest T -score levels.

Figure 8.17 presents a graph of the percentage of the sample scoring at each RE score level that reported each limitation or characteristic defined or evaluated in Table 8.17.

Mental Health (MH)

As would be expected, the percentage of respondents reporting problems in their emotional health and well-being quickly increased with decreasing MH scale scores. Table 8.18 demonstrates that reports of feeling calm (Column 3) or happy (Column 5) *little or none of the time* can be useful in interpreting score differences across almost all of the MH score levels. Of interest is the fact that these feelings were reported even in 6.2% and 2.1%, respectively, of respondents at score Level 3, which includes the U.S. general population MH mean score. Reports of being very nervous (Column 1), being down in the dumps (Column 2), and being downhearted and depressed (Column 4) *most or all of the time* generally began to appear at the next lowest score level (Level 4, T -score range = 45.0–49.9), making these items useful for interpreting score differences in the middle and lowest MH score levels.

Figure 8.18 presents a graph of the percentage of the sample scoring at each MH score level that reported each limitation or characteristic defined or evaluated in Table 8.18.

Content-Based Interpretation of the Acute Form Component Summary Measures

As with the standard form results, content-based interpretations of the SF-36v2 acute (1-week recall) form PCS and MCS measures are facilitated through an examination of the percentage of respondents from the 2009 normative sample at each of 8 levels of PCS

and MCS T scores whose responses to items from those health domain scales most closely associated with each health dimension were indicative of problems or limitations imposed by the respondents' health status.

Physical Component Summary (PCS)

Tables 8.19 through 8.23 provide data for the content-based interpretations of SF-36v2 acute (1-week recall) form PCS T scores relative to limitations in physical and role functioning activities, pain severity and interference, and ratings of general health.

Physical functioning and PCS. As shown in Column 1 in Table 8.19, 14.2% of those scoring at the highest PCS level reported at least some limitations in vigorous activities. The percentage nearly doubled to 26.7% at Level 2 and then more than doubled at Level 3 (72.5%), which includes the mean T score for PCS (T -score range = 50.0–54.9). At least 90% of those at each of the lower levels (Levels 4–8) reported such limitations. Overall, this item is most useful in explaining score differences at the highest and middle PCS score levels. Meanwhile, there was a slow but then rapidly increasing percentage of respondents reporting limitations in performing moderate activities (Column 2), reaching 72.5% at Level 5 (T -score range = 40.0–44.9) and making this item useful in explaining score differences at all score levels.

Table 8.19 also shows linear or near linear increases in the percentages of those reporting limitations in carrying groceries (Column 3), climbing several flights of stairs (Column 4), and climbing one flight of stairs (Column 5) with decreasing PCS scores, indicating that each of these three items is also useful in explaining score differences across all score levels. A rather rapid increase in the percentage reporting limitations in climbing several flights of stairs was seen, going from 4.5% at the highest level (Level 1), increasing more than sixfold to 31.0% at Level 3, and then doubling to 62.6% at score Level 4. A slower, more moderate increase in the percentage of respondents reporting limitations from the highest to the lowest score levels was observed for the other two items.

Examining limitations in the remaining five physical activities evaluated by the PF scale relative to PCS score levels, Table 8.20 reveals relatively slow but steady increases in the percentage of respondents reporting limitations in walking several hundred yards (Column 8), walking 100 yards (Column 9), and bathing oneself (Column 10) with decreasing PCS scores. This indicates that each of these three items is useful in explaining score differences across all PCS score levels. More rapidly progressing increases in reported limitations through the

Table 8.10

Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 4,034)

PF T-Score Level	T Scores		n	Limited in vigorous activities ^a	Limited in moderate activities ^b	Limited in lifting or carrying groceries ^c	Limited in climbing several flights of stairs ^d	Limited in climbing one flight of stairs ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	55+	56.99	1,848	24.3	0.2	0.2	1.7	0.0
2	50-54.9	52.90	721	91.4	13.6	3.2	37.6	2.5
3	45-49.9	48.41	456	96.0	43.4	19.8	77.5	13.0
4	40-44.9	42.11	324	94.7	77.6	54.2	91.3	50.3
5	35-39.9	37.66	177	97.2	89.7	69.5	96.6	81.4
6	30-34.9	32.62	219	98.6	96.3	88.1	98.2	89.0
7	25-29.9	26.98	167	96.3	99.4	98.8	99.4	99.4
8	< 25	21.21	122	100.0	100.0	100.0	100.0	100.0

^a% reporting any limitations in vigorous activities (Item 3a).

^b% reporting any limitations in moderate activities (Item 3b).

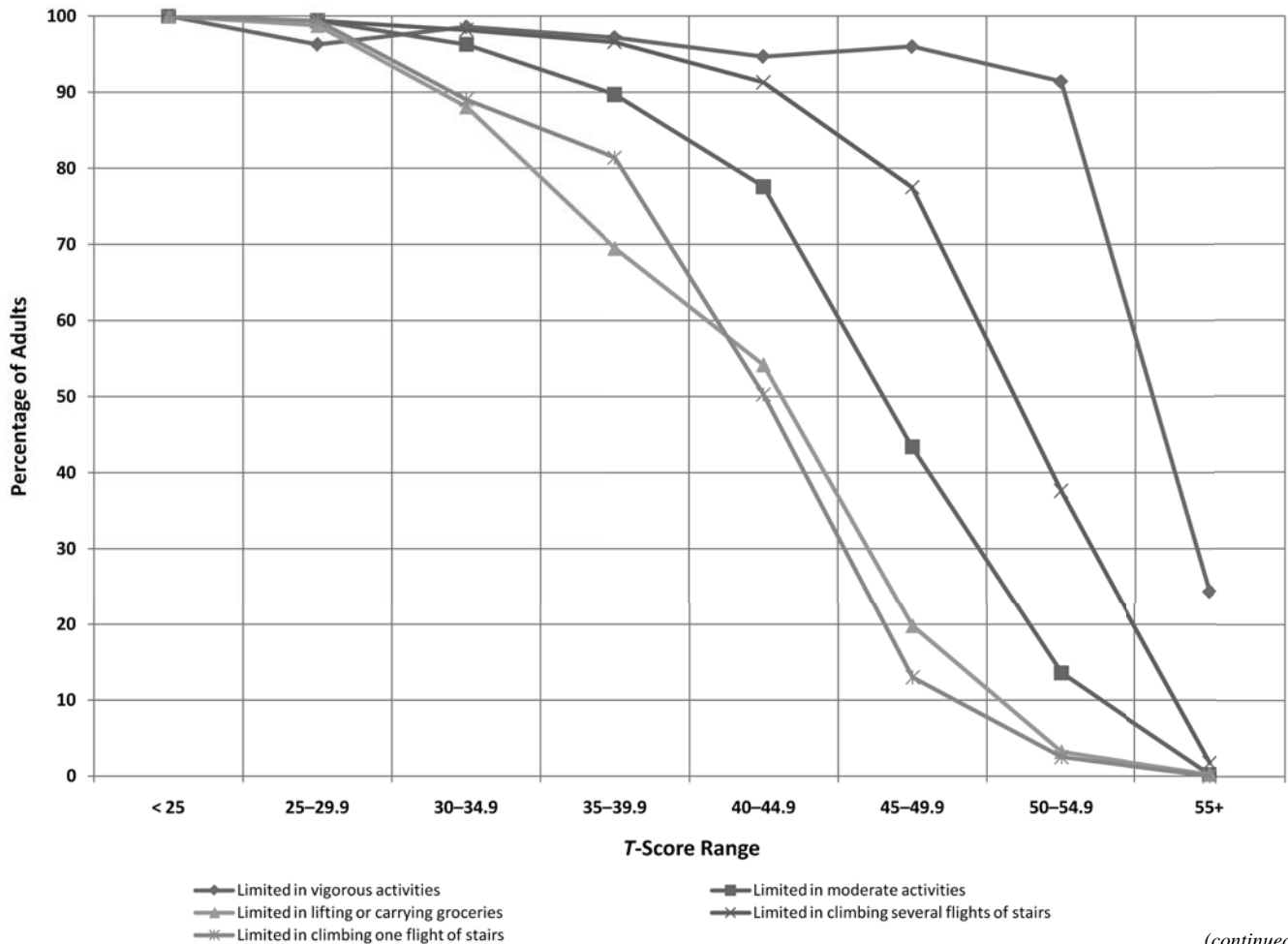
^c% reporting any limitations in lifting or carrying groceries (Item 3c).

^d% reporting any limitations in climbing several flights of stairs (Item 3d).

^e% reporting any limitations in climbing one flight of stairs (Item 3e).

(continued)

Figure 8.10 Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 4,034)



(continued)

Table 8.11

Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 4,034) (continued)

PF T-Score Level	T Scores		n	Limited in bending, kneeling, or stooping ^f	Limited in walking more than a mile ^g	Limited in walking several hundred yards ^h	Limited in walking 100 yards ⁱ	Limited in bathing yourself ^j
	Range	Mean		(6) %	(7) %	(8) %	(9) %	(10) %
1	55+	56.99	1,848	4.3	1.0	0.0	0.1	0.1
2	50–54.9	52.90	721	50.1	25.7	3.0	1.7	0.3
3	45–49.9	48.41	456	72.9	68.6	19.0	9.9	1.8
4	40–44.9	42.11	324	85.4	91.0	59.6	33.4	10.0
5	35–39.9	37.66	177	92.0	97.1	87.9	72.4	21.6
6	30–34.9	32.62	219	95.4	99.5	98.2	87.2	29.7
7	25–29.9	26.98	167	99.4	100.0	100.0	98.2	66.3
8	< 25	21.21	122	100.0	100.0	100.0	100.0	94.3

^f% reporting any limitations in bending, kneeling, or stooping (Item 3f).

^g% reporting any limitations in walking more than a mile (Item 3g).

^h% reporting any limitations in walking several hundred yards (Item 3h).

ⁱ% reporting any limitations in walking 100 yards (Item 3i).

^j% reporting any limitations in bathing or dressing oneself (Item 3j).

Figure 8.11 Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 4,034) (continued)

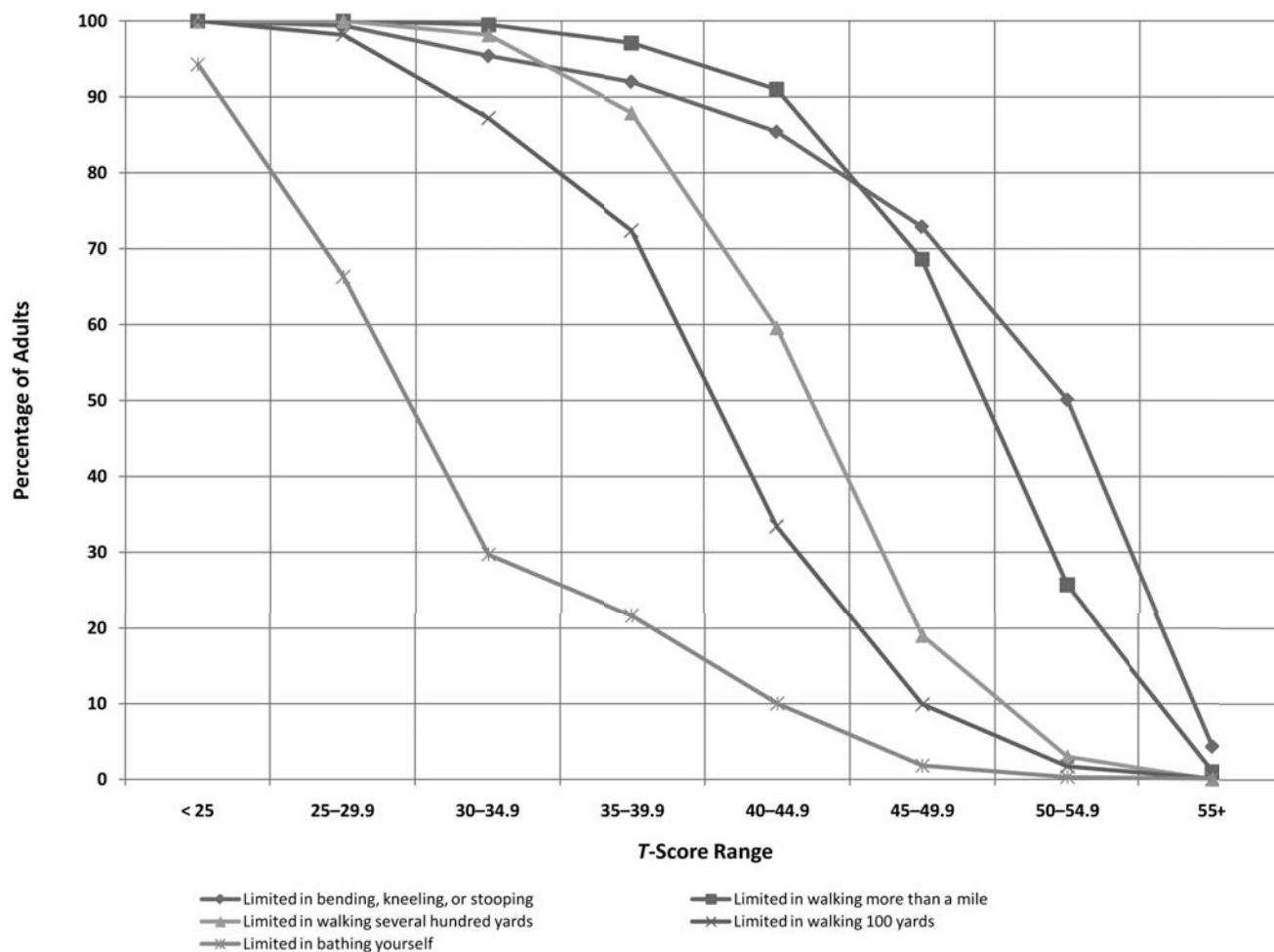


Table 8.12

Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (N = 4,027)

RP T-Score Level	T Scores		n	Cut down time at work most or all of the time ^a	Accomplished less most or all of the time ^b	Limited in the kind of work or other activities most or all of the time ^c	Difficulty at work most or all of the time ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	55+	57.16	1,861	0.0	0.0	0.0	0.0
2	50–54.9	52.86	754	0.4	0.4	0.0	0.1
3	45–49.9	47.43	399	0.3	3.8	0.3	0.5
4	40–44.9	42.58	232	0.0	11.2	5.2	7.4
5	35–39.9	38.35	277	6.9	21.5	18.8	13.8
6	30–34.9	32.12	264	53.2	79.9	81.0	81.0
7	25–29.9	26.83	100	88.0	96.0	100.0	99.0
8	< 25	21.86	140	100.0	100.0	100.0	100.0

^a% reporting having cut down amount of time spent on work or other activities *most or all of the time* (Item 4a).

^b% reporting having accomplished less than they would like *most or all of the time* (Item 4b).

^c% reporting being limited in the kind of work or other activities *most or all of the time* (Item 4c).

^d% reporting having had difficulty performing work or other activities *most or all of the time* (Item 4d).

Figure 8.12 Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (N = 4,027)

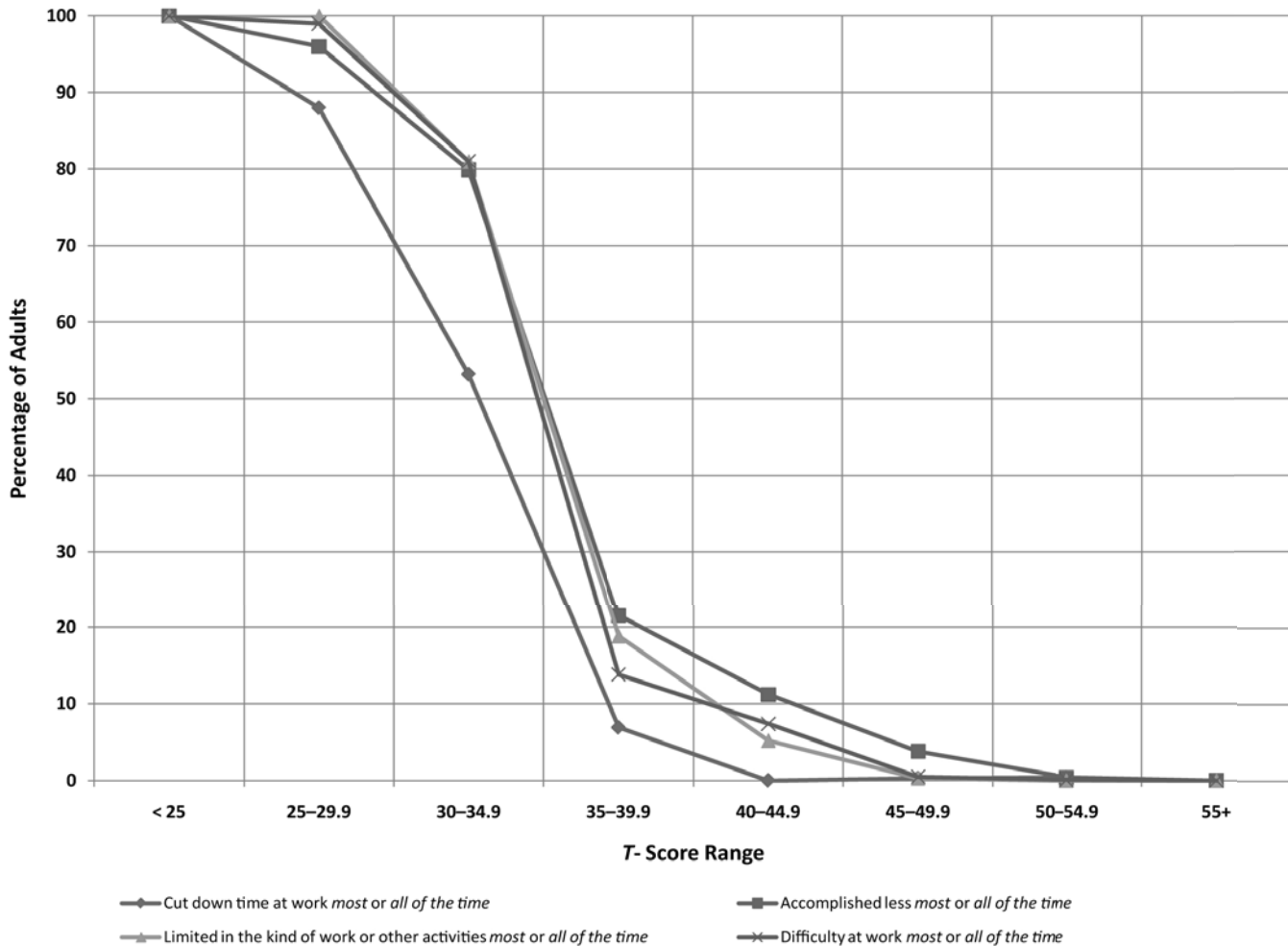


Table 8.13

Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (N = 4,027)

BP T-Score Level	T Scores		n	Severe or very severe pain ^a	No or very mild pain ^b	Quite a lot or extreme interference with normal work ^c	Little or no interference with work ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	62.00	811	0.0	100.0	0.0	100.0
2	55–59.9	55.55	965	0.0	100.0	0.0	100.0
3	50–54.9	51.28	617	0.0	60.7	0.0	100.0
4	45–49.9	46.68	522	0.0	3.3	0.0	96.7
5	40–44.9	42.35	333	1.8	0.0	0.0	75.1
6	35–39.9	38.25	350	7.1	0.0	0.9	7.2
7	30–34.9	32.46	312	60.6	0.0	85.4	0.0
8	< 30	24.48	117	100.0	0.0	99.2	0.0

^a% reporting severe or very severe bodily pain (Item 7).

^b% reporting no or very mild pain (Item 7).

^c% reporting that pain interferes with normal work (inside and outside the home) quite a lot or extremely (Item 8).

^d% reporting that pain interferes with normal work (inside and outside the home) a little bit or not at all (Item 8).

Figure 8.13 Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (N = 4,027)

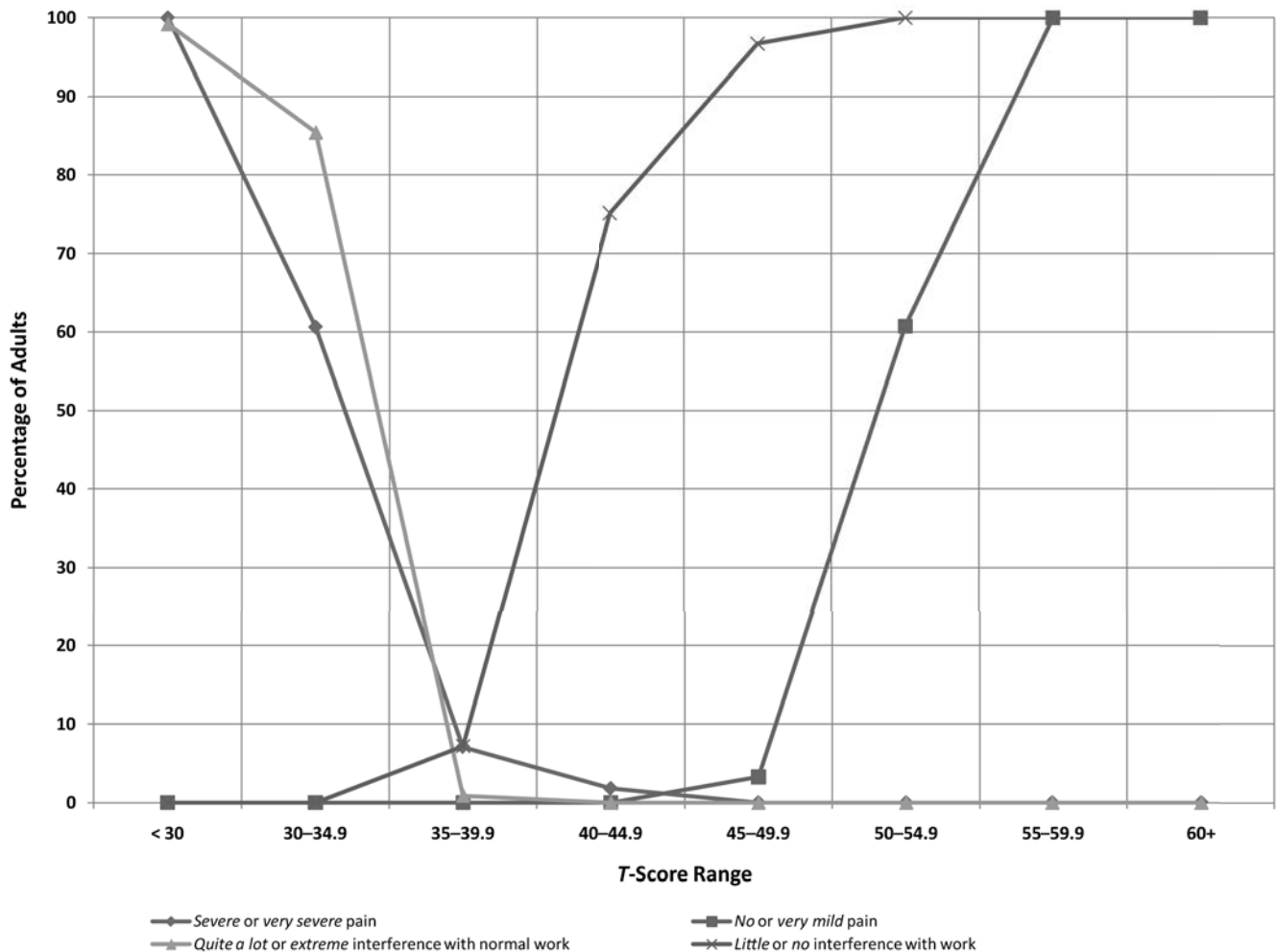


Table 8.14

Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (N = 4,036)

GH T-Score Level	T Scores		n	Fair or poor health ^a (1) %	Getting sick easier mostly or definitely true ^b (2) %	As healthy as anybody mostly or definitely false ^c (3) %	Health expected to get worse mostly or definitely true ^d (4) %	Health is excellent mostly or definitely false ^e (5) %
	Range	Mean						
1	65+	65.94	258	0.0	0.0	0.0	0.0	0.0
2	60–64.9	61.63	537	0.0	0.0	0.0	0.0	0.0
3	55–59.9	56.83	827	0.1	0.7	1.1	4.5	0.4
4	50–54.9	52.05	745	1.1	3.5	4.3	14.8	5.7
5	45–49.9	47.34	547	7.7	6.5	13.0	22.9	37.0
6	40–44.9	42.32	457	27.8	12.3	31.9	28.7	68.1
7	35–39.9	37.61	330	48.8	24.1	64.4	40.6	91.7
8	30–34.9	32.32	184	80.4	37.4	91.8	51.9	98.4
9	< 30	25.69	151	95.4	66.4	96.6	76.0	100.0

^a% reporting fair or poor health (Item 1).

^b% reporting getting sick easier as mostly true or definitely true (Item 11a).

^c% reporting being as healthy as anybody they know as mostly false or definitely false (Item 11b).

^d% reporting health expected to get worse as mostly true or definitely true (Item 11c).

^e% reporting health is excellent as mostly false or definitely false (Item 11d).

Figure 8.14 Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (N = 4,036)

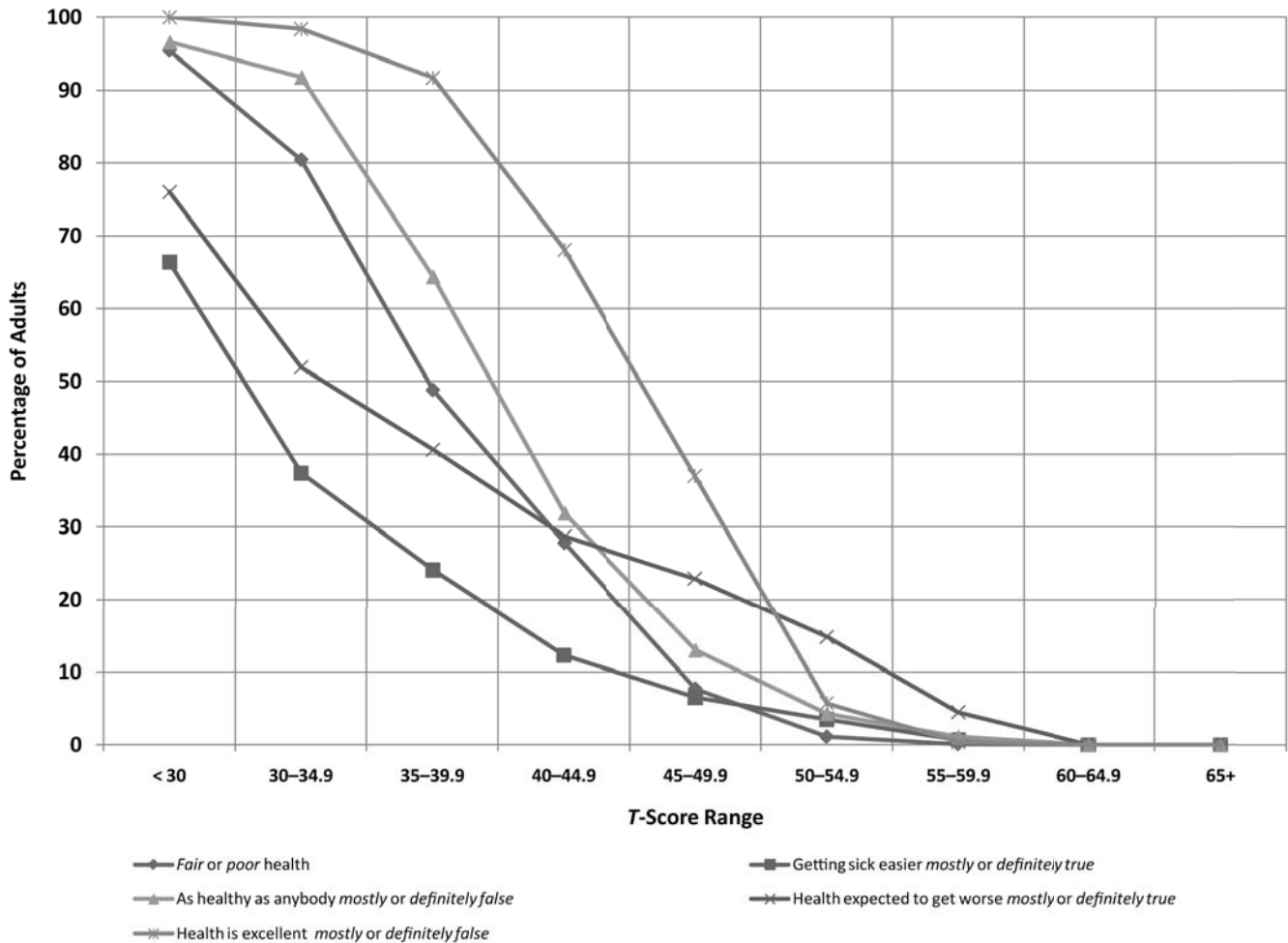


Table 8.15

Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (N = 4,028)

VT T-Score Level	T Scores		n	Feeling full of life little or none of the time ^a	Having a lot of energy little or none of the time ^b	Feeling worn out most or all of the time ^c	Feeling tired most or all of the time ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	65+	68.84	167	0.0	0.0	0.0	0.0
2	60–64.9	62.56	443	0.0	0.0	0.0	0.0
3	55–59.9	57.25	995	0.6	1.0	0.0	0.7
4	50–54.9	52.59	464	3.5	3.5	0.4	2.0
5	45–49.9	48.13	860	11.4	16.9	4.7	11.4
6	40–44.9	42.34	491	34.4	62.1	24.2	48.5
7	35–39.9	37.75	173	47.7	83.2	58.4	83.1
8	30–34.9	33.39	263	82.8	93.1	82.5	95.8
9	< 30	26.08	172	98.3	100.0	99.4	99.4

^a% reporting feeling full of life *little or none of the time* (Item 9a).

^b% reporting having a lot of energy *little or none of the time* (Item 9e).

^c% reporting feeling worn out *most or all of the time* (Item 9g).

^d% reporting feeling tired *most or all of the time* (Item 9i).

Figure 8.15 Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (N = 4,028)

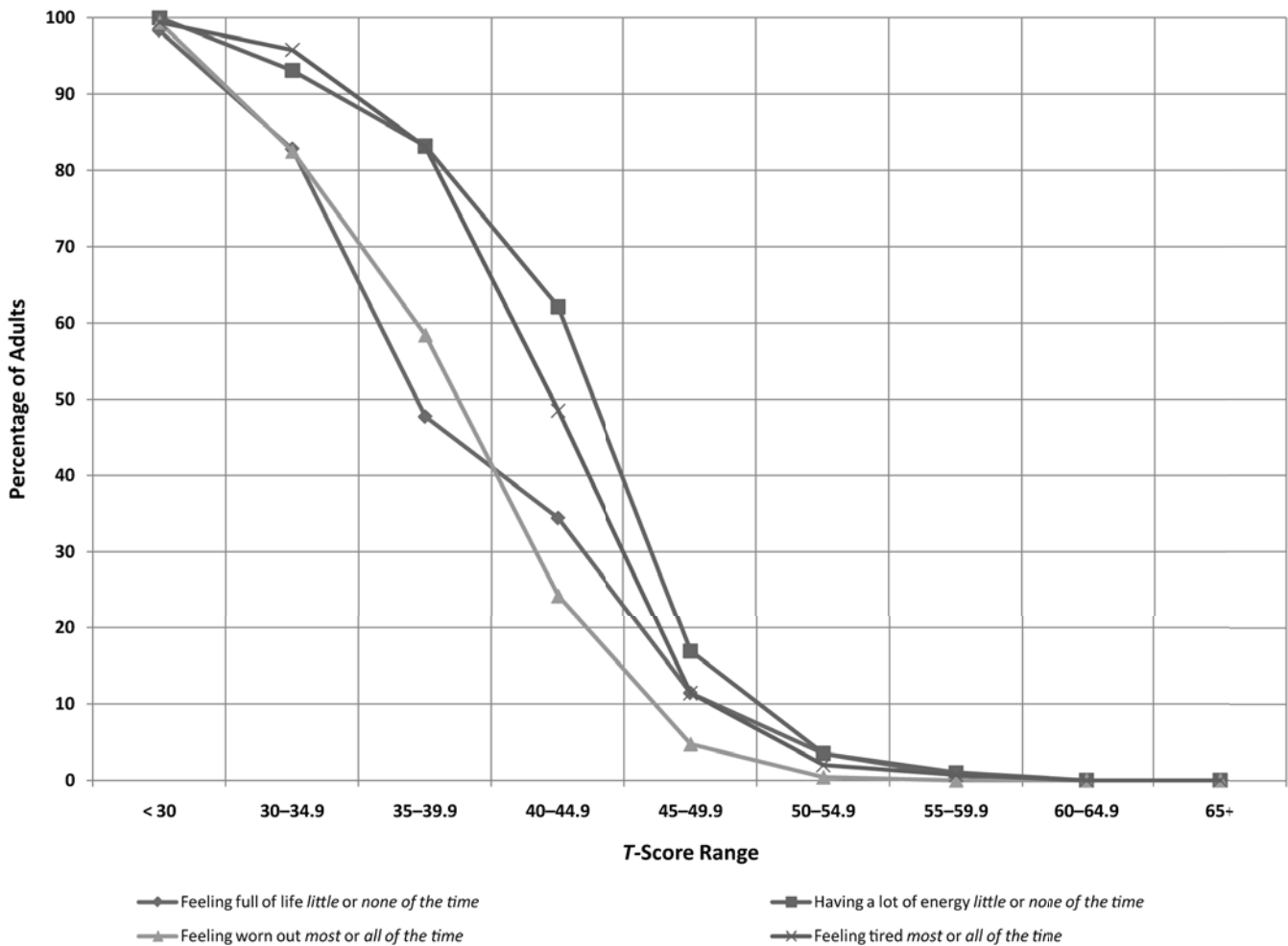


Table 8.16

Percentage of Adults Reporting Limitations in Social Activities at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (N = 4,029)

SF T-Score Level	T Scores		n	Health interfered with social activities moderately, quite a bit, or extremely ^a	Health interfered with social activities slightly or not at all ^b	Health interfered with social activities most or all of the time ^c	Health interfered with social activities little or none of the time ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	55+	57.34	2,187	0.0	100.0	0.0	100.0
2	50–54.9	52.33	435	0.0	100.0	0.0	100.0
3	45–49.9	47.31	484	6.6	93.4	0.0	90.4
4	40–44.9	42.30	269	33.1	66.9	4.1	33.1
5	35–39.9	37.29	300	78.9	21.1	21.0	11.0
6	30–34.9	32.27	125	95.2	4.8	46.4	4.8
7	25–29.9	27.26	118	100.0	0.0	95.7	0.0
8	< 25	20.04	111	100.0	0.0	100.0	0.0

^a% reporting physical or emotional problems interfering with social activities moderately, quite a bit, or extremely (Item 6).

^b% reporting physical or emotional problems interfering with social activities slightly or not at all (Item 6).

^c% reporting physical or emotional problems interfering with social activities most or all of the time (Item 10).

^d% reporting physical or emotional problems interfering with social activities little or none of the time (Item 10).

Figure 8.16 Percentage of Adults Reporting Limitations in Social Activities at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (N = 4,029)

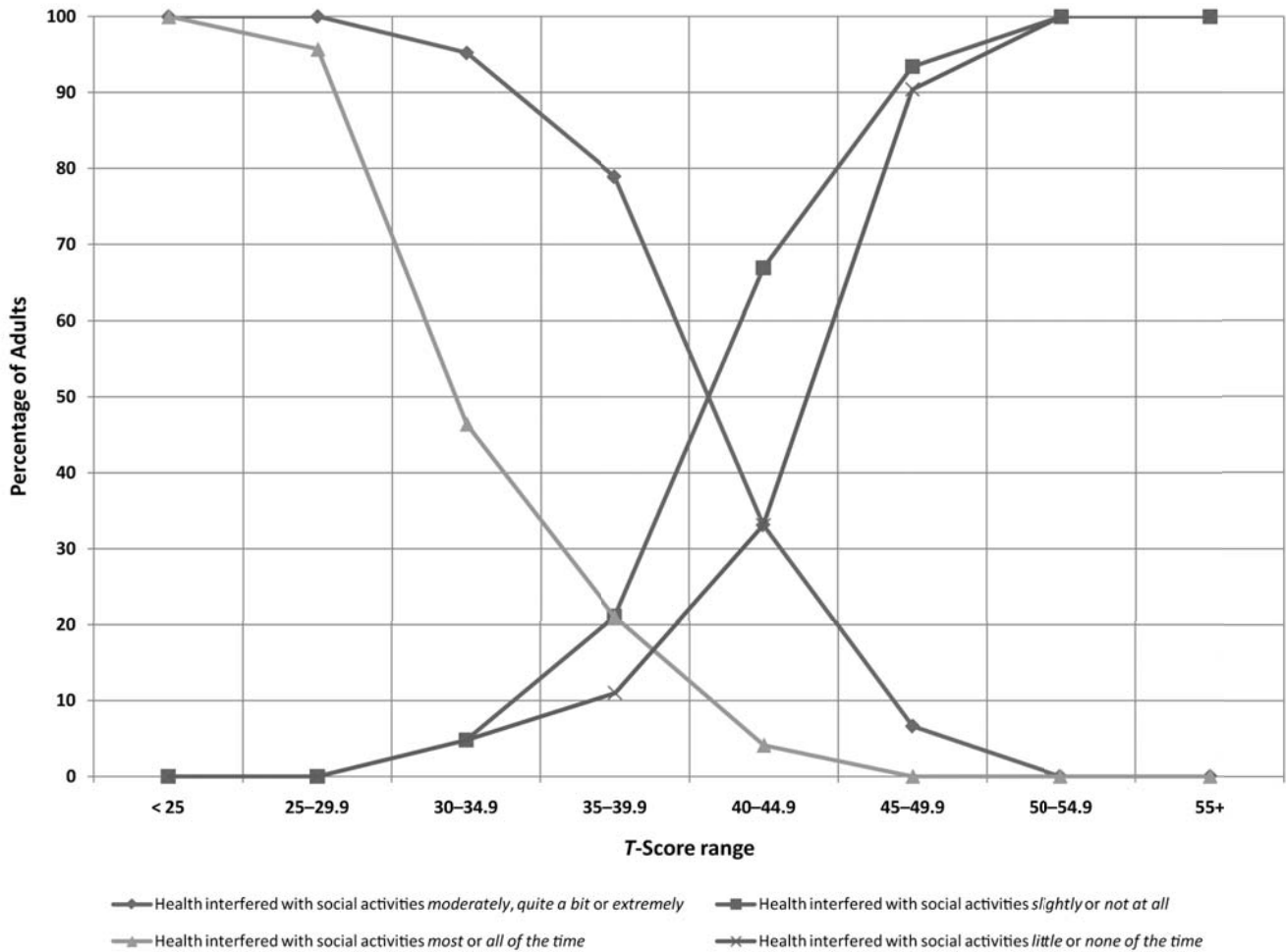


Table 8.17

Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (N = 4,026)

RE T-Score Level	T Scores		n	Cut down time at work most or all of the time ^a	Accomplished less most or all of the time ^b	Did work less carefully most or all of the time ^c
	Range	Mean		(1) %	(2) %	(3) %
1	55+	56.17	2,461	0.0	0.0	0.0
2	50–54.9	52.66	246	0.0	0.0	0.0
3	45–49.9	47.41	592	0.2	0.7	0.2
4	40–44.9	42.24	134	0.8	4.5	0.8
5	35–39.9	36.40	311	4.9	11.9	1.9
6	30–34.9	31.78	60	42.4	68.3	20.0
7	25–29.9	28.31	55	69.1	90.9	32.7
8	20–24.9	23.88	97	97.9	99.0	72.3
9	< 20	14.96	70	100.0	100.0	100.0

^a% reporting cutting down amount of time spent on work or other activities *most* or *all of the time* (Item 5a).

^b% reporting accomplished less than they would like *most* or *all of the time* (Item 5b).

^c% reporting did work or other activities less carefully *most* or *all of the time* (Item 5c).

Figure 8.17 Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (N = 4,026)

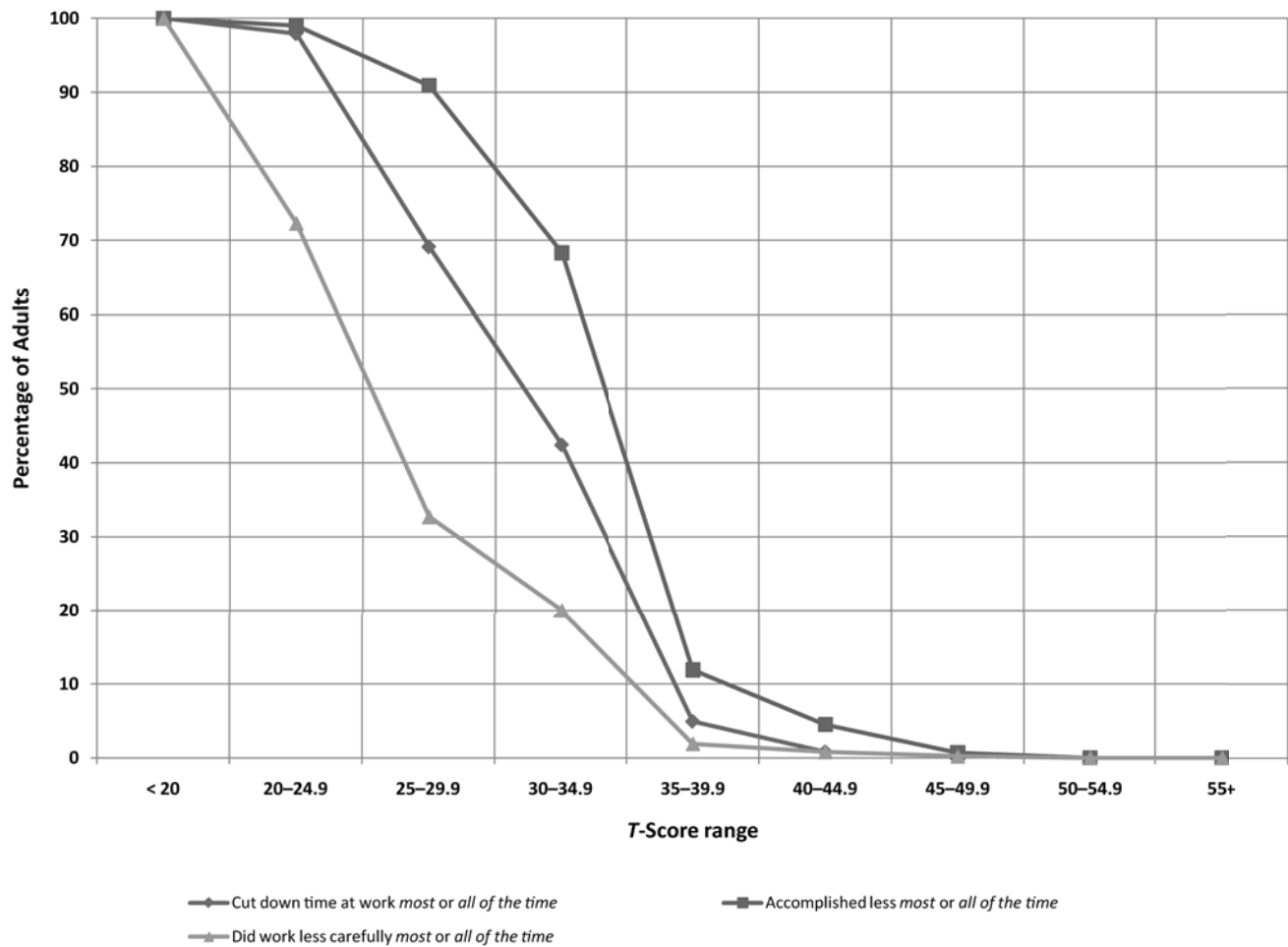


Table 8.18

Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (N = 4,028)

MH T-Score Level	T Scores		n	Been very nervous most or all of the time ^a	Down in dumps most or all of the time ^b	Calm little or none of the time ^c	Downhearted and depressed most or all of the time ^d	Happy little or none of the time ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	60+	62.52	446	0.0	0.0	0.0	0.0	0.0
2	55-59.9	57.54	1,320	0.0	0.0	0.1	0.0	0.0
3	50-54.9	52.35	876	0.1	0.4	6.2	0.0	2.1
4	45-49.9	47.13	496	2.0	0.6	19.0	0.2	9.5
5	40-44.9	41.90	336	9.6	1.8	43.3	1.5	25.1
6	35-39.9	36.87	273	18.1	5.2	56.3	12.6	37.7
7	30-34.9	32.52	85	37.7	22.6	79.8	44.6	49.4
8	25-29.9	28.72	100	48.0	47.0	85.0	72.0	81.8
9	< 25	21.06	96	81.1	88.5	97.9	95.7	95.8

^a% reporting being very nervous most or all of the time (Item 9b).

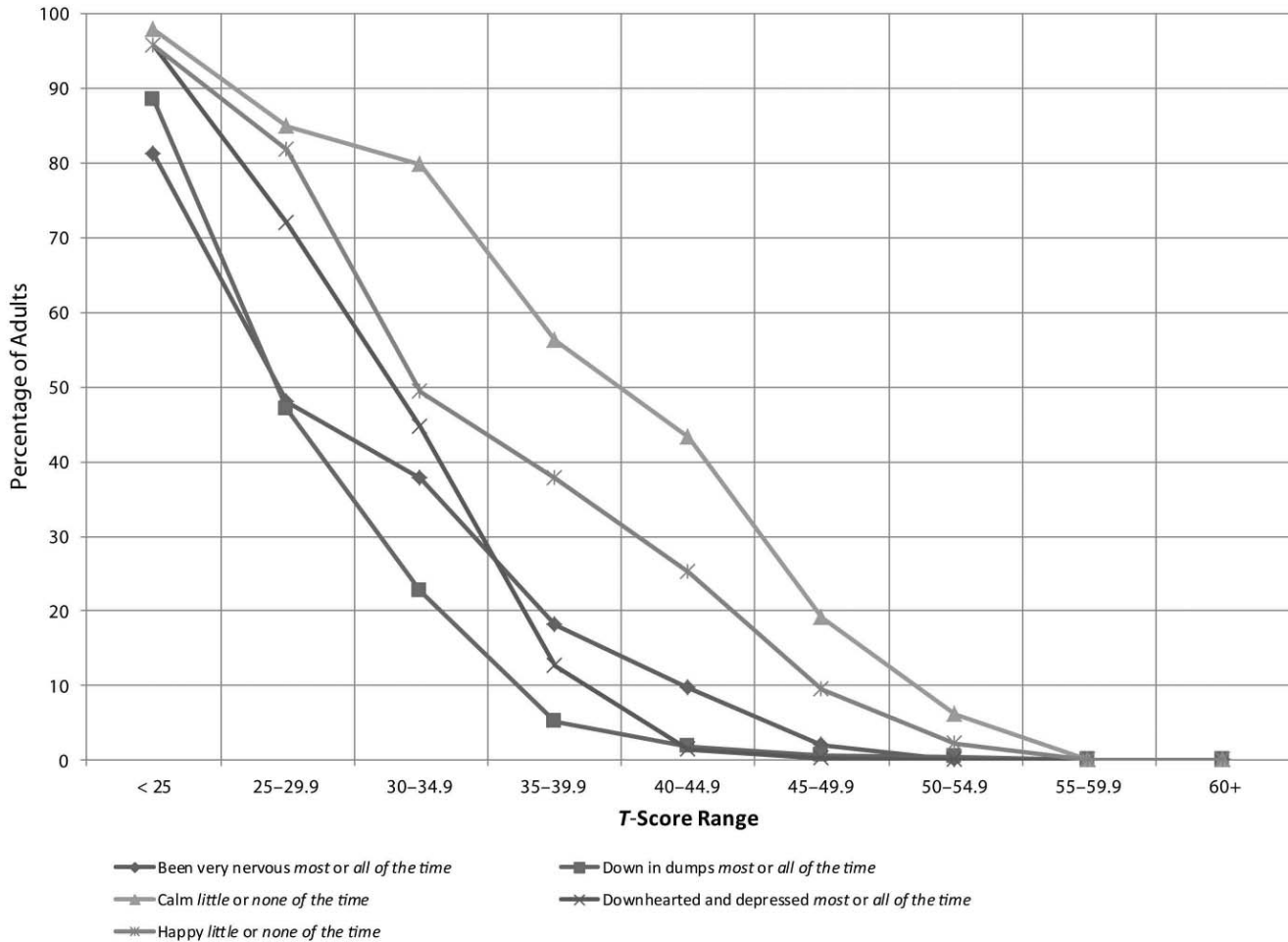
^b% reporting being so down in the dumps that nothing could cheer them up most or all of the time (Item 9c).

^c% reporting being calm and peaceful little or none of the time (Item 9d).

^d% reporting being downhearted and depressed most or all of the time (Item 9f).

^e% reporting being happy little or none of the time (Item 9h).

Figure 8.18 Percentage of Adults Reporting Emotional Distress at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (N = 4,028)



lower PCS score levels was seen for bending/kneeling/stooping (Column 6) and walking more than a mile (Column 7), making these two items useful in interpreting all PCS score ranges as well. It was notable that even at the upper half of the average PCS score range (Level 3), one third and one quarter of the respondents, respectively, reported having limitations in these two activities.

Figures 8.19 and 8.20 present graphs of the percentage of the sample scoring at each PCS score level that reported each limitation defined or evaluated in Tables 8.19 and 8.20, respectively.

Role functioning and PCS. With one minor exception (Column 1, Level 2), Table 8.21 reveals that decreasing levels of PCS scores are associated with a linear increase in the percentage of respondents who reported physical health-related problems in carrying out the role-related activities measured by the RP scale items. In general, significant percentages of respondents reporting problems did not occur until Level 5 (*T*-score range = 40.0–44.9), the first level that is considered to be below the average score range. Reports of cutting down time at work (Column 1), accomplishing less (Column 2), being limited in the kind of work or activities performed (Column 3), and having difficulty in doing work or other activities (Column 4) *most or all of the time* slowly increased through the highest and middle score ranges and then quickly increased through the lowest score levels (Levels 6–8); however, even the largest percentages of those reporting such limitations fell below 90%. Overall, all four RE items appear to be most useful in interpreting the middle- and low-range PCS score levels.

Figure 8.21 presents a graph of the percentage of the sample scoring at each PCS score level that reported each limitation or characteristic defined or evaluated in Table 8.21.

Bodily pain and PCS. As shown in Table 8.22, both of the BP items are considered useful in interpreting the highest and middle ranges of the acute form PCS levels, with increasing percentages of respondents reporting *severe* or *very severe* pain (Column 1) or pain interfering with work *quite a lot* or *extremely* (Column 3) from the highest to the lowest PCS score levels. The opposite is true for those reporting *no* or *very mild* pain (Column 2) or *little* or *no interference* of pain with work (Column 4), making these two variables useful in interpreting PCS score differences across all score levels.

Figure 8.22 presents a graph of the percentage of the sample scoring at each PCS score level that reported each limitation or characteristic defined or evaluated in Table 8.22.

General health and PCS. Table 8.23 reveals perceptions of general health-related problems throughout

all the PCS score levels, with interesting differences being noted. Reports of *fair* or *poor* health (Column 1) and reports of being as healthy as anybody they know (Column 3) and health being excellent (Column 5) as *mostly* or *definitely false* increased in a linear manner as PCS scores decreased. Overall, these items are generally useful in interpreting differences in PCS scores across the middle and lowest score levels. The percentages of those reporting getting sick easier (Column 2) and expecting health to get worse (Column 4) as *mostly* or *definitely true* generally increased with decreasing PCS scores in a linear manner. Also, the percentages of those reporting these two problems at the lowest score levels were relatively low: 21.1% and 31.1% for getting sick easier and 38.9% and 53.8% for health expected to get worse (Levels 7 and 8, respectively). Overall, these two variables are useful in interpreting PCS score differences across all score levels.

Figure 8.23 presents a graph of the percentage of the sample scoring at each PCS score level that reported each limitation or characteristic defined or evaluated in Table 8.23.

Mental Component Summary (MCS)

Tables 8.24 through 8.27 provide data for the content-based interpretations of MCS *T* scores relative to reported limitations in vitality, social and role functioning, and mental health.

Vitality and MCS. Table 8.24 reveals a relationship between decreasing MCS scores and increasing reports of feeling worn out (Column 3) and tired (Column 4) *most or all of the time*. A similar, general trend was also found with regard to feeling full of life (Column 1) and having a lot of energy (Column 2) *little* or *none of the time*. It's notable that reports of these problems were present at the highest of the MCS score levels. Also, a substantial portion of those scoring at Level 3, which includes the mean MCS *T* score of 50, reported problems affecting their vitality. Generally, all of the VT items are useful in interpreting the entire range of MCS levels.

Figure 8.24 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.24.

Social functioning and MCS. Table 8.25 reveals a general trend for increasing percentages of respondents reporting health as interfering with social activities *moderately*, *quite a bit*, or *extremely* (Column 1) and *most* or *all of the time* (Column 3) as MCS score levels go from high to low. Conversely, the percentages reporting that health interfered with social activities *slightly* or *not at all* (Column 2) and *little* or *none of the time* (Column

4) decreased with the lowering of the MCS score levels. Generally, the two SF items are most useful in interpreting the middle and lower MCS score levels.

Figure 8.25 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.25.

Role functioning and MCS. As shown in Table 8.26, the three RE items proved to be the most useful in interpreting MCS score at the lowest levels. Reports of problems in accomplishing less (Column 2) and working less carefully (Column 3) were not present in any of the three highest score levels. As for cutting down time spent on work or other activities (Column 1), a significant percentage of respondents (13.2%) reporting problems in this area did not appear until MCS score Level 6 (*T*-score range = 35.0–39.9). Beginning at score Level 7, the percentage of those reporting any of the RE scale problems quickly accelerated.

Figure 8.26 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.26.

Mental health and MCS. As would be expected, responses to each of the five MH scale items had a fairly clear relationship with the MCS measure, with progressively more respondents reporting emotional problems generally from the middle score levels to the lower levels (see Table 8.27). It is notable that there were no reports of either feeling down in the dumps (Column 2) or downhearted and depressed (Column 4) *most or all of the time* until MCS scores fell into the score levels that are below the average range (Levels 5–8). The fact that reports of experiencing either of these two indicators of depression did not appear until MCS *T* scores were below 45 is consistent with Ware and Kosinski's (2001b) recommendation for using an MCS *T*-score cutoff of 42 as a first-stage screen for depression. Beginning slowly and then accelerating, the percentage of those reporting each of these problems then increased through the middle and lower MCS score ranges, demonstrating their usefulness in interpreting MCS scores at the middle and lower score levels. It is interesting that at the lowest MCS score level, only 68.4% reported feeling down in the dumps and 86.0% reported feeling downhearted and depressed. Overall, these two variables are most useful in interpreting MCS score differences at the lowest *T*-score levels.

Furthermore, the percentages of those reporting feeling calm and peaceful (Column 3) and happy (Column 5) *little or none of the time* steadily increased from the highest to the lowest MCS score levels in a linear fash-

ion. These items, therefore, can be considered useful in interpreting MCS scores throughout all the MCS score levels. Although the nervousness item (Column 1) appears useful in interpreting MCS scores at the middle and lower levels, the relatively low percentages of those reporting this symptom *most or all of the time* at Level 7 (29.6%) and Level 8 (64.3%) suggest that MCS scores may be more sensitive to the presence of symptoms of depression than symptom of nervousness.

Figure 8.27 presents a graph of the percentage of the sample scoring at each MCS score level that reported each limitation or characteristic defined or evaluated in Table 8.27.

Content-Based Interpretation of the Acute Form Health Domain Scales

Content-based interpretations of the SF-36v2 acute form health domain scales are facilitated through an examination of the percentage of respondents from the 2009 U.S. general population normative sample whose responses to each item from each health domain scale were indicative of problems or limitations imposed by the respondents' health status, at each score level of a given health domain scale.

Physical Functioning (PF)

Tables 8.28 and 8.29 present the percentages of respondents reporting limitations at the seven PF scale score levels. As revealed in Table 8.28, limitations in vigorous activities (Column 1) were found in significant percentages throughout all PF score ranges. Even at the highest score level (Level 1, *T*-score range = 55+), almost one quarter (23.4%) of the respondents reported some limitations in these types of activities. The percentage jumps to 90.7% at Level 2, which represents the upper half of the average score range. Overall, limitations in vigorous activities are useful in interpreting PF score differences at only the highest score levels. Meanwhile, the percentages of those reporting limitations in moderate activities (Column 2), lifting or carrying groceries (Column 3), climbing several flights of stairs (Column 4), and climbing one flight of stairs (Column 5) all increase linearly and rapidly through the PF score levels, with all items increasing to 98.5% or higher by score Level 7 and each proving most useful in interpreting PF score differences across all score levels. Along with the prevalence of limitations in vigorous activities, notable is the fact that by score Level 2 (which includes the average PF *T* score), 40.6% of the respondents reported limitations in climbing several flights of stairs (Column 4).

Table 8.19

Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

PCS T-Score Level	T Scores		n	Limited in vigorous activities ^a	Limited in moderate activities ^b	Limited in lifting or carrying groceries ^c	Limited in climbing several flights of stairs ^d	Limited in climbing one flight of stairs ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	60+	61.80	155	14.2	1.3	0.0	4.5	1.9
2	55-59.9	57.55	686	26.7	1.9	0.9	6.2	0.4
3	50-54.9	52.73	452	72.5	13.1	6.0	31.0	3.1
4	45-49.9	47.77	229	90.4	36.7	18.9	62.6	18.9
5	40-44.9	42.58	196	99.0	72.5	45.6	81.5	38.8
6	35-39.9	37.55	116	99.1	81.0	56.9	88.7	69.0
7	30-34.9	32.65	90	97.8	96.7	77.8	95.5	77.8
8	< 30	24.39	132	100.0	99.2	96.2	99.2	93.9

^a% reporting any limitations in vigorous activities (Item 3a).

^b% reporting any limitations in moderate activities (Item 3b).

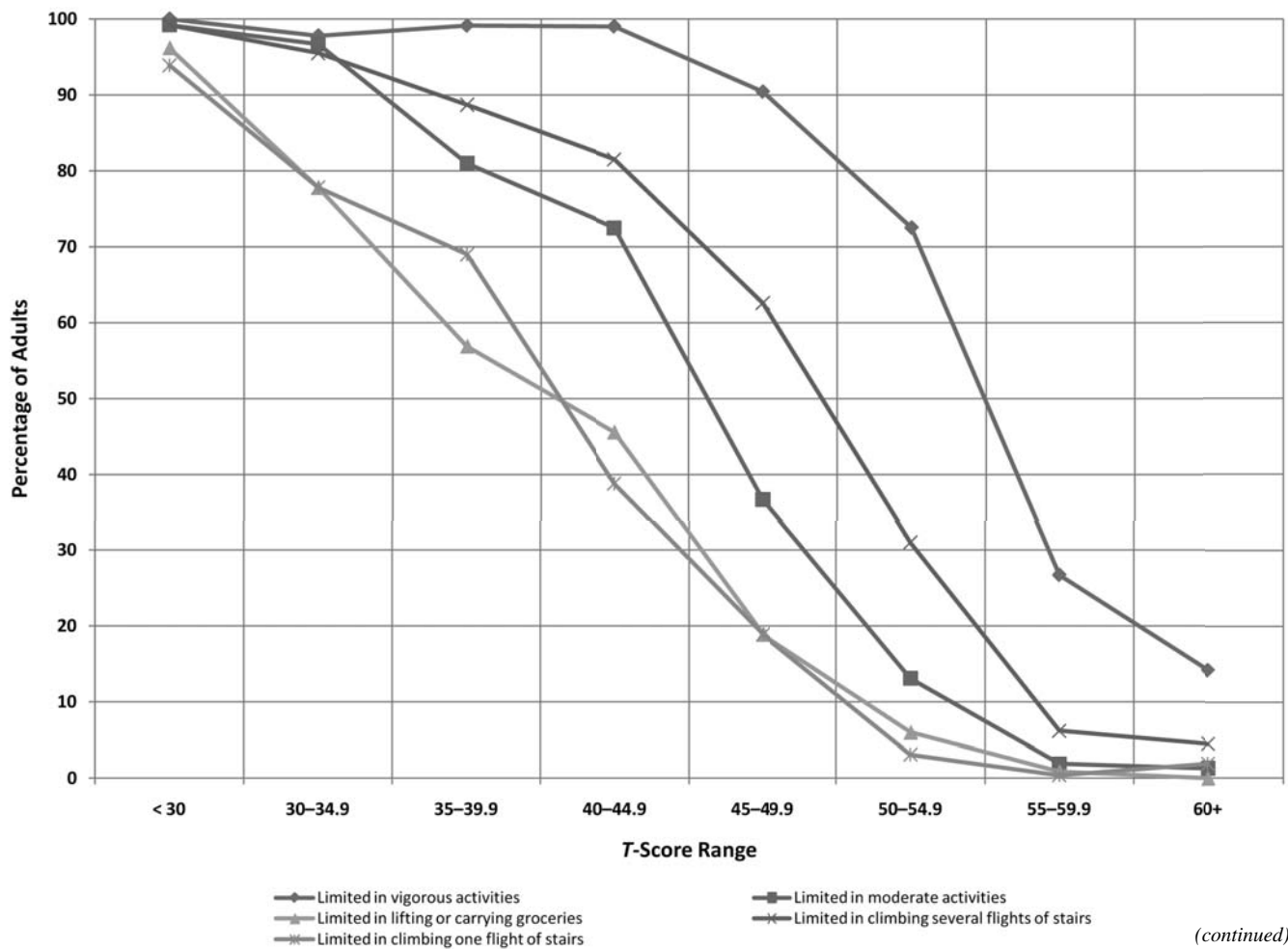
^c% reporting any limitations in lifting or carrying groceries (Item 3c).

^d% reporting any limitations in climbing several flights of stairs (Item 3d).

^e% reporting any limitations in climbing one flight of stairs (Item 3e).

(continued)

Figure 8.19 Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)



(continued)

Table 8.20

Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056) (continued)

PCS T-Score Level	T Scores		n	Limited in bending, kneeling, or stooping ^f	Limited in walking more than a mile ^g	Limited in walking several hundred yards ^h	Limited in walking 100 yards ⁱ	Limited in bathing yourself ^j
	Range	Mean		(6) %	(7) %	(8) %	(9) %	(10) %
1	60+	61.80	155	3.9	5.8	1.3	0.7	0.0
2	55-59.9	57.55	686	8.2	4.1	1.2	0.9	0.2
3	50-54.9	52.73	452	33.2	25.8	6.0	4.0	0.7
4	45-49.9	47.77	229	64.5	52.8	16.2	13.2	3.1
5	40-44.9	42.58	196	75.0	77.8	41.8	28.9	12.3
6	35-39.9	37.55	116	87.1	88.8	69.0	54.4	18.1
7	30-34.9	32.65	90	89.8	95.5	83.0	73.0	25.6
8	< 30	24.39	132	97.7	99.2	95.4	90.1	61.4

^f% reporting any limitations in bending, kneeling, or stooping (Item 3f).

^g% reporting any limitations in walking more than a mile (Item 3g).

^h% reporting any limitations in walking several hundred yards (Item 3h).

ⁱ% reporting any limitations in walking 100 yards (Item 3i).

^j% reporting any limitations in bathing or dressing oneself (Item 3j).

Figure 8.20 Percentage of Adults Reporting Limitations in Physical Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056) (continued)

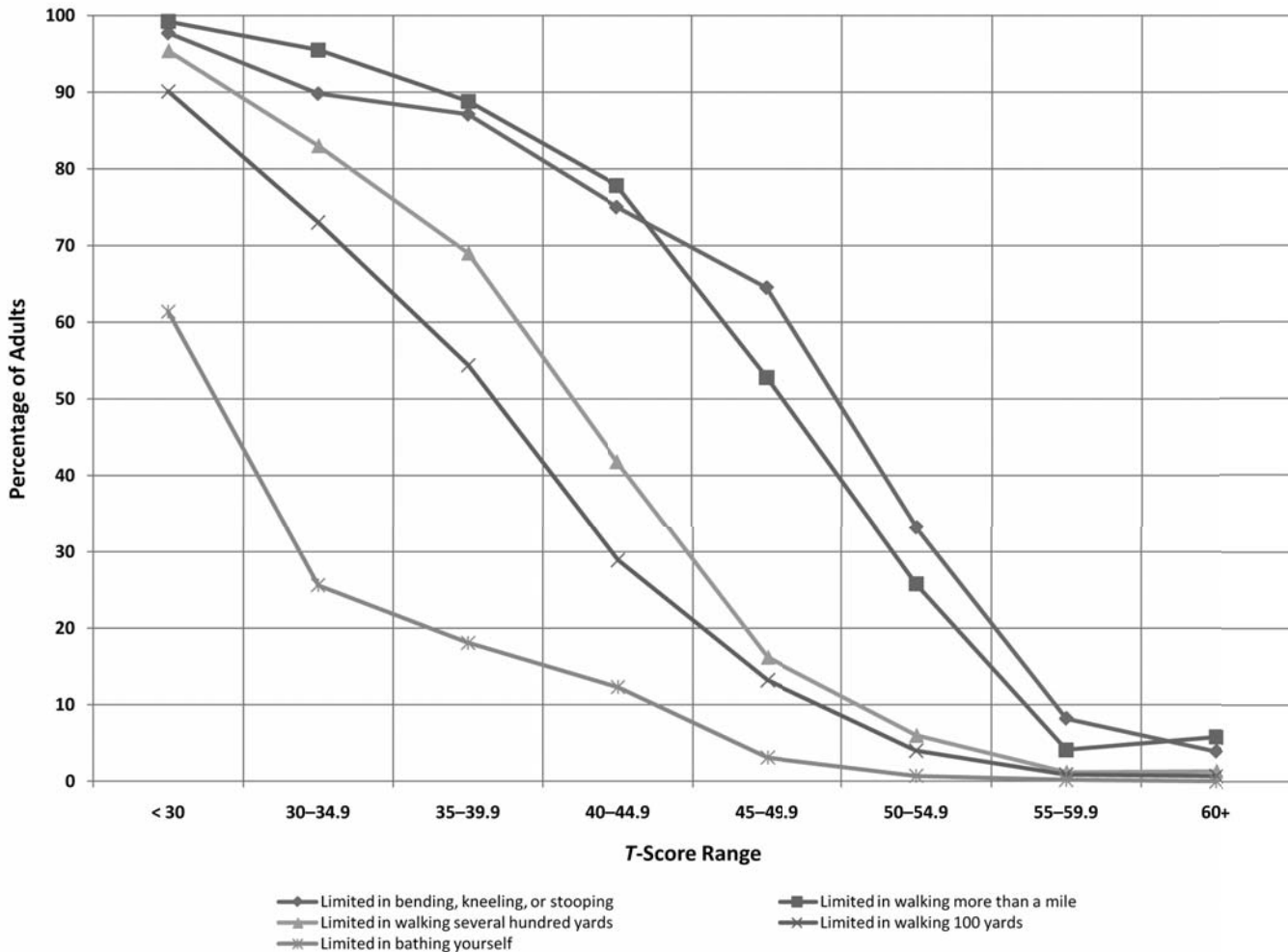


Table 8.21

Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

PCS T-Score Level	T Scores		n	Cut down time at work most or all of the time ^a	Accomplished less most or all of the time ^b	Limited in the kind of work or other activities most or all of the time ^c	Difficulty at work most or all of the time ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	61.80	155	0.0	0.0	0.0	0.0
2	55-59.9	57.55	686	0.4	0.2	0.0	0.0
3	50-54.9	52.73	452	0.2	1.3	0.0	0.9
4	45-49.9	47.77	229	2.6	5.3	3.5	4.4
5	40-44.9	42.58	196	7.7	11.3	13.9	12.2
6	35-39.9	37.55	116	17.2	24.1	23.3	29.6
7	30-34.9	32.65	90	47.8	62.2	67.8	66.7
8	< 30	24.39	132	65.2	85.6	87.8	89.3

^a% reporting having cut down amount of time spent on work or other activities most or all of the time (Item 4a).

^b% reporting having accomplished less than they would like most or all of the time (Item 4b).

^c% reporting being limited in the kind of work or other activities most or all of the time (Item 4c).

^d% reporting having had difficulty performing work or other activities most or all of the time (Item 4d).

Figure 8.21 Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

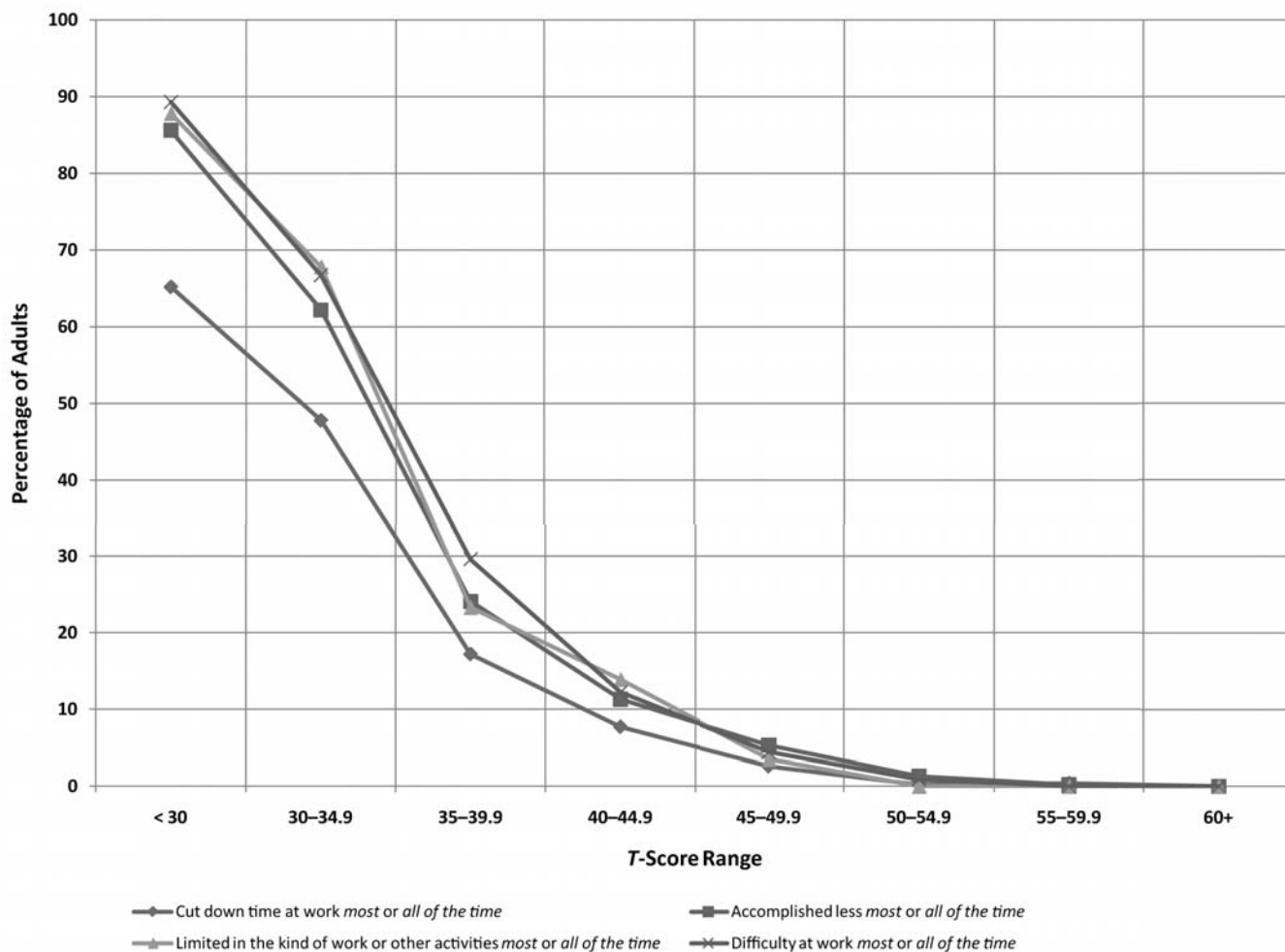


Table 8.22

Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

PCS T-Score Level	T Scores		n	Severe or very severe pain ^a	No or very mild pain ^b	Quite a lot or extreme interference with normal work ^c	Little or no interference with work ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	61.80	155	0.0	96.8	0.0	99.4
2	55–59.9	57.55	686	0.0	91.1	0.0	99.9
3	50–54.9	52.73	452	0.5	62.6	0.5	98.0
4	45–49.9	47.77	229	3.5	36.6	3.9	84.7
5	40–44.9	42.58	196	9.2	22.6	9.3	61.3
6	35–39.9	37.55	116	14.7	13.8	22.4	44.0
7	30–34.9	32.65	90	35.6	8.9	42.2	21.1
8	< 30	24.39	132	57.7	3.9	65.2	12.9

^a% reporting severe or very severe bodily pain (Item 7).

^b% reporting no or very mild pain (Item 7).

^c% reporting that pain interferes with normal work (inside and outside the home) quite a lot or extremely (Item 8).

^d% reporting that pain interferes with normal work (inside and outside the home) a little bit or not at all (Item 8).

Figure 8.22 Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

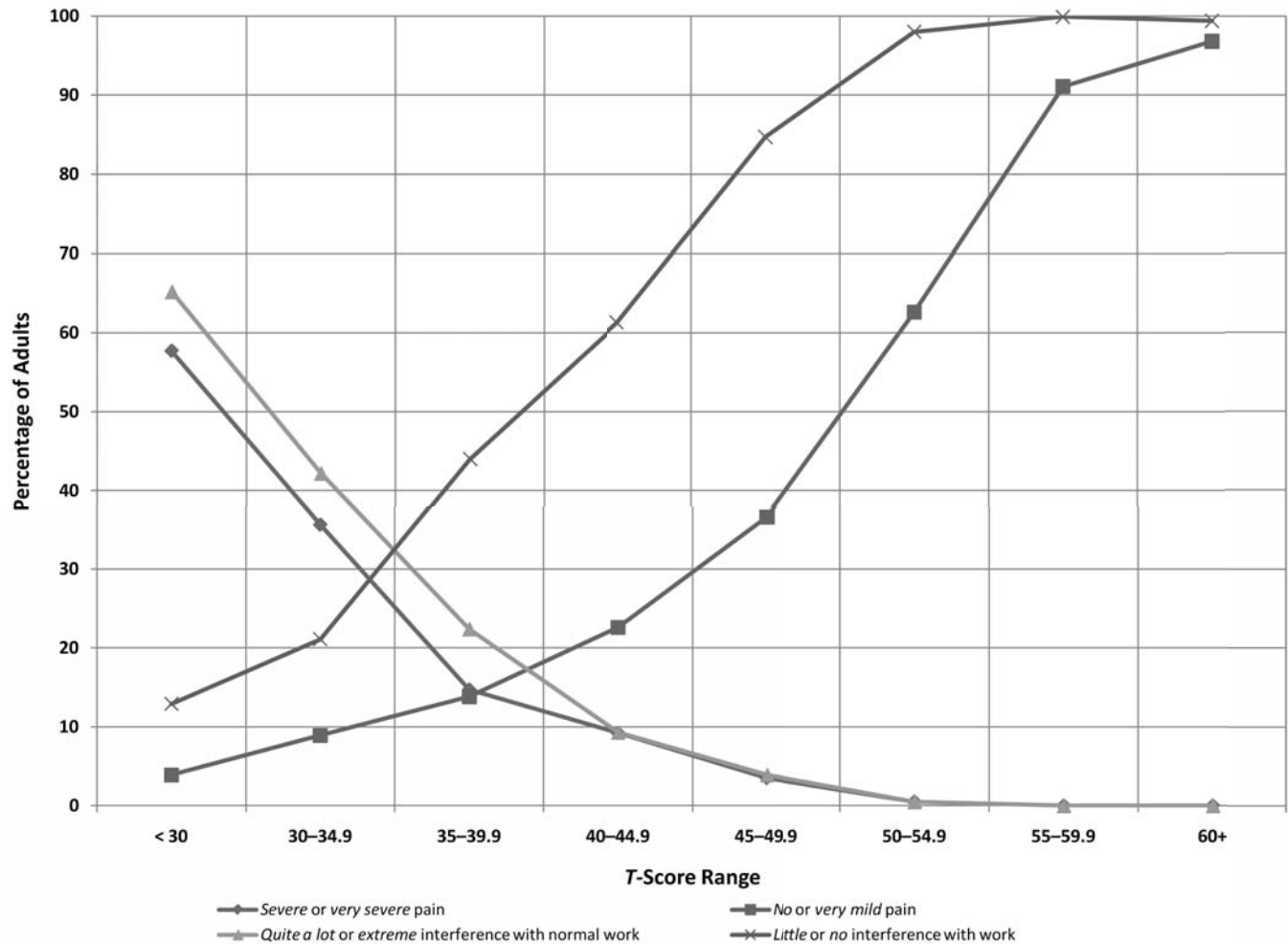


Table 8.23

Percentage of Adults Reporting General Health Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

PCS T-Score Level	T Scores		n	Fair or poor health ^a	Getting sick easier mostly or definitely true ^b	As healthy as anybody mostly or definitely false ^c	Health expected to get worse mostly or definitely true ^d	Health is excellent mostly or definitely false ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	60+	61.80	155	0.7	0.7	3.3	3.3	4.6
2	55-59.9	57.55	686	0.7	2.0	2.9	6.4	4.7
3	50-54.9	52.73	452	6.9	6.0	11.8	15.6	22.7
4	45-49.9	47.77	229	18.8	7.9	20.5	22.7	39.7
5	40-44.9	42.58	196	27.6	14.9	34.4	30.3	55.9
6	35-39.9	37.55	116	38.8	18.1	46.1	29.3	64.7
7	30-34.9	32.65	90	62.9	21.1	53.3	38.9	81.1
8	< 30	24.39	132	81.7	31.1	79.4	53.8	95.5

^a% reporting fair or poor health (Item 1).

^b% reporting getting sick easier as mostly true or definitely true (Item 11a).

^c% reporting being as healthy as anybody they know as mostly false or definitely false (Item 11b).

^d% reporting health expected to get worse as mostly true or definitely true (Item 11c).

^e% reporting health is excellent as mostly false or definitely false (Item 11d).

Figure 8.23 Percentage of Adults Reporting General Health Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

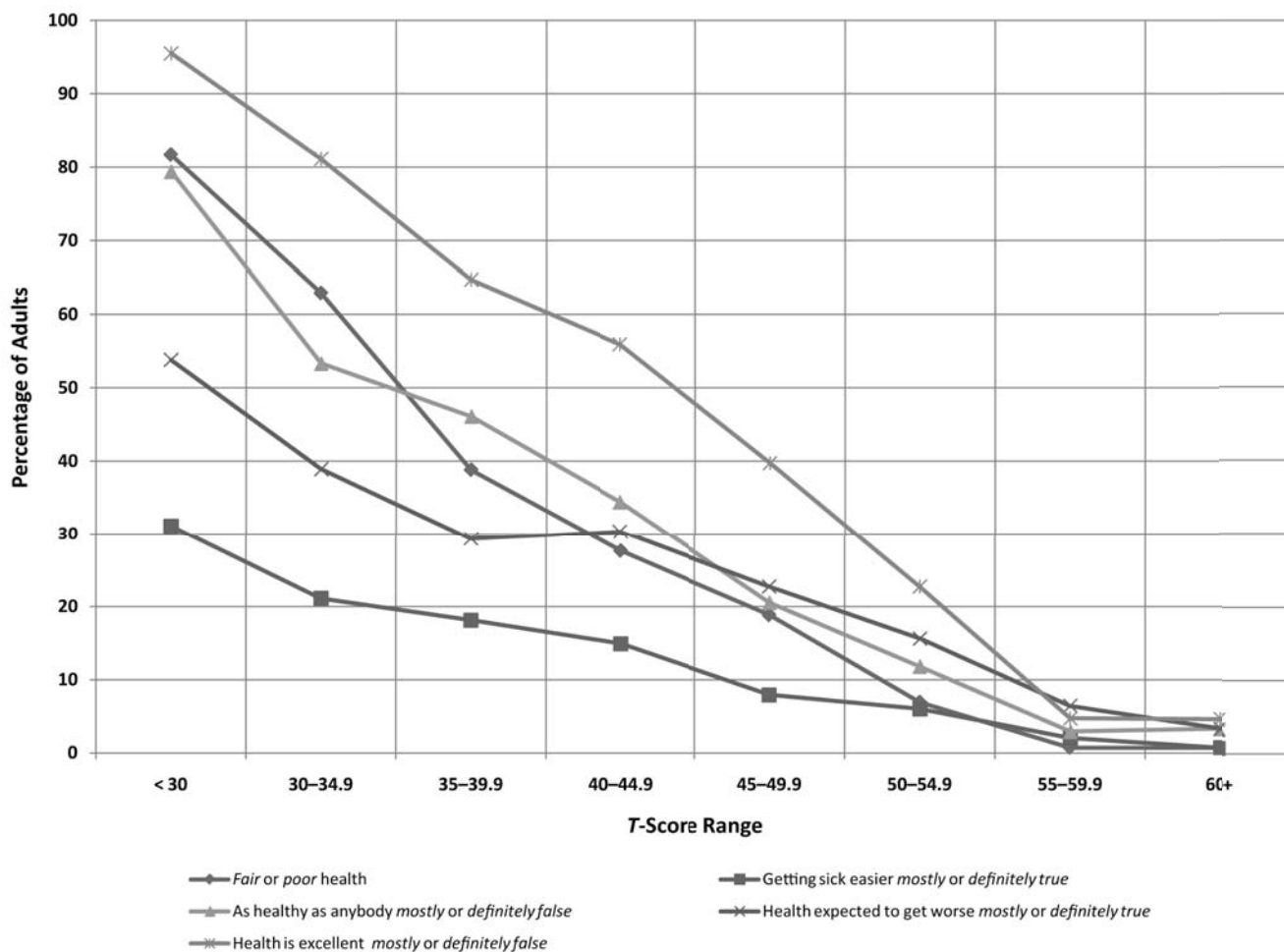


Table 8.24

Percentage of Adults Reporting Limitations in Vitality at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

MCS T-Score Level	T Scores		n	Feeling full of life little or none of the time ^a	Having a lot of energy little or none of the time ^b	Feeling worn out most or all of the time ^c	Feeling tired most or all of the time ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	61.95	239	3.4	8.8	4.2	5.0
2	55-59.9	57.21	721	5.7	6.5	2.9	3.3
3	50-54.9	52.80	419	14.6	23.2	12.2	17.3
4	45-49.9	47.69	234	31.6	44.2	20.9	33.8
5	40-44.9	42.54	179	44.1	53.1	33.7	51.1
6	35-39.9	37.64	107	58.9	67.9	37.4	51.9
7	25-34.9	31.19	100	65.0	74.5	52.5	59.6
8	< 25	19.46	57	91.2	93.0	86.0	93.0

^a% reporting feeling full of life little or none of the time (Item 9a).

^b% reporting having a lot of energy little or none of the time (Item 9e).

^c% reporting feeling worn out all or most of the time (Item 9g).

^d% reporting feeling tired most or all of the time (Item 9i).

Figure 8.24 Percentage of Adults Reporting Limitations in Vitality at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

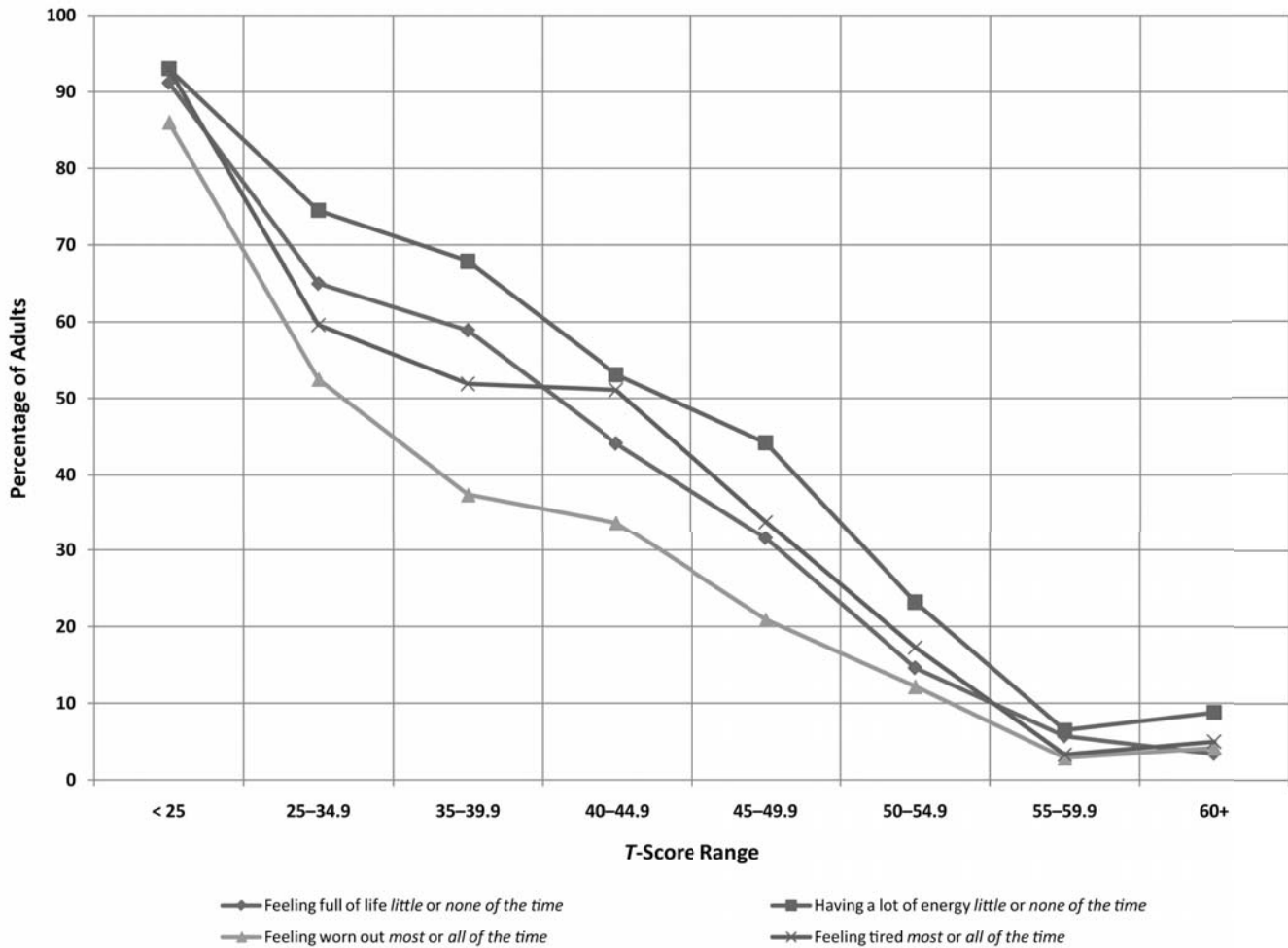


Table 8.25

Percentage of Adults Reporting Limitations in Social Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

MCS T-Score Level	T Scores		n	Health interfered with social activities <i>moderately, quite a bit, or extremely</i> ^a	Health interfered with social activities <i>slightly or not at all</i> ^b	Health interfered with social activities <i>most or all of the time</i> ^c	Health interfered with social activities <i>little or none of the time</i> ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	60+	61.95	239	3.4	96.7	1.3	97.1
2	55–59.9	57.21	721	2.1	97.9	0.4	97.6
3	50–54.9	52.80	419	8.6	91.4	2.2	90.9
4	45–49.9	47.69	234	16.3	83.7	6.0	83.3
5	40–44.9	42.54	179	30.2	69.8	16.8	59.8
6	35–39.9	37.64	107	42.5	57.6	18.7	38.3
7	25–34.9	31.19	100	69.0	31.0	32.3	21.2
8	< 25	19.46	57	93.0	7.0	87.5	1.8

^a% reporting physical or emotional problems interfering with social activities *moderately, quite a bit, or extremely* (Item 6).

^b% reporting physical or emotional problems interfering with social activities *slightly or not at all* (Item 6).

^c% reporting physical or emotional problems interfering with social activities *most or all of the time* (Item 10).

^d% reporting physical or emotional problems interfering with social activities *little or none of the time* (Item 10).

Figure 8.25 Percentage of Adults Reporting Limitations in Social Functioning at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

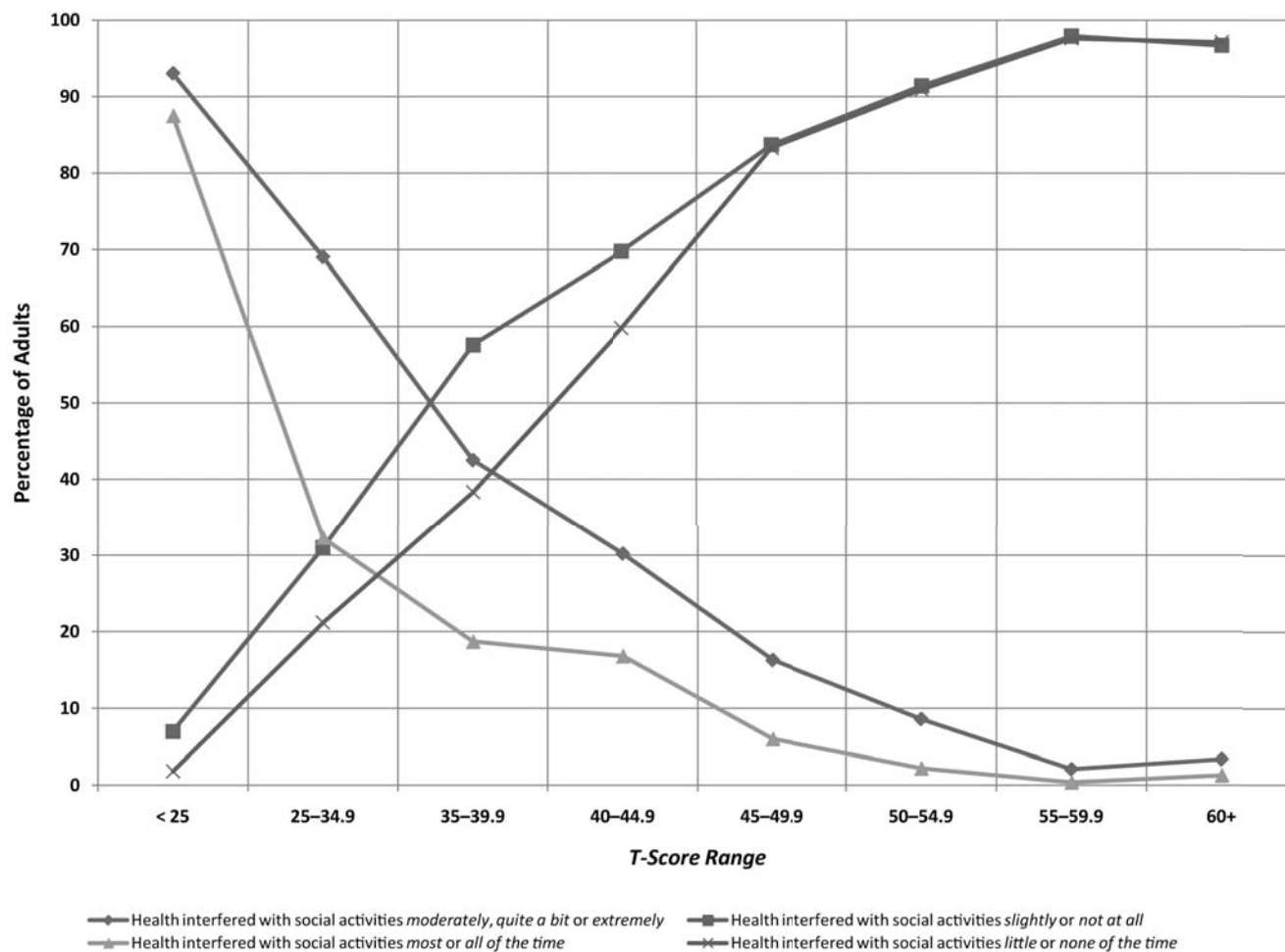


Table 8.26

Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

MCS T-Score Level	T Scores		n	Cut down time at work most or all of the time ^a	Accomplished less most or all of the time ^b	Did work less carefully most or all of the time ^c
	Range	Mean		(1) %	(2) %	(3) %
1	60+	61.95	239	0.0	0.0	0.0
2	55–59.9	57.21	721	0.4	0.0	0.0
3	50–54.9	52.80	419	0.7	0.0	0.0
4	45–49.9	47.69	234	2.2	2.6	1.3
5	40–44.9	42.54	179	2.8	1.7	0.0
6	35–39.9	37.64	107	13.2	8.4	2.8
7	25–34.9	31.19	100	31.0	32.0	19.0
8	< 25	19.46	57	78.6	89.5	77.2

^a% reporting cutting down amount of time spent on work or other activities *most or all of the time* (Item 5a).

^b% reporting accomplished less than they would like *most or all of the time* (Item 5b).

^c% reporting did work or other activities less carefully *most or all of the time* (Item 5c).

Figure 8.26 Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

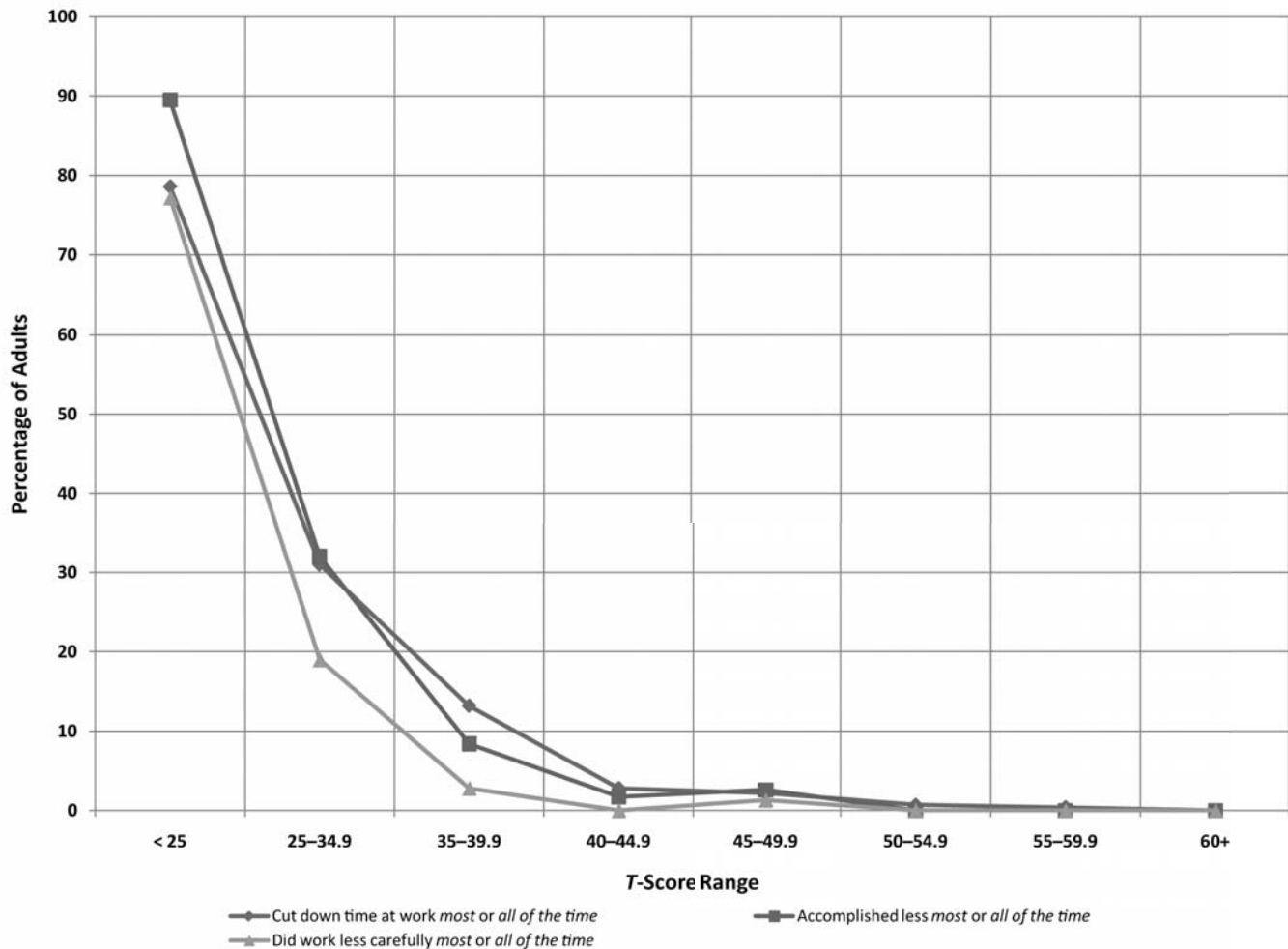


Table 8.27

Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

MCS T-Score Level	T Scores		n	Been very nervous most or all of the time ^a	Down in dumps most or all of the time ^b	Calm little or none of the time ^c	Downhearted and depressed most or all of the time ^d	Happy little or none of the time ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	60+	61.95	239	0.4	0.0	0.4	0.0	0.0
2	55–59.9	57.21	721	0.1	0.0	0.7	0.0	0.7
3	50–54.9	52.80	419	1.4	0.0	7.7	0.0	3.3
4	45–49.9	47.69	234	3.0	0.0	26.3	0.0	22.7
5	40–44.9	42.54	179	8.4	3.4	43.5	7.9	31.8
6	35–39.9	37.64	107	8.4	3.7	53.3	12.2	42.1
7	25–34.9	31.19	100	29.6	26.0	72.0	44.4	68.7
8	< 25	19.46	57	64.3	68.4	94.7	86.0	91.2

^a% reporting being very nervous most or all of the time (Item 9b).

^b% reporting being so down in the dumps that nothing could cheer them up most or all of the time (Item 9c).

^c% reporting being calm and peaceful little or none of the time (Item 9d).

^d% reporting being downhearted and depressed most or all of the time (Item 9f).

^e% reporting being happy little or none of the time (Item 9h).

Figure 8.27 Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 2,056)

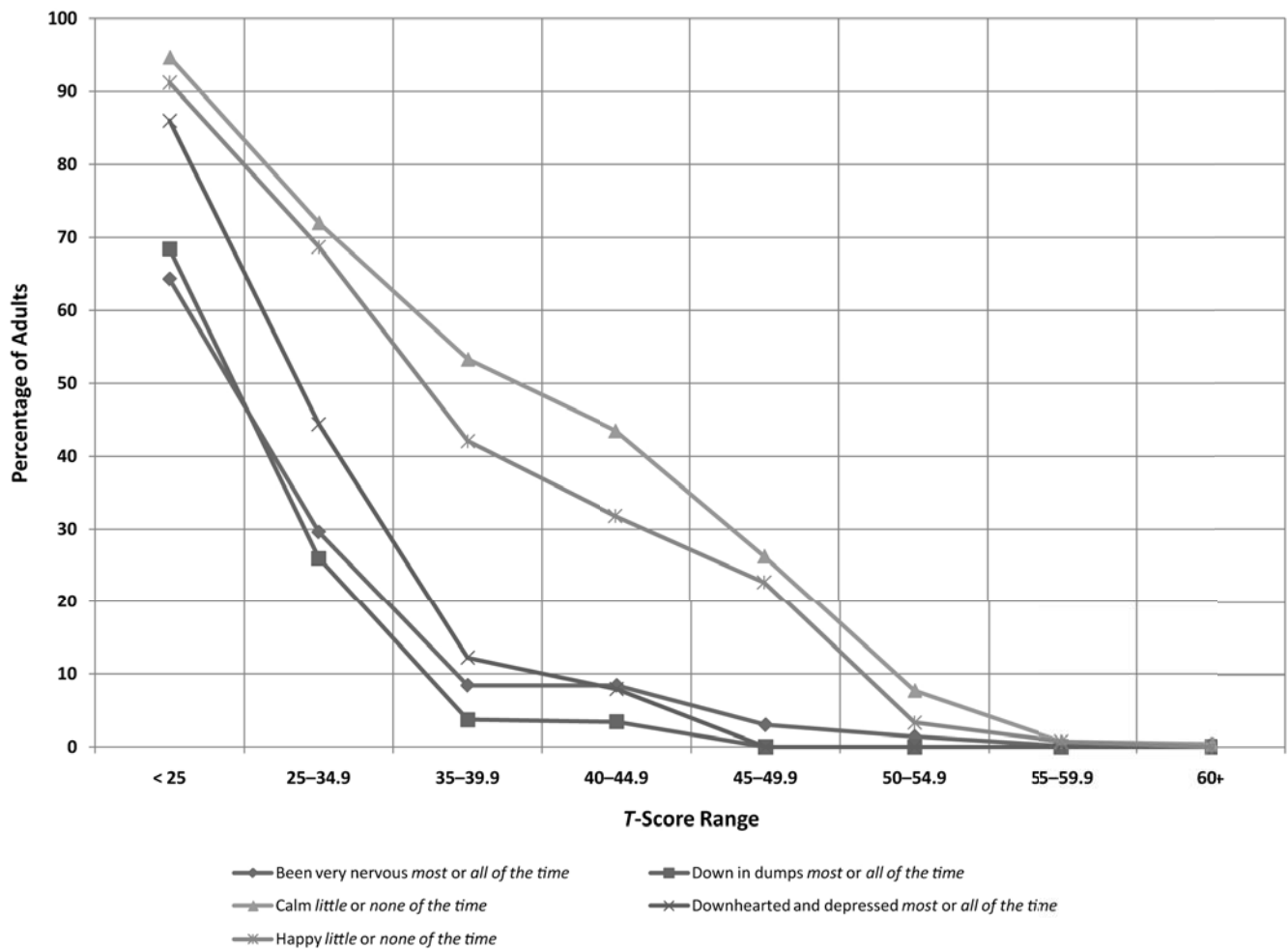


Table 8.29 reveals that a high percentage (45.3%) of respondents at PF score Level 2, which includes the average PF *T* score, reported limitations in bending, kneeling, or stooping (Column 6). Like other activities reported in Table 8.28, the percentages of respondents reporting limitations in walking more than a mile (Column 7), walking several hundred yards (Column 8), walking 100 yards (Column 9), and bathing oneself (Column 10) all increased linearly and rapidly through the PF score levels, with the percentages for three of the five variables increasing to 100% by score Level 7 (Columns 6–8). Overall, limitations in walking more than a mile is most useful in interpreting score differences in the highest and middle score ranges, while the other four items in this table are useful in interpreting PF scores differences across all score levels.

Figures 8.28 and 8.29 present graphs of the percentage of the sample scoring at each PF score level that reported each limitation or characteristic defined or evaluated in Tables 8.28 and 8.29, respectively.

Role-Physical (RP)

Table 8.30 provides data for content-based interpretation of the RP scale, with *most* or *all of the time* responses to any of this scale's four items indicating limitations in role functioning due to the respondent's physical health. It is notable that large percentages of respondents (more than 25%) reporting problems with accomplishing less (Column 2), being limited in the kinds of work or activities (Column 3), and difficulty at work (Column 4) were not seen until score Level 5 (*T*-score range = 35.0–39.9), with having to cut down on time at work *most* or *all of the time* (Column 1) increasing more than eightfold from Level 5 to Level 6 (7.8% and 65.6%, respectively). Generally, all of the items appear to be most useful in interpreting RP score differences in the middle and lowest score levels.

Figure 8.30 presents a graph of the percentage of the sample scoring at each RP score level that reported each limitation or characteristic defined or evaluated in Table 8.30.

Bodily Pain (BP)

As indicated in Table 8.31, reports of *severe* or *very severe* pain (Column 1) did not begin to occur until BP score Level 4 (1.4%, *T*-score range = 40.0–44.9), but then quickly increased to 63.8% at Level 6 and to 97.7% at Level 7. A somewhat similar pattern of increasing percentages was seen in pain interfering with normal work *quite a lot* or *extremely* (Column 3). Overall, both of these items appear most useful in interpreting BP score differences in lowest score levels.

As expected, reports of *no* or *mild* pain (Column 2) and *little* or *no* interference with work (Column 4) were more common at the highest BP score levels but decreased as scores decreased. Both items appear to be most useful in interpreting scores in the middle score ranges. A notable and rapid drop occurred in reports of *no* or *mild* pain from Level 2 to Level 3 (100% to 1.4%, respectively), and then through Levels 3 to 7 (1.4% to 0.0%). A rapid but more moderated drop in percentages was also apparent when considering *little* or *no* interference with work, making these responses to the two items most useful with only those scoring in the middle BP score levels.

Figure 8.31 presents a graph of the percentage of the sample scoring at each BP score level that reported each limitation or characteristic defined or evaluated in Table 8.31.

General Health (GH)

Table 8.32 presents reports of respondents' negative perceptions of their health in all of the five areas, which did not occur or were infrequent at the three highest GH score levels. Reports of *fair* or *poor* health (Column 1), getting sick easier (Column 2), and expecting health to get worse (Column 4) as *mostly* or *definitely true* and reports of feeling as healthy as anybody (Column 3) and excellent health (Column 5) as *mostly* or *definitely false* generally became more common at GH score Level 5 (*T* = 45.0–49.9) and then increased in a linear manner through the lower score levels. Overall, the five items are most useful in interpreting GH score differences in the middle and lowest score levels.

Figure 8.32 presents a graph of the percentage of the sample scoring at each GH score level that reported each limitation or characteristic defined or evaluated in Table 8.32.

Vitality (VT)

Inspection of Table 8.33 reveals the same pattern of scores for the four VT items that was seen with the five GH items just previously discussed. Similarly, the indicated responses to the VT items are most useful in interpreting score differences in the middle and lowest score levels.

Figure 8.33 presents a graph of the percentage of the sample scoring at each VT score level that reported each limitation or characteristic defined or evaluated in Table 8.33.

Social Functioning (SF)

Table 8.34 indicates that the frequency and degree to which health interfered with social activities generally was related to the seven SF score levels. One hundred

percent of respondents scoring at the two highest SF score levels (Levels 1 and 2) reported that their health interfered with social activities either *slightly* or *not at all* (Column 2) and either *little* or *none of the time* (Column 4). There was a rapid decrease in the percentages reporting limited or no interference and interference occurring infrequently or never, beginning at SF score Level 3 (T -score range = 45.0–49.9), ending in 0.0% at Level 7 (T -score range < 25). The opposite trend was seen for those reporting health interfering with social activities as *moderately*, *quite a bit* or *extremely* (Column 1) and *most* or *all of the time* (Column 3). Notable were the rapidly increasing percentages of those reporting significant interference at the lowest three SF score levels (Levels 5–7), with 100.0% reporting problems at the lowest level (Level 7) for both variables. In all, these items are most helpful in interpreting SF score differences in the middle and lowest score levels.

Figure 8.34 presents a graph of the percentage of the sample scoring at each SF score level that reported each limitation or characteristic defined or evaluated in Table 8.34.

Role-Emotional (RE)

The percentages of those reporting limitations in role functioning due to emotional problems are presented in Table 8.35. No respondents scoring at or above the average range (Levels 1–3) on the RE scale reported having to cut down on time spent at work or other activities (Column 1), accomplishing less than they would like (Column 2), and working less carefully (Column 3) *most* or *all of the time*, while all respondents scoring at the lowest RE score level (Level 7) reported experiencing these problems at either of those frequencies. Overall, these items are most helpful in interpreting RE score differences in the middle and lowest score levels.

Figure 8.35 presents a graph of the percentage of the sample scoring at each RE score level that reported each limitation or characteristic defined or evaluated in Table 8.35.

Mental Health (MH)

Table 8.36 shows that problems in emotional health and well-being generally increased with decreasing MH scale scores. Beginning at about score Level 3 (T -score range = 50.0–54.9), reports of being very nervous *most* or *all of the time* (Column 1) and calm (Column 3) and happy (Column 5) *little* or *none of the time* began to appear. However, the two MH scale items that are most indicative of the presence of depression—feeling down in the dumps (Column 2) and downhearted and depressed (Column 4) *most* or *all of the time*—did not

have notable percentages (5.4% and 15.3%, respectively) of respondents reporting these symptoms until MH score Level 6 (T -score range = 35.0–39.9), percentages which then quickly increased through the lower score levels. Reports of problems with feeling downhearted and depressed and in not feeling calm, peaceful, and happy all reached 100% at the lowest score level (Level 8), while reports of being very nervous and feeling down in the dumps reached 100% at the lowest score level. Overall, item responses are most helpful in interpreting MH score differences in the middle and lowest levels for the nervous, downhearted and depressed, and happiness items. Whereas feeling down in the dumps is most helpful in interpreting score differences at the lowest levels, not feeling calm and peaceful is useful throughout all score ranges.

Figure 8.36 presents a graph of the percentage of the sample scoring at each MH score level that reported each limitation or characteristic defined or evaluated in Table 8.36.

Interpolation of Score-Related Percentages

Because only the score ranges and means within those ranges for the component summary measure scores, health domain scales, and item response criteria are printed in Tables 8.1 through 8.36, users must calculate ratios of differences and interpolate to estimate the percentage that is associated with a specific score. Percentages interpolated for two specific scores can then be used to relate differences in scores to differences in criterion percentages that occur within or across the levels on the table. The remainder of this section illustrates the use of this method through an example in which the percentage of the population reporting having a lot of energy only *a little* or *none of the time* is estimated based on interpolation of MCS scores using data from a portion of Table 8.6 and reproduced in Table 8.37.

Suppose that a group of respondents undergoing treatment for depression had an average MCS T score of 35.3 before the treatment and 39.5 after 4 weeks of treatment. It would be useful to know how this change in the MCS score is related to predicted differences in the percentage of respondents who reported having a lot of energy only *a little* or *none of the time*. Table 8.37 shows the percentage of respondents reporting having a lot of energy *a little* or *none of the time* at each of 9 T -score ranges on the MCS measure, as well as the mean MCS score within each of those ranges. In this

Table 8.28

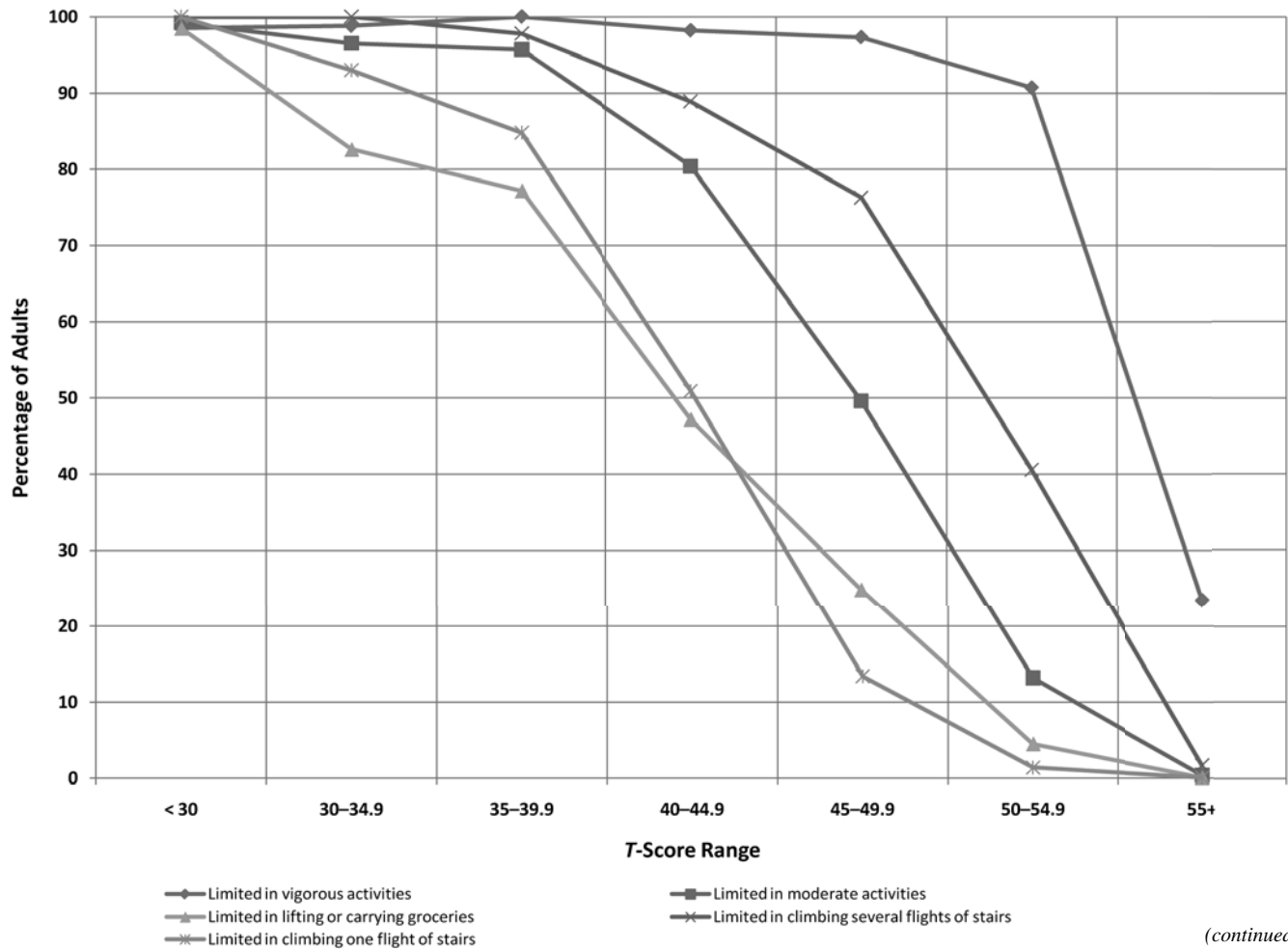
Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 2,059)

PF T-Score Level	T Scores		n	Limited in vigorous activities ^a	Limited in moderate activities ^b	Limited in lifting or carrying groceries ^c	Limited in climbing several flights of stairs ^d	Limited in climbing one flight of stairs ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	55+	57.02	969	23.4	0.4	0.1	1.7	0.1
2	50–54.9	52.90	355	90.7	13.2	4.5	40.6	1.4
3	45–49.9	48.18	263	97.3	49.6	24.8	76.3	13.4
4	40–44.9	42.37	163	98.2	80.4	47.2	88.9	50.9
5	35–39.9	37.72	92	100.0	95.7	77.2	97.8	84.8
6	30–34.9	32.53	86	98.8	96.5	82.6	100.0	93.0
7	< 30	23.57	131	98.5	99.2	98.5	100.0	100.0

^a% reporting any limitations in vigorous activities (Item 3a).
^b% reporting any limitations in moderate activities (Item 3b).
^c% reporting any limitations in lifting or carrying groceries (Item 3c).
^d% reporting any limitations in climbing several flights of stairs (Item 3d).
^e% reporting any limitations in climbing one flight of stairs (Item 3e).

(continued)

Figure 8.28 Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 2,059)



(continued)

Table 8.29

Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 2,059) (continued)

PF T-Score Level	T Scores		n	Limited in bending, kneeling, or stooping ^f	Limited in walking more than a mile ^g	Limited in walking several hundred yards ^h	Limited in walking 100 yards ⁱ	Limited in bathing yourself ^j
	Range	Mean		(6) %	(7) %	(8) %	(9) %	(10) %
1	55+	57.02	969	4.4	1.7	0.1	0.1	0.0
2	50–54.9	52.90	355	45.3	26.3	3.4	3.7	0.9
3	45–49.9	48.18	263	69.5	69.2	13.7	9.5	1.9
4	40–44.9	42.37	163	84.7	92.0	57.1	30.0	8.0
5	35–39.9	37.72	92	92.4	97.8	90.0	69.2	24.4
6	30–34.9	32.53	86	91.9	100.0	94.2	91.9	26.7
7	< 30	23.57	131	100.0	100.0	100.0	98.5	71.8

^f% reporting any limitations in bending, kneeling, or stooping (Item 3f).

^g% reporting any limitations in walking more than a mile (Item 3g).

^h% reporting any limitations in walking several hundred yards (Item 3h).

ⁱ% reporting any limitations in walking 100 yards (Item 3i).

^j% reporting any limitations in bathing or dressing oneself (Item 3j).

Figure 8.29 Percentage of Adults Reporting Limitations in Physical Functioning at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scale Scores, 2009 U.S. General Population (N = 2,059) (continued)

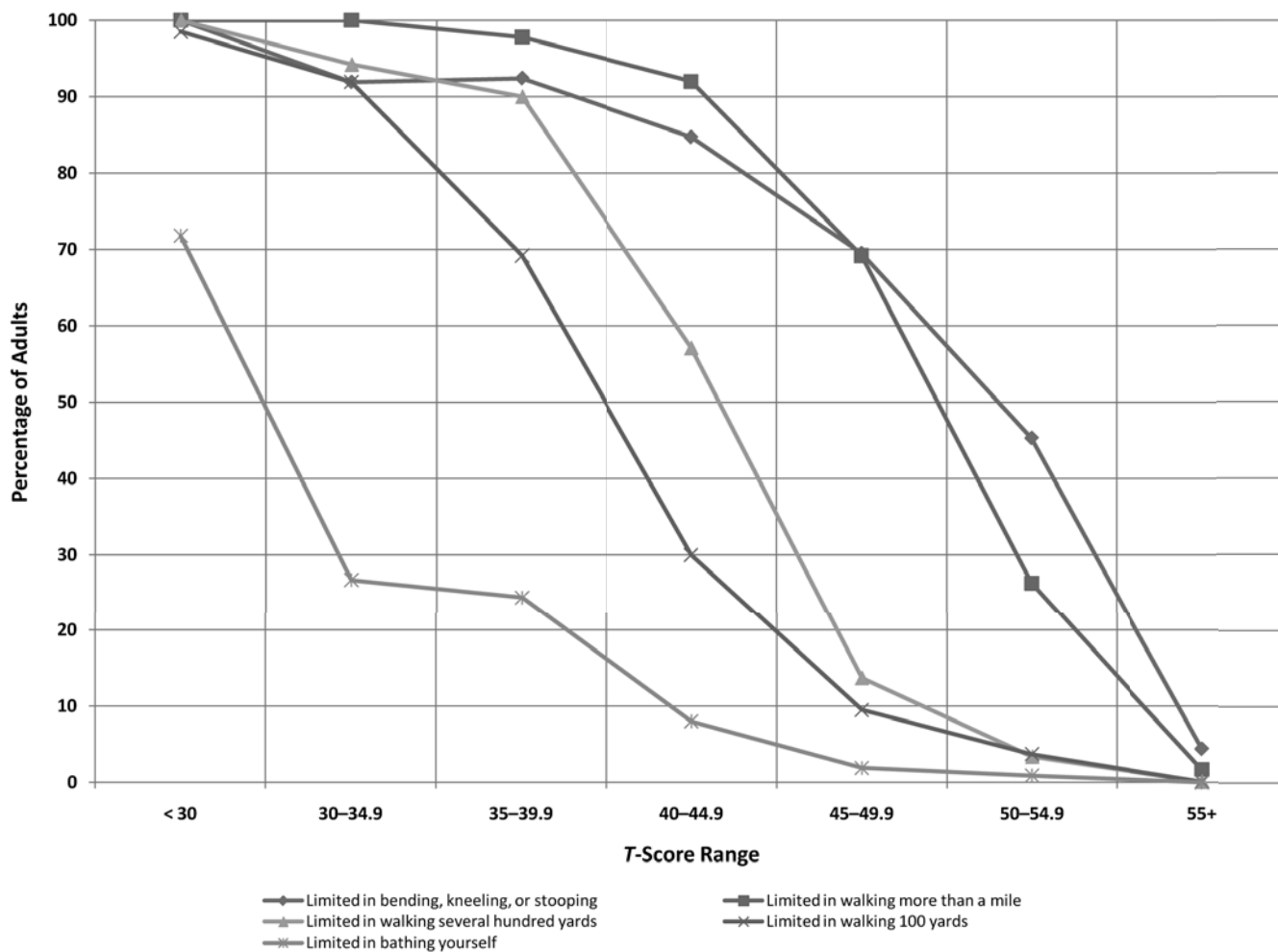


Table 8.30

Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (N = 2,057)

RP T-Score Level	T Scores		n	Cut down time at work <i>most</i> or <i>all of the time</i> ^a	Accomplished less <i>most</i> or <i>all of the time</i> ^b	Limited in the kind of work or other activities <i>most</i> or <i>all of the time</i> ^c	Difficulty at work <i>most</i> or <i>all of the time</i> ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	55+	57.12	1,025	0.0	0.0	0.0	0.0
2	50–54.9	52.85	361	1.1	0.3	0.0	0.0
3	45–49.9	47.53	205	0.0	0.5	0.0	1.0
4	40–44.9	42.83	102	4.9	4.9	5.9	6.9
5	35–39.9	37.94	166	7.8	25.3	25.3	30.9
6	30–34.9	31.50	96	65.6	90.6	92.7	91.7
7	< 30	24.56	102	87.3	100.0	100.0	100.0

^a% reporting having cut down amount of time spent on work or other activities *most* or *all of the time* (item 4a).

^b% reporting having accomplished less than they would like *most* or *all of the time* (item 4b).

^c% reporting being limited in the kind of work or other activities *most* or *all of the time* (item 4c).

^d% reporting having had difficulty performing work or other activities *most* or *all of the time* (item 4d).

Figure 8.30 Percentage of Adults Reporting Limitations in Role Functioning Due to Physical Health at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Physical Scale Scores, 2009 U.S. General Population (N = 2,057)

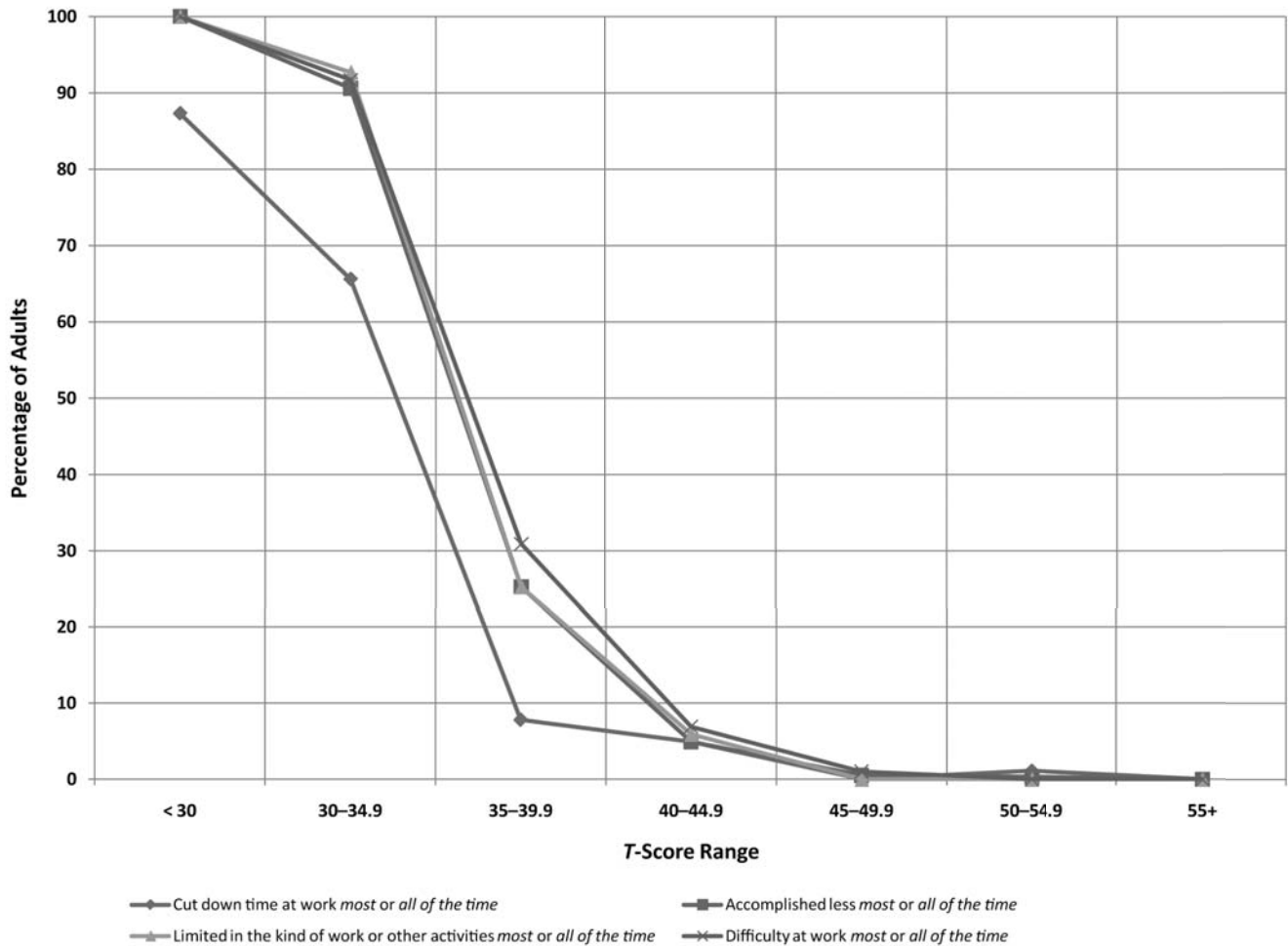


Table 8.31

Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (N = 2,056)

BP T-Score Level	T Scores		n	Severe or very severe pain ^a	No or very mild pain ^b	Quite a lot or extreme interference with normal work ^c	Little or no interference with work ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	55+	60.85	544	0.0	100.0	0.0	100.0
2	50–54.9	53.42	667	0.0	100.0	0.0	100.0
3	45–49.9	47.22	351	0.0	1.4	0.0	98.6
4	40–44.9	41.60	146	1.4	1.4	1.4	78.1
5	35–39.9	37.58	151	7.3	0.0	0.0	7.3
6	30–34.9	31.83	153	63.8	0.0	89.5	0.0
7	< 30	23.77	44	97.7	0.0	95.4	0.0

^a% reporting *severe* or *very severe* bodily pain (Item 7).

^b% reporting *no* or *very mild* pain (Item 7).

^c% reporting that pain interferes with normal work (inside and outside the home) *quite a lot* or *extremely* (Item 8).

^d% reporting that pain interferes with normal work (inside and outside the home) a *little bit* or *not at all* (Item 8).

Figure 8.31 Percentage of Adults Reporting Bodily Pain or Impact of Pain on Work at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Bodily Pain Scale Scores, 2009 U.S. General Population (N = 2,056)

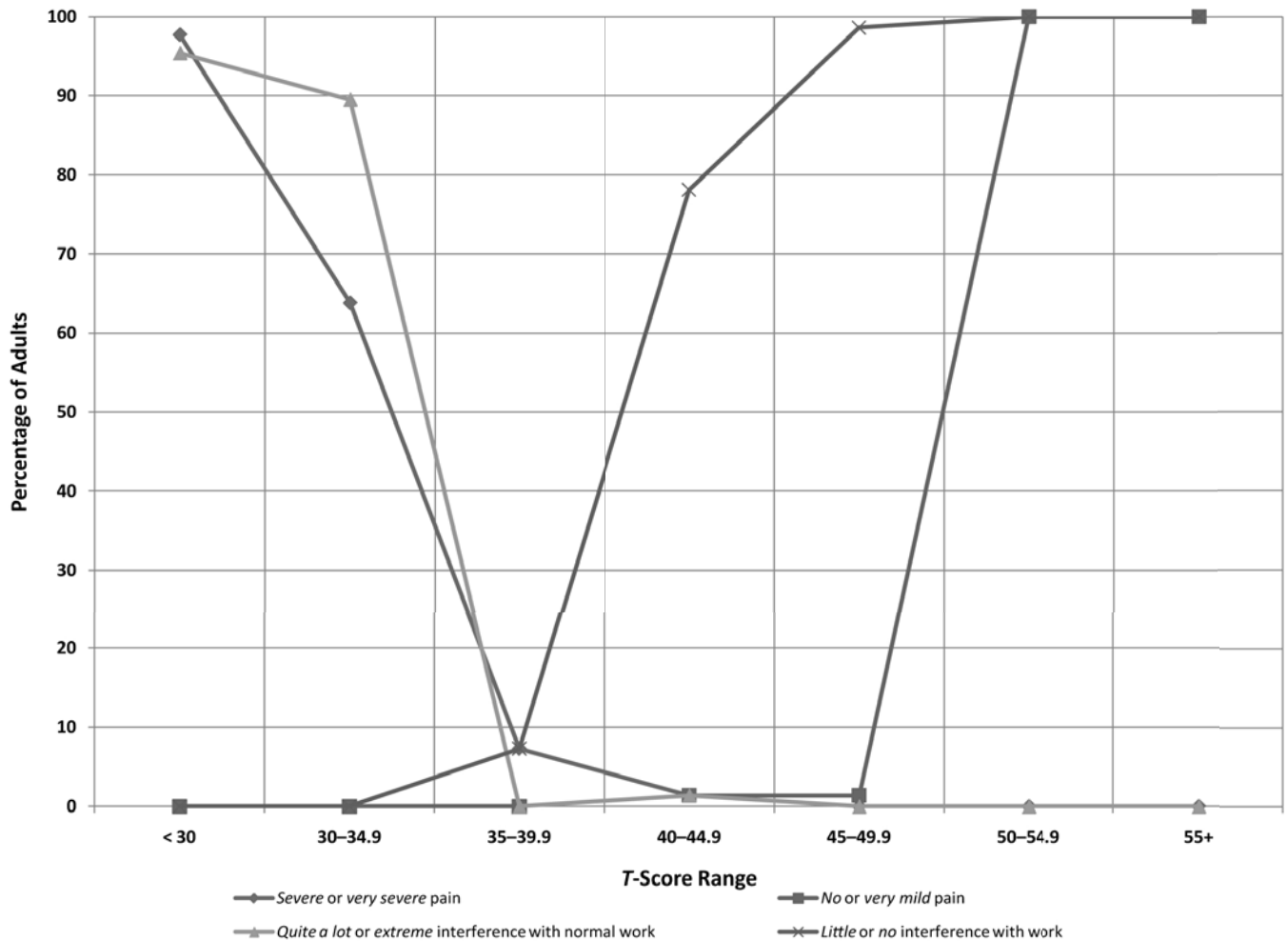


Table 8.32

Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (N = 2,061)

GH T-Score Level	T Scores		n	Fair or poor health ^a	Getting sick easier mostly or definitely true ^b	As healthy as anybody mostly or definitely false ^c	Health expected to get worse mostly or definitely true ^d	Health is excellent mostly or definitely false ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	65+	65.40	84	0.0	0.0	0.0	0.0	0.0
2	60–64.9	62.23	211	0.0	0.0	0.0	0.0	0.0
3	55–59.9	57.25	587	0.0	0.2	1.0	2.2	0.3
4	50–54.9	52.15	346	0.6	2.6	5.2	14.5	5.5
5	45–49.9	47.61	258	8.1	7.8	9.0	19.0	34.4
6	40–44.9	43.09	210	23.6	8.6	30.4	30.5	72.3
7	35–39.9	38.09	183	53.6	19.1	63.9	39.9	96.2
8	30–34.9	32.23	128	93.8	36.7	92.2	59.4	99.2
9	< 30	25.36	54	98.1	75.5	98.1	87.0	100.0

^a% reporting fair or poor health (Item 1).

^b% reporting getting sick easier as mostly true or definitely true (Item 11a).

^c% reporting being as healthy as anybody they know as mostly false or definitely false (Item 11b).

^d% reporting health expected to get worse as mostly true or definitely true (Item 11c).

^e% reporting health is excellent as mostly false or definitely false (Item 11d).

Figure 8.32 Percentage of Adults Reporting General Health Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scale Scores, 2009 U.S. General Population (N = 2,061)

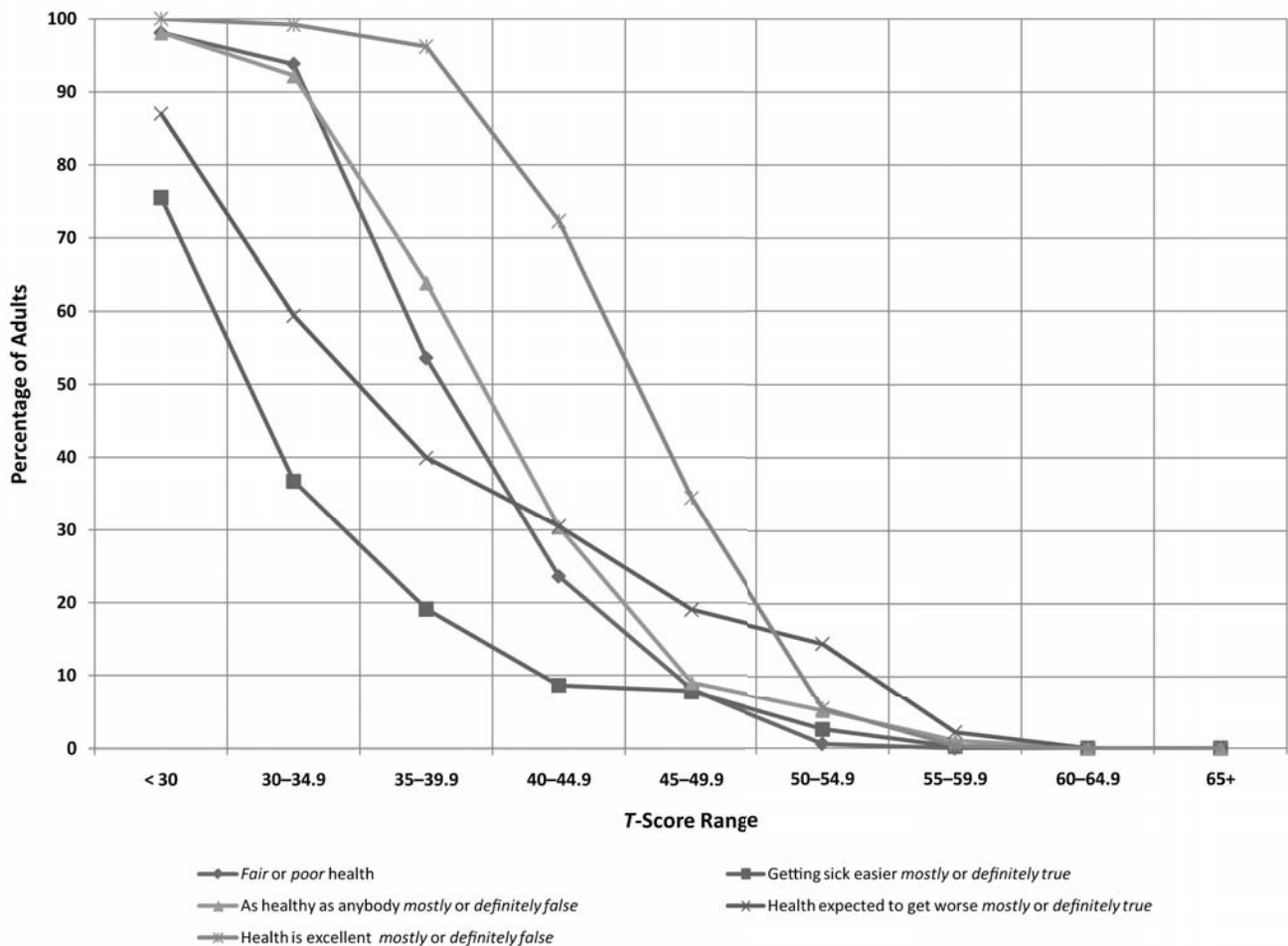


Table 8.33

Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (N = 2,057)

VT T-Score Level	T Scores		n	Feeling full of life little or none of the time ^a	Having a lot of energy little or none of the time ^b	Feeling worn out most or all of the time ^c	Feeling tired most or all of the time ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	65+	67.93	93	0.0	0.0	0.0	0.0
2	60–64.9	61.77	238	0.0	0.0	0.0	0.0
3	55–59.9	57.15	522	1.0	0.8	0.2	0.6
4	50–54.9	51.38	430	4.2	8.4	1.4	3.5
5	45–49.9	47.38	178	13.6	25.8	7.3	13.0
6	40–44.9	43.51	250	45.4	60.0	24.0	38.2
7	35–39.9	38.00	193	70.8	91.2	57.8	82.3
8	30–34.9	32.45	98	92.9	97.9	88.7	99.0
9	< 30	26.80	55	100.0	100.0	100.0	100.0

^a% reporting feeling full of life little or none of the time (Item 9a).

^b% reporting having a lot of energy little or none of the time (Item 9e).

^c% reporting feeling worn out most or all of the time (Item 9g).

^d% reporting feeling tired most or all of the time (Item 9i).

Figure 8.33 Percentage of Adults Reporting Limitations in Vitality at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Vitality Scale Scores, 2009 U.S. General Population (N = 2,057)

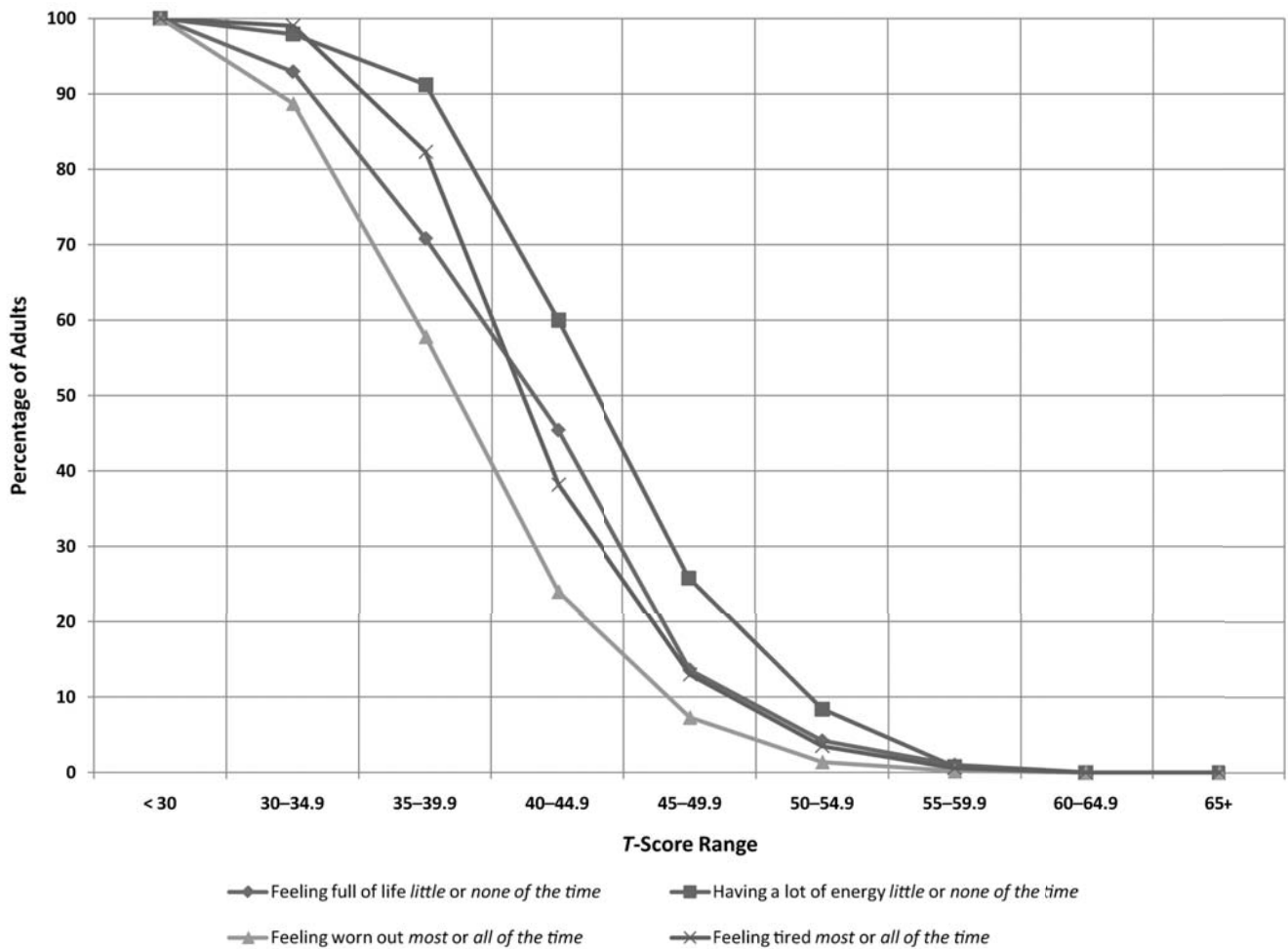


Table 8.34

Percentage of Adults Reporting Limitations in Social Activities at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (N = 2,057)

SF T-Score Level	T Scores		n	Health interfered with social activities <i>moderately, quite a bit, or extremely</i> ^a	Health interfered with social activities <i>slightly or not at all</i> ^b	Health interfered with social activities <i>most or all of the time</i> ^c	Health interfered with social activities <i>little or none of the time</i> ^d
	Range	Mean		(1) %	(2) %	(3) %	(4) %
1	55+	56.74	1,232	0.0	100.0	0.0	100.0
2	50–54.9	51.79	211	0.0	100.0	0.0	100.0
3	45–49.9	46.85	192	9.9	90.1	0.0	87.0
4	40–44.9	41.91	139	33.1	66.9	5.8	33.1
5	35–39.9	36.97	123	79.5	20.5	20.5	18.9
6	25–34.9	29.87	111	97.3	2.7	70.9	3.6
7	< 25	19.83	49	100.0	0.0	100.0	0.0

^a% reporting physical or emotional problems interfering with social activities *moderately, quite a bit, or extremely* (Item 6).

^b% reporting physical or emotional problems interfering with social activities *slightly or not at all* (Item 6).

^c% reporting physical or emotional problems interfering with social activities *most or all of the time* (Item 10).

^d% reporting physical or emotional problems interfering with social activities *little or none of the time* (Item 10).

Figure 8.34 Percentage of Adults Reporting Limitations in Social Activities at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Social Functioning Scale Scores, 2009 U.S. General Population (N = 2,057)

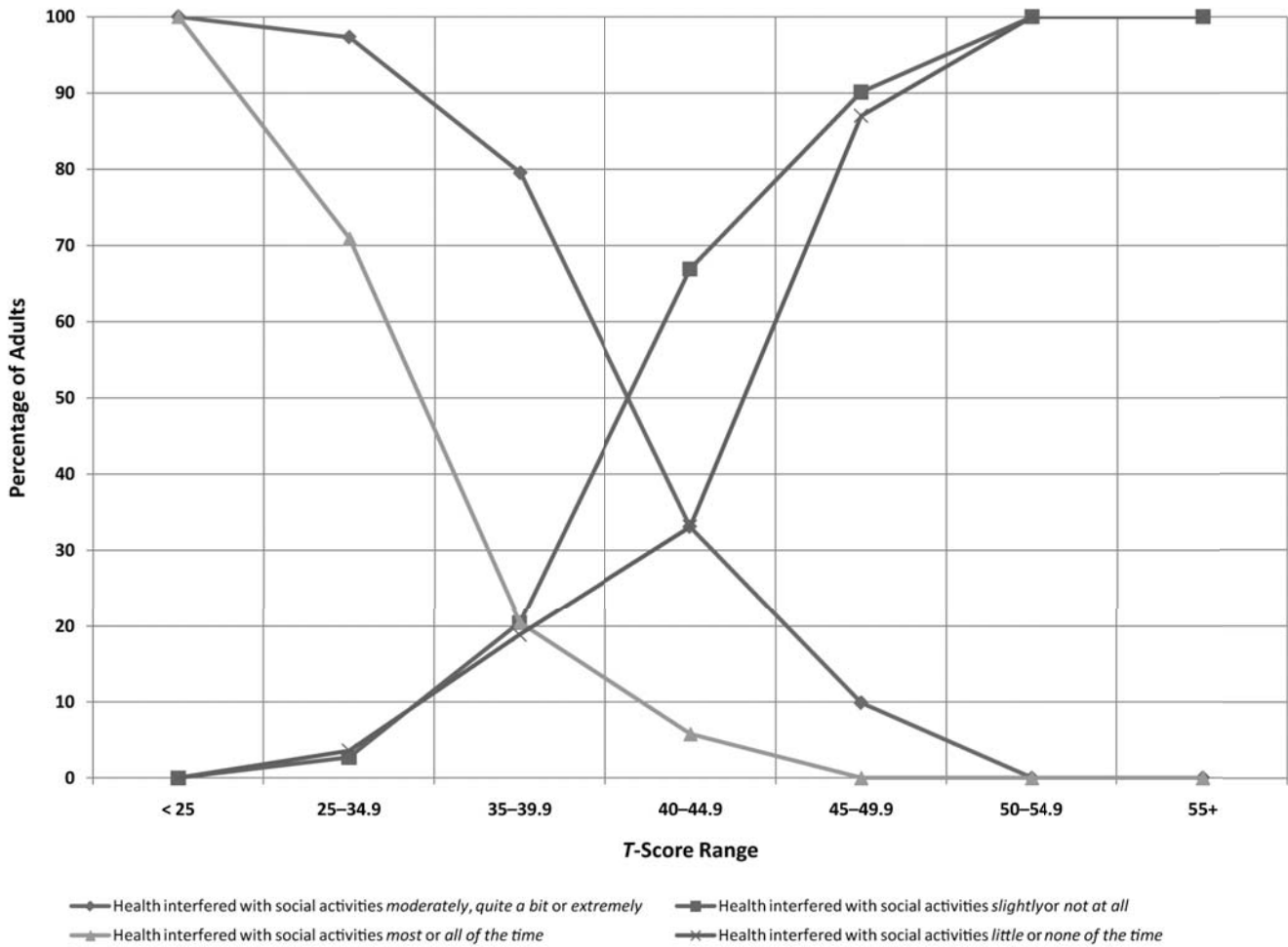


Table 8.35

Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (N = 2,057)

RE T-Score Level	T Scores		n	Cut down time at work most or all of the time ^a	Accomplished less most or all of the time ^b	Did work less carefully most or all of the time ^c
	Range	Mean		(1) %	(2) %	(3) %
1	55+	55.64	1,388	0.0	0.0	0.0
2	50–54.9	51.82	102	0.0	0.0	0.0
3	45–49.9	48.01	166	0.0	0.0	0.0
4	40–44.9	42.92	179	2.8	0.6	0.0
5	30–39.9	34.18	125	13.7	11.3	0.8
6	20–29.9	24.07	66	80.0	83.3	56.1
7	< 20	11.84	31	100.0	100.0	100.0

^a% reporting cutting down amount of time spent on work or other activities *most or all of the time* (Item 5a).

^b% reporting accomplished less than they would like *most or all of the time* (Item 5b).

^c% reporting did work or other activities less carefully *most or all of the time* (Item 5c).

Figure 8.35 Percentage of Adults Reporting Limitations in Role Functioning Due to Emotional Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scale Scores, 2009 U.S. General Population (N = 2,057)

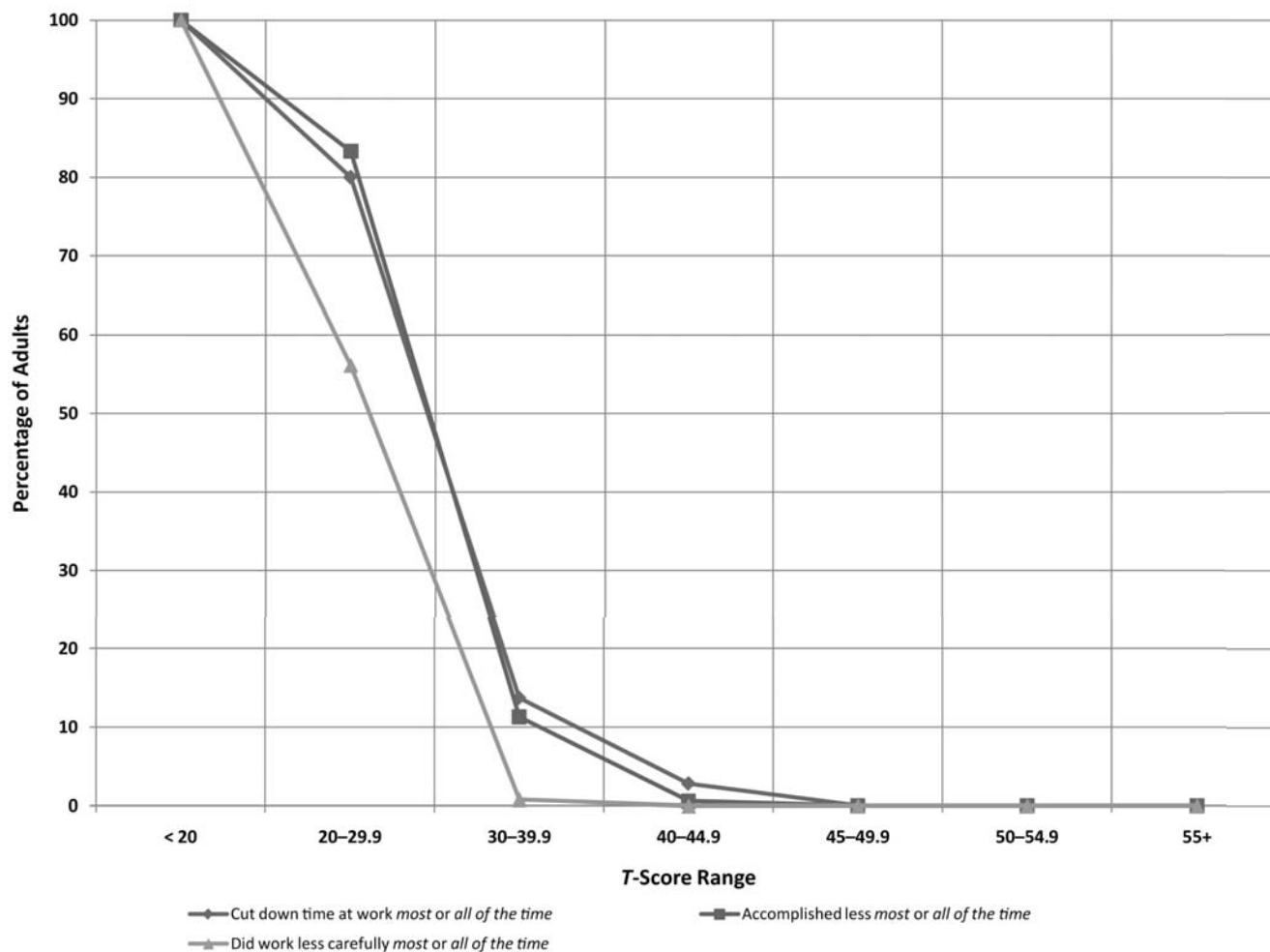


Table 8.36

Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (N = 2,060)

MH T-Score Level	T Scores		n	Been very nervous most or all of the time ^a	Down in dumps most or all of the time ^b	Calm little or none of the time ^c	Downhearted and depressed most or all of the time ^d	Happy little or none of the time ^e
	Range	Mean		(1) %	(2) %	(3) %	(4) %	(5) %
1	60+	61.44	280	0.0	0.0	0.0	0.0	0.0
2	55–59.9	56.74	769	0.0	0.0	0.1	0.0	0.0
3	50–54.9	51.65	343	1.5	0.0	5.0	0.0	1.8
4	45–49.9	46.65	238	2.1	0.0	23.6	0.0	15.1
5	40–44.9	41.67	178	5.7	0.6	50.0	2.3	38.2
6	35–39.9	36.93	112	14.3	5.4	58.2	15.3	53.6
7	25–34.9	29.56	111	39.5	36.9	92.8	63.6	86.4
8	< 25	20.14	29	89.3	93.1	100.0	100.0	100.0

^a% reporting being very nervous most or all of the time (Item 9b).

^b% reporting being so down in the dumps that nothing could cheer them up most or all of the time (Item 9c).

^c% reporting being calm and peaceful little or none of the time (Item 9d).

^d% reporting being downhearted and depressed most or all of the time (Item 9f).

^e% reporting being happy little or none of the time (Item 9h).

Figure 8.36 Percentage of Adults Reporting Emotional Distress at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scale Scores, 2009 U.S. General Population (N = 2,060)

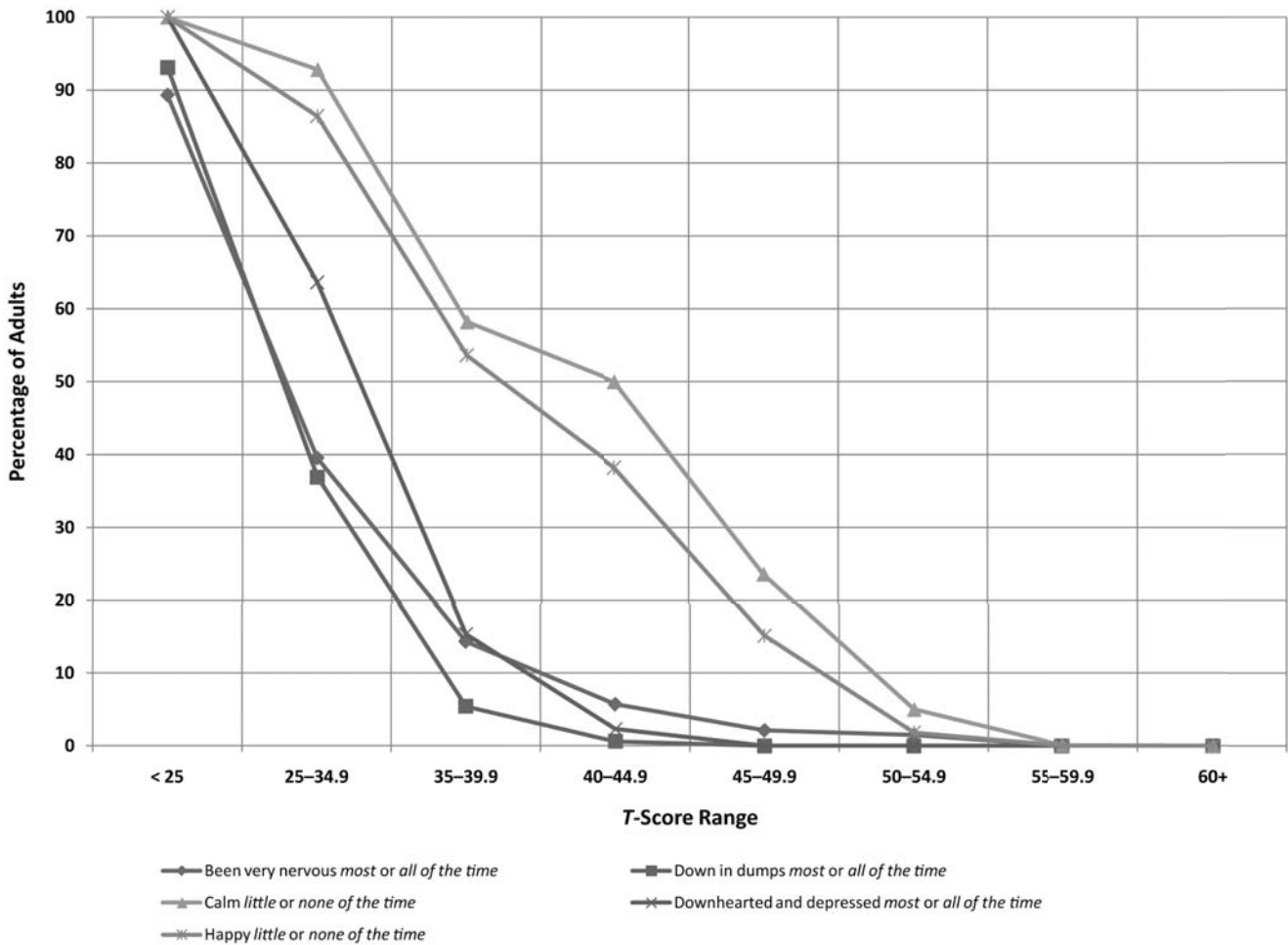


Table 8.37

Percentage of Adults Reporting Feeling Energetic Little or None of the Time at 9 Levels of SF-36v2 Mental Component Summary Measure Scores, 2009 U.S. General Population (N = 4,024)

MCS T-Score Level	T Scores		n	Having a lot of energy little or none of the time %
	Range	Mean		
1	60+	62.24	492	7.0
2	55–59.9	57.39	1,251	9.0
3	50–54.9	52.82	886	18.7
4	45–49.9	47.79	509	33.8
5	40–44.9	42.56	338	49.3
6	35–39.9	37.69	219	58.1
7	30–34.9	32.65	129	71.3
8	25–29.9	27.53	103	78.4
9	< 25	19.42	97	89.7

example, the mean MCS scores both before (35.3) and after (39.5) treatment fall within one score level (Level 6). If either of these scores were equal to the mean for that level (37.69), then the percentage of respondents having a lot of energy *a little or none of the time* (58.1%) could be determined directly from the table. However, because the percentages associated with the pre- and posttreatment scores are not presented in the table, they must be estimated by interpolation. First, the appropriate difference ratio (change in percentage per unit change in MCS score) must be calculated. Then, this ratio is used to estimate the percentage for a specific MCS score.

Thus, to estimate the predicted change in the percentage of respondents reporting having a lot of energy *a little or none of the time* that corresponds to the change in this example's MCS scores before and after treatment:

1. *Identify the values needed to calculate the change in the criterion (i.e., having a lot of energy a little or none of the time) per unit change in MCS (i.e., the difference ratio).* Look at the means for each level, and choose those levels in which the means are just lesser than and greater than the lower and higher scores, respectively. Using Table 8.37, the MCS scores for before (35.3) and after (39.5) treatment both fall within the range of the mean MCS scores for Levels 5 (42.56) and 7 (32.65).
2. *Calculate the change in the percentage meeting the criterion per unit change in MCS score.* At Level 5, the mean MCS score is 42.56 and the percentage having a lot of energy *a little or none of the time* is 49.3. At Level 7, the mean MCS score is 32.65 and the percentage is 71.3.

Determine the difference in the percentages associated with the mean MCS scores at Levels 7 and 5 ($71.3\% - 49.3\% = 22.0\%$), determine the difference in the means at Levels 5 and 7 ($42.56 - 32.65 = 9.91$), and divide the resulting percentage by the mean difference ($22.0\% \div 9.91$). Thus, the percentage change per unit of change in MCS score is 2.22%.

3. *Calculate the percentage meeting the criterion at one score value.* The percentage of respondents reporting having a lot of energy *a little or none of the time* at the pretreatment score of 35.3 should be less than the mean percentage for Level 7 (32.65) because the trend in the data demonstrates that the rate for having a lot of energy *a little or none of the time* goes down as MCS scores go up (i.e., higher scores indicate better health). To estimate the percentage associated with the pretreatment score, subtract the mean MCS score at Level 7 (32.65) from the pretreatment MCS score (35.3), and multiply this result by the percentage change in having a lot of energy *a little or none of the time* per unit of change in MCS score ($[35.3 - 32.65] \times 2.22\% = 5.88\%$). Subtract this result (5.88%) from the percentage associated with a score of 32.65 (71.3%) to determine the percentage of respondents who would be expected to have a lot of energy *a little or none of the time* at a score of 35.3 (65.42%).
4. *Calculate the percentage meeting the criterion at the other score value.* To estimate the percentage of respondents reporting a lot of energy *a little or none of the time* at the average posttreatment MCS score of 39.5, multiply the difference between the pre- and posttreatment MCS scores ($39.5 - 35.3 = 4.2$) by the percentage change in having a lot of energy *a little or none of the time* per unit of change in the MCS score ($4.2 \times 2.22\% = 9.32\%$). Subtract this result (9.32%) from the percentage of respondents having a lot of energy *a little or none of the time* at the pretreatment score (65.42%), which results in a posttreatment percentage of 56.1%.

Thus, in this example, the percentage of respondents who would have a lot of energy *a little or none of the time* at a pretreatment MCS score of 35.3 is estimated to be 65.42%, whereas the percentage of respondents expected to earn a posttreatment score of 39.5 is estimated to be 56.1%. Moreover, the percentage change, or reduction, of those meeting the criterion from pre- to

posttreatment is estimated to be 9.32% ($65.42\% - 56.1\% = 9.32\%$). When interested in estimating the difference in the percentage meeting the criterion between two scores that do not fall within the same score level, compute each score-related percentage according to Steps 1 through 3 above, and then subtract one from the other. For example, to estimate the percentage change in MCS scores of 36 and 47: (a) first, using Steps 1 through 3, compute the difference ratio and estimated percentage associated with a score of 36; (b) then, again using Steps 1 through 3, compute the difference ratio and estimated percentage associated with a score of 47; and (c) finally, subtract the estimated percentage associated with a score of 47 from the estimated percentage associated with a score of 36.

Note that this method yields approximations and that simpler rather than more complex calculations are used to promote better understanding. For example, the difference ratios derived in the preceding example are based on simple averaging and assume a linear relationship between score levels. When the values associated with score levels greatly differ, a more accurate approach would be to put greater weight on the values that are closer to the score of interest. The calculations described here provide simple averages for difference ratios for each of several levels, thereby capturing some of the variation of change in criterion associated with change in scores at different levels of scale scores.

9

Criterion-Based Interpretation

Criterion-based interpretation guidelines are based on analyses of the relationships between the measures in question and other variables, referred to as *criteria*, measured either concurrently or after a period of time. The empirical strategy for evaluating the meaningfulness of SF-36v2 *T* scores derived from the 2009 norming study has been to link health domain scale and component summary measure scores to important clinical and nonclinical variables. By this logic, information about *importance* is gathered by linking differences in scores to these variables that have well-understood effects on the domains and components of health measured by the SF-36v2 and by showing how differences in scores of a certain magnitude predict important clinical and nonclinical variables. Previously published examples of criterion-based interpretation of the SF-36 and SF-36v2 surveys can be found in Ware, Snow, Kosinski, and Gandek (1993) and Ware et al. (2007), respectively. The purpose of this chapter is to present the results of analyses that were designed to yield interpretation guidelines for differences in SF-36v2 health domain and component summary measure scores, based on their relationships with other variables that were assessed along with and/or 3 to 4 months after the SF-36v2 during the 2009 norming study.

Interpretation of Scales and Measures Across All Score Ranges

Like the content-based interpretation guidelines presented in Chapter 8, criterion-based interpretation guidelines for each component summary measure and health domain scale, across all score ranges, were developed in several steps. First, the background, validation, chronic condition, and health care utilization variables included on the 2009 normative study forms containing SF-36v2 items—Forms A, B, and C—that were deter-

mined to be conceptually related to the health domain scales and component summary measures were initially considered for recommendation. Specifically, those considered for inclusion in the guidelines were variables that (a) were clinically or socially important (e.g., clinical diagnosis, employment status); (b) represented plausible outcomes of the variations in physical, social, and role functioning and in pain, vitality, and mental health; and (c) were measured independent of the SF-36v2 health domain scales and component summary measures. The data collected (including SF-8 item data) from the 2009 U.S. general population sample were then used for the analyses to develop these criterion-based interpretation guidelines.

Second, responses to each of the criterion variables selected for consideration were dichotomized in a meaningful way that was thought to reveal differences across levels of the scale or measure in the SF-36v2 score ranges of interest. For each component summary measure, the same variables and associated responses that were examined for the health domain scales most highly correlated with each measure—the Physical Functioning (PF), Role-Physical (RP), Bodily Pain (BP), and General Health (GH) scales with the PCS measure, and the Vitality (VT), Social Functioning (SF), Role-Emotional (RE), and Mental Health (MH) scales with the MCS measure—were selected for examination as potential bases for interpretation of that measure.

Third, the percentage of the 2009 normative sample that entered a criterion-related response for a given external variable at each respective SF-36v2 health domain scale and component summary measure *T*-score level being interpreted was determined. Generally, the score levels, which range from 7 to 11 depending on the scale or measure, represent 5-point *T*-score intervals throughout the range of scores observed in the 2009 U.S. general population for each component summary measure and health domain scale. Often, the highest

score levels and the lowest score levels were combined to encompass larger, more meaningful ranges of scores at each end of the score range.

Fourth, the percentage of the 2009 normative sample that endorsed each variable response at each summary measure or scale *T*-score level was evaluated. Those variables that provided useful interpretations across the entire continuum or at particular summary or scale *T*-score levels were retained as recommended sources of criterion-based SF-36v2 interpretation. The 2009 SF-36v2 standard (4-week recall) form percentages are presented in Tables 9.1 through 9.28 and 2009 SF-36v2 acute (1-week recall) form percentages are presented in Tables 9.29 through 9.54.

To facilitate the interpretation of the results, the tables presented here follow the same format as the content-based interpretation tables in Chapter 8. Note that sample size is indicated for each individual variable, as not all sample members responded to all items pertaining to external (non-SF-36v2) variables.

Criterion-Based Interpretation of the Standard Form Component Summary Measures

Criterion-based interpretation of the SF-36v2 standard (4-week recall) form PCS measure is facilitated through an examination of the percentage of respondents from the 2009 normative sample at each level of PCS *T* scores whose responses to background, validation, chronic condition, and health care utilization items—thought to be conceptually related to the physical health dimension and likely to covary with changes in PCS scores—were indicative of problems or limitations imposed by the respondents' physical health status. Similarly, criterion-based interpretation of the MCS measure is facilitated through an examination of the percentage of respondents from the 2009 normative sample at each level of MCS *T* scores whose responses to background, validation, chronic condition, and health care utilization items—thought to be conceptually related to the mental health dimension and likely to covary with changes in MCS scores—were indicative of problems or limitations imposed by the respondents' mental health status.

Physical Component Summary (PCS)

Tables 9.1 through 9.5 provide data for the criterion-based interpretations of SF-36v2 standard form PCS *T* scores relative to limitations in physical and role-

functioning activities, pain interference, health care utilization, employment status, presence of chronic conditions, future health, and work-related problems, as well as ratings of quality of life, general health, and job performance.

General health, HRQOL, and PCS. Table 9.1 presents data related to the general health and quality of life criterion variables. For each of the five variables examined, there was a near perfect ordering across the 9 PCS *T*-score levels of the percentages of respondents reporting health-related problems or HRQOL as *fair* or *poor*. In other words, the percentages reporting such problems increased from the highest PCS score level (Level 1) to the lower levels (Levels 8 and 9). For example, from Levels 1 to 9, there was anywhere from a fivefold increase (see column labeled 5) to a 229-fold increase (Column 2). Notable percentages of respondents reporting chronic conditions experienced limitations in usual activities or enjoyment (Column 5) across the 9 score levels. Even at the highest score level (Level 1, *T*-score range = 55+), 17.8% of the respondents reported significant limitations in usual activities or enjoyment as a result of having a chronic condition.

Performance of work and other activities and PCS. With regard to the ability to work or engage in other activities (Table 9.2, Columns 1–3), there again was a linear ordering of increasing reports of problems from the highest to the lowest levels of PCS *T*-scores. In general, however, reports of being *disabled* were less common at the lowest PCS level (57.1% at Level 9) than having *quite a lot* of difficulty or being unable to do usual activities (100%) or do daily work (97.1%) due to physical problems.

On the other hand, further examination of Table 9.2 reveals that reports of significant negative deviations from the mean with regard to overall job performance rating (Column 4) and days of missed work due to illness or injury (Column 5) varied in a nonlinear manner across the PCS *T*-score levels. For example, the percentage of respondents reporting missing a significant number of workdays steadily increased from Level 1 (1.0%) through Level 5 (12.2%), followed by a slight decrease at Level 6 (11.4%), then increased more than threefold at Level 7 (38.1%), and declined thereafter to 0% at Level 9. Variability was also noted with ratings of overall job performance, but to a much lesser degree.

Health problems, treatment, and PCS. As Table 9.3 reveals, the amount of both inpatient and outpatient treatment gradually increased with decreasing PCS scores. Of those respondents scoring at the highest PCS score level, only 2.6% (Column 2) and 0.8% (Column 3) reported significantly more outpatient visits and inpatient stays,

respectively, than the general population. At PCS score Level 9, the percentage of those reporting significantly more outpatient visits increased 14-fold to 37.1%, while the percentage reporting significantly more inpatient stays increased 33-fold to 26.5%. Similarly, there was a perfect ordering of increasing percentages of respondents reporting significantly more days in bed than the general population, increasing 65-fold from 0.7% at Level 1 to 45.7% at Level 9.

Pain-related interference and PCS. Table 9.4 clearly demonstrates the increased probability that pain *very often* or *always* impacted or interfered with daily functioning and activities as PCS *T* scores decreased in the 2009 normative sample. This trend was particularly evident in the effect that pain had on the ability to enjoy life (Column 1), making simple tasks hard to complete (Column 2), and participating in leisure activities (Column 3), with the percentage of those reporting such problems and the rate at which the percentages increased from the higher to the lower PCS score levels being quite comparable. Meanwhile, a perfect ordering of increasing percentages from the highest level (1.0% at Level 1) to the lowest level (53.9% at Level 9) was also evident with regard to reports of feeling fed up and frustrated *very often* or *always* (Column 4); however, this generally appeared to be a less frequent problem than is the case for the other criterion variables dealing with pain interference.

Future health, work-related problems, and PCS. Table 9.5 demonstrates the relationship between a respondent's baseline PCS score level and the occurrence of health-related events assessed 3 to 4 months later. Generally, those scoring at the highest PCS score level (Level 1) were less likely to report one or more outpatient visits (Column 1) and/or one or more bed days due to illness or injury (Column 2) during the 4 weeks preceding reassessment, as well as were less likely to report not working at a paying job because of health (Column 3) at the time of reassessment, than those scoring at the lowest PCS score level (Level 7).

Mental Component Summary (MCS)

Tables 9.6 through 9.14 provide data for the criterion-based interpretations of SF-36v2 standard form MCS *T* scores relative to behavioral health; emotional, personal, and physical problems; interference of pain on functioning; ratings of quality of life, general health, and job performance; problems with sleep, cognitive functioning, and energy level; employment status; and future mental health problems.

Behavioral health problems and MCS. Table 9.6 illustrates the near perfect ordering across the 10 MCS score levels in the percentages of the general population

that reported feeling down/depressed/hopeless *more than half the days* (Column 1), was currently experiencing depression (Column 2), and was currently experiencing anxiety (Column 3). The percentage of those reporting each of these problems began to increase at MCS score Level 4 (*T*-score range = 45.0–49.9), which was the lower half of the average MCS range for the general population. On the other hand, the percentages of those reporting significantly more occasions of having five or more drinks than the general population (Column 4) was relatively high at the highest score levels (11.5% at Level 1 and 11.1% at Level 2), increased to 19.4% at Level 5, and then began a cycle of increasing and decreasing, with 15.4% reporting this behavior at the lowest MCS score level.

Effects of personal, emotional, and physical problems and MCS. As expected, Table 9.7 shows a perfect ordering across the 10 MCS score levels of increasing percentages of respondents reporting mental health-related problems: physical or emotional problems limiting social activities (Column 1), being bothered by emotional problems (Column 2), unhappiness or dissatisfaction with one's personal life (Column 3), and feeling little interest or pleasure in doing things (Column 4). Notable, however, is the relatively high percentage of those who reported problems with happiness or satisfaction in their personal lives (Column 3), beginning with 10.2% at score Level 3, which includes the mean MCS score, and then more than doubling (25.7%) at Level 4, which includes the lower half of the average range of scores. At least 90% of those scoring at the lowest MCS score level reported problems in each of these four criterion variables.

Job performance and the effects of stress and MCS. Table 9.8 reveals a perfect ordering of increasing percentages from MCS score Level 1 to 10 for the two stress-related variables (Columns 4 and 5). Similar to the happiness/life satisfaction variable in Table 9.7, stress-related problems were common across the score ranges, with almost one third (32.2%) reporting that they experienced *a good bit*, *quite a bit*, or *a great deal* of stress or pressure during the previous 4 weeks at Level 3, which includes the mean MCS score. Also, more than one fifth (21.5% at Level 4, which includes average range scores) reported that stress or pressure had affected their health *moderately*, *quite a lot*, or *extremely*.

Meanwhile, the results for the work-related variables reported in Table 9.8 reveal a somewhat different picture. With the exception of Level 1, there was a consistent increase in the percentage of respondents who reported that they *could not do* or were kept *quite a lot* from doing usual work, school, or other activities (Column 1) from the higher to lower MCS score levels. The below-average

Table 9.1

Percentage of Adults Reporting General Health and Quality of Life Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

Level	PCST Scores		Overall quality of life rated as <i>fair</i> or <i>poor</i> ^a			Health rated as <i>poor</i> or <i>very poor</i> ^b			Health rated significantly below the mean ^c			Number of chronic conditions significantly above the mean ^d			Chronic condition(s) limit usual activities/enjoyment <i>moderately</i> , <i>quite a lot</i> , or <i>extremely</i> ^e		
	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	
1	55+	1,518	58.27	2.8	1,520	58.27	0.3	1,514	58.27	2.0	1,520	58.27	2.9	898	58.20	17.8	
2	50-54.9	911	52.67	5.8	910	52.67	1.0	907	52.68	5.0	908	52.67	7.4	683	52.67	22.0	
3	45-49.9	501	47.71	10.6	501	47.72	2.4	495	47.71	8.5	500	47.71	15.4	430	47.68	34.0	
4	40-44.9	357	42.41	16.8	358	42.42	4.5	353	42.42	18.7	354	42.41	22.9	297	42.33	52.5	
5	35-39.9	264	37.54	33.0	265	37.54	9.8	262	37.52	31.7	264	37.54	25.8	240	37.49	68.8	
6	30-34.9	201	32.79	42.8	200	32.78	19.5	200	32.78	44.0	201	32.79	37.8	190	32.77	77.4	
7	25-29.9	150	27.65	54.0	150	27.65	37.3	148	27.66	43.2	149	27.65	45.0	146	27.64	90.4	
8	20-24.9	79	22.93	68.4	79	22.93	51.9	78	22.98	57.7	79	22.93	58.2	77	22.96	97.4	
9	< 20	35	16.64	74.3	35	16.64	68.6	34	16.71	70.6	35	16.64	54.3	35	16.64	94.3	

^a% rating overall quality of life as *fair* or *poor*.

^b% rating health as *poor* or *very poor* during the past 4 weeks.

^c% rating health 1 *SD* or more below the general population mean 0-100 rating during the past 4 weeks.

^d% reporting the number of chronic conditions now has as being 1 *SD* or more above the mean for the general population.

^e% reporting one or more chronic condition(s) now has that limit usual activities/enjoyment *moderately*, *quite a lot*, or *extremely*.

Table 9.2

Percentage of Adults Reporting Problems in Work Performance and Other Activities at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

Level	PCS T Scores		Current employment status is disabled ^a			Could not do or had quite a lot of difficulty doing usual activities due to physical conditions ^b			Could not do or had quite a lot of difficulty doing daily work ^c			Rating of overall job performance significantly below the mean ^d			Days of missed work due to illness/injury significantly above the mean ^e		
	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	
1	55+	1,516	58.28	0.8	1,517	58.27	0.3	1,510	58.27	0.2	1,032	58.18	6.6	1,037	58.20	1.0	
2	50-54.9	909	52.67	1.8	911	52.67	0.6	907	52.67	0.3	523	52.64	8.0	520	52.64	3.7	
3	45-49.9	500	47.71	4.0	500	47.71	4.6	499	47.71	3.0	251	47.77	13.2	251	47.78	4.4	
4	40-44.9	356	42.41	13.5	356	42.43	7.9	358	42.42	6.4	137	42.55	17.5	137	42.55	11.7	
5	35-39.9	265	37.54	21.5	265	37.54	29.8	265	37.54	19.3	73	37.43	12.3	74	37.44	12.2	
6	30-34.9	200	32.76	30.5	200	32.78	56.5	201	32.79	38.8	34	32.72	35.3	35	32.72	11.4	
7	25-29.9	150	27.65	38.0	150	27.65	79.3	149	27.64	61.1	21	26.96	19.1	21	26.96	38.1	
8	20-24.9	79	22.93	50.6	78	22.92	96.2	79	22.93	92.4	4	21.47	50.0	4	21.47	25.0	
9	< 20	35	16.64	57.1	35	16.64	100.0	35	16.64	97.1	2	19.13	50.0	2	19.13	0.0	

^a% reporting current work status as disabled.

^b% reporting could not do or were limited quite a lot in usual activities during the past 4 weeks.

^c% reporting could not do or had quite a lot of difficulty doing daily work during the past 4 weeks.

^d% rating overall job performance as being 1 SD or more below the mean 0-100 rating for the general population during the past 4 weeks.

^e% reporting number of days of work missed due to illness or injury as being 1 SD or more above the mean for the general population during the past 4 weeks (mean = 0.39, SD = 1.82).

Table 9.3

Percentage of Adults Reporting Health Problems and Treatment at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Days in bed due to illness/injury significantly above the mean ^a			Number of outpatient visits significantly above the mean ^b			Number of hospital stays significantly above the mean ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	1,515	58.27	0.7	1,514	58.27	2.6	1,513	58.28	0.8
2	50–54.9	907	52.67	1.1	906	52.68	5.4	907	52.68	1.9
3	45–49.9	498	47.70	4.4	496	47.70	13.1	499	47.71	5.0
4	40–44.9	355	42.41	7.9	351	42.41	17.7	351	42.41	7.7
5	35–39.9	261	37.56	11.1	262	37.56	19.9	260	37.57	7.7
6	30–34.9	199	32.76	17.1	201	32.79	27.9	198	32.79	8.6
7	25–29.9	149	27.65	28.2	148	27.65	37.2	149	27.65	14.1
8	20–24.9	77	22.96	40.3	79	22.93	32.9	79	22.93	16.5
9	< 20	35	16.64	45.7	35	16.64	37.1	34	16.71	26.5

^a% reporting number of days in bed due to illness or injury during the past 4 weeks as being 1 SD or more above the mean for the general population (mean = 1.04 SD = 3.60).

^b% reporting number of outpatient visits during past 4 weeks as being 1 SD or more above the mean for the general population (mean = 0.82, SD = 1.57).

^c% reporting number of hospital stays during the past 12 months as being 1 SD or more above the mean for the general population (mean = 0.22, SD = 0.89).

Table 9.4

Percentage of Adults Reporting Pain Interference Problems at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Pain interfered with enjoyment in life very often or always ^a			Pain made simple tasks hard to complete very often or always ^b			Leisure activities affected by pain very often or always ^c			Felt fed up and frustrated by pain very often or always ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	721	58.27	0.8	718	58.27	0.1	719	58.26	0.6	720	58.25	1.0
2	50–54.9	470	52.61	1.1	467	52.62	0.6	467	52.61	1.5	469	52.61	2.6
3	45–49.9	269	47.91	6.3	266	47.90	3.8	268	47.91	4.9	269	47.91	8.9
4	40–44.9	179	42.46	10.6	179	42.46	8.9	178	42.45	11.8	179	42.46	13.4
5	35–39.9	130	37.71	24.6	130	37.71	20.0	130	37.71	20.0	130	37.71	23.9
6	30–34.9	105	32.96	44.8	105	32.96	42.9	103	32.96	45.6	105	32.96	34.3
7	25–29.9	73	27.60	58.9	73	27.60	53.4	73	27.60	56.2	72	27.62	40.3
8	20–24.9	30	22.89	73.3	30	22.89	70.0	30	22.89	76.7	30	22.89	46.7
9	< 20	13	17.76	84.6	13	17.76	92.3	13	17.76	84.6	13	17.76	53.9

^a% reporting that pain interfered with their enjoyment in life very often or always during the past 4 weeks.

^b% reporting that pain made simple tasks hard to complete very often or always during the past 4 weeks.

^c% reporting that leisure activities were affected by pain very often or always during the past 4 weeks.

^d% reporting that pain made them feel fed up and frustrated very often or always during the past 4 weeks.

Table 9.5

Percentage of Adults Reporting Future Health and Work-Related Problems at 7 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Outpatient visits with health professional ^a			Bed days due to illness/injury ^b			Not working because of health ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	115	58.05	23.5	115	58.05	5.2	115	58.05	26.1
2	50–54.9	73	52.72	45.2	72	52.75	11.1	73	52.72	38.4
3	45–49.9	45	47.79	51.1	44	47.80	18.2	45	47.79	42.2
4	40–44.9	20	42.26	45.0	21	42.23	14.3	21	42.23	71.4
5	35–39.9	16	37.72	68.8	16	37.72	25.0	16	37.72	75.0
6	30–34.9	16	32.39	81.3	16	32.39	56.3	16	32.39	81.3
7	< 30	15	23.35	93.3	15	23.35	46.7	15	23.35	80.0

^a% reporting one or more outpatient visits with a health professional during the 4 weeks preceding survey readministration.

^b% reporting one or more days in bed because of illness or injury during the 4 weeks preceding survey readministration.

^c% reporting not working at a paying job because of health at the time of survey readministration.

ratings of job performance variable (Column 2) showed a relatively minor inconsistency in increasing percentages (see Level 8), while additional inconsistency was seen in the disability variable (Column 3). Note that higher percentages of disability were reported at the two highest MCS score levels (Levels 1 and 2) as compared to the same two PCS score levels (see Table 9.2), while the reverse was true at the lower MCS and PCS score levels.

Health, quality of life, pain-related interference, energy level, and MCS. Table 9.9 reveals a pattern of consistently increasing percentages of respondents reporting relatively low health (Column 1) and HRQOL (Column 2) ratings and *little* or *no* energy (Column 5). Note that 90.5% of respondents at the lowest MCS score level (Level 10) rated their overall HRQOL as *fair* or *poor*. Meanwhile, the percentages of those reporting pain-related problems (Columns 3 and 4) generally increased as MCS scores decreased except at the lowest level (Level 10), which reported lower percentages than those reported at Level 9.

Cognitive functioning and MCS. Overall, relative to the other criterion variables, those related to cognitive functioning are reported to be less problematic across the range of MCS scores. As Table 9.10 reveals, the percentages of respondents who reported difficulty in reasoning and problem solving (Column 1), concentration and thinking (Column 2), confusion (Column 3), or forgetfulness (Column 4) did not reach double digits until MCS score Level 7 and below (T -score range < 35). Even at the lowest level, only about half (52.6%) of the respondents reported concentration and thinking problems.

Sleep disturbance and MCS. Table 9.11 shows that reports of various aspects of sleep disturbance generally increased as MCS scores began to fall through to the lower levels, with significant percentages of respondents beginning to report such problems even at MCS score Level 3, which includes the mean T score of 50. The most striking of the sleep disturbance findings were those having to do with the number of minutes needed to fall asleep (Column 1). Even at the highest MCS score level, 36.6% of respondents reported this problem. The percentages quickly increased to almost 60% at Level 3. Significant percentages of other sleep disturbance problems—sleep not being quiet (Column 2), frequency of having trouble falling asleep (Column 3), and awakening and having trouble falling back to sleep (Column 4)—occurred across the 10 MCS score levels; however, no more than 65.2% report any such problems even at the highest level.

Sleep somnolence and MCS. As with the sleep disturbance criterion variables, Table 9.12 reveals that no more than 65.2% of the respondents reported somnolence problems at the highest MCS score level. There

was a near perfect ordering across the 10 score levels in the percentages of respondents who reported feeling drowsy during the day (Column 1) or having trouble staying awake during the day (Column 2). The percentages of those reporting taking naps during the day *most* or *all of the time* (Column 3) also increased through the lower score levels, albeit in a much more inconsistent manner. It is interesting to note that 10.3% reported this nap-taking behavior even at the highest MCS score level.

Sleep quantity and adequacy and headache/shortness of breath and MCS. Table 9.13 presents the findings pertaining to MCS score levels and a variety of other sleep-related problems. Upon examination of these findings, it is immediately apparent that getting more or less than the average number of hours of sleep—7 to 8 hours—was a fairly common problem across the 10 MCS score levels (Column 1). For example, at Level 1, 44.9% of respondents reported nightly average number of hours of sleep falling outside the average range. Although not to the same degree, getting enough sleep to feel rested (Column 2) and getting the needed amount of sleep (Column 4) were also fairly frequently reported across the MCS score levels, with the percentages generally increasing as MCS scores decreased. Unlike the other criterion variables examined in Table 9.13, awakening short of breath or with a headache (Column 3) was not reported as that problematic. To wit, the percentages of those reporting this problem did not reach double digits until MCS score Level 7 (T -score range = 30.0–34.9) or lower, with only 30.4% at the lowest level (Level 10) reporting this occurrence *most* or *all of the time*.

Future mental health problems and MCS. Table 9.14 looks at the relationship between a respondent's baseline MCS score level and the occurrence of health-related events assessed 3 to 4 months later. Generally, respondents scoring at the highest MCS score level (Level 1) at baseline were much less likely to report feeling down/depressed/hopeless (Column 1) or having little interest or pleasure in doing things (Column 2) at reassessment than those scoring at the lowest MCS score level (Level 7).

Criterion-Based Interpretation of the Standard Form Health Domain Scales

Tables 9.15 through 9.28 present the findings from the 2009 normative study regarding reported problems on relevant criterion variables at each of the standard form health domain T -score levels.

Physical Functioning (PF)

Quality of life and performance of work and other activities. As shown in Table 9.15, a perfect or near perfect

Table 9.6

Percentage of Adults Reporting Behavioral Health Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Feeling down/ depressed/hopeless more than half the days or nearly every day ^a			Depression is a current chronic condition ^b			Anxiety is a current chronic condition ^c			Number of occasions having 5+ drinks significantly above the mean ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	488	62.25	0.0	487	62.18	1.2	486	62.16	1.4	217	62.38	11.5
2	55–59.9	1,247	57.39	0.2	1,245	57.39	2.2	1,244	57.39	3.5	624	57.31	11.1
3	50–54.9	885	52.82	0.8	880	52.83	7.2	875	52.82	7.7	471	52.86	13.8
4	45–49.9	505	47.79	4.0	502	47.79	14.7	502	47.79	21.7	251	47.88	18.7
5	40–44.9	334	42.57	13.8	335	42.57	29.3	335	42.57	29.6	170	42.54	19.4
6	35–39.9	216	37.69	28.2	217	37.69	51.2	217	37.69	44.7	100	37.79	15.0
7	30–34.9	129	32.65	40.3	125	32.64	54.4	124	32.63	57.3	54	32.61	27.8
8	25–29.9	102	27.57	70.6	103	27.53	70.9	103	27.53	61.2	39	27.72	23.1
9	20–24.9	54	22.91	90.7	54	22.91	74.1	54	22.91	66.7	20	22.82	35.0
10	< 20	42	15.55	97.6	42	15.55	92.9	42	15.55	83.3	13	16.63	15.4

^a% reporting feeling down/depressed/hopeless more than half the days or nearly every day during the past 2 weeks.

^b% reporting depression as a current chronic condition.

^c% reporting anxiety as a current chronic condition.

^d% reporting the number of occasions of having 5 or more drinks during the past 4 weeks as being 1 SD or more above the general population mean (mean = 1.59, SD = 2.27).

Table 9.7

Percentage of Adults Reporting Negative Effects of Personal, Emotional, and Physical Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Could not do or was limited quite a lot in usual social activities due to physical health/emotional problems ^a			Bothered by emotional problems moderately, quite a lot, or extremely ^b			Happiness/satisfaction with personal life rated as sometimes fairly satisfied or generally dissatisfied ^c			Felt little interest/ pleasure in doing things more than half the days or nearly every day ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	492	62.24	0.8	489	62.27	0.0	492	62.24	1.0	487	62.25	1.4
2	55–59.9	1,248	57.39	0.8	1,246	57.40	0.1	1,249	57.39	2.7	1,247	57.39	2.4
3	50–54.9	884	52.83	2.5	881	52.82	0.2	885	52.82	10.2	884	52.83	3.6
4	45–49.9	508	47.79	4.7	507	47.79	3.6	505	47.79	25.7	506	47.79	7.7
5	40–44.9	336	42.57	10.7	334	42.55	9.6	334	42.56	39.5	331	42.58	23.3
6	35–39.9	219	37.69	17.8	217	37.70	22.1	218	37.69	57.8	217	37.69	32.3
7	30–34.9	128	32.65	25.0	129	32.65	47.3	128	32.64	64.1	126	32.62	44.4
8	25–29.9	103	27.53	55.3	103	27.53	72.8	103	27.53	80.6	102	27.57	59.8
9	20–24.9	54	22.91	72.2	54	22.91	87.0	54	22.91	90.7	54	22.91	79.6
10	< 20	42	15.55	92.9	42	15.55	100.0	42	15.55	90.5	42	15.55	97.6

^a% reporting they could not do or were limited quite a lot in usual social activities due to physical health/emotional problems during the past 4 weeks.

^b% reporting being bothered by emotional problems moderately, quite a lot, or extremely during the past 4 weeks.

^c% reporting happiness/satisfaction with personal life rated as sometimes fairly satisfied or generally dissatisfied during the past 4 weeks.

^d% reporting taking little interest/pleasure in doing things more than half the days or nearly every day during the past 2 weeks.

ordering of the percentages of respondents reporting fair or poor HRQOL (Column 1), limitations in performing usual activities due to physical limitations (Column 2), a significant number of bed days due to illness or injury (Column 3), and being disabled (Column 4) was apparent across the 9 PF score levels. Of interest is the fact that among those scoring at the lowest score level (Level 9, T-score range < 20), only 45.5% reported a significant number of bed days and only 43.8% reported employ-

ment status as disabled; however, 81.8% indicated that they could not do or had quite a lot of difficulty doing usual activities due to physical conditions at the same score level.

Role-Physical (RP)

Work performance, illness/injury, and limitations due to pain and physical conditions. Table 9.16 reveals a perfect or near perfect ordering of the increasing

Table 9.8

Percentage of Adults Reporting Job Performance Problems and the Effects of Stress at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores Level	Could not do or were kept quite a lot from doing usual work/school/other activities ^a			Rating of overall job performance significantly below the mean ^b			Employment status is disabled ^c			Experienced a good bit, quite a bit, or a great deal of stress/pressure in daily living ^d			Stress/pressure has affected health moderately, quite a lot, or extremely ^e		
	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	491	62.24	0.8	179	61.89	1.7	490	62.24	6.7	490	62.25	8.6	490	62.25	1.6
2	1,244	57.40	0.5	705	57.33	2.8	1,250	57.39	3.0	1,246	57.39	14.7	1,248	57.38	3.3
3	882	52.83	0.5	521	52.90	6.3	882	52.82	4.9	885	52.82	32.2	883	52.82	9.6
4	507	47.79	3.0	298	47.82	10.7	508	47.79	7.3	506	47.79	44.9	507	47.79	21.5
5	334	42.57	6.0	160	42.76	18.8	334	42.56	12.3	334	42.56	57.8	337	42.56	37.1
6	219	37.69	14.6	106	37.95	32.1	219	37.69	16.0	218	37.69	71.1	219	37.69	49.8
7	128	32.64	27.3	50	32.85	40.0	128	32.64	24.2	125	32.62	84.8	126	32.64	65.1
8	103	27.53	44.7	36	27.56	30.6	103	27.53	35.9	102	27.57	86.3	103	27.53	78.6
9	53	22.92	56.6	14	23.14	42.9	54	22.91	40.7	54	22.91	94.4	54	22.91	90.7
10	42	15.55	73.8	8	17.36	75.0	42	15.55	35.7	42	15.55	97.6	42	15.55	95.2

^a% reporting they could not do or were kept quite a lot from doing usual work/school/other activities during the past 4 weeks due to personal/emotional/physical problems.

^b% rating overall job performance as being 1 SD or more below the general population mean 0–10 rating during the past 4 weeks (mean = 8.37, SD = 1.53).

^c% reporting employment status as disabled.

^d% reporting having experienced a good bit, quite a bit, or a great deal of stress/pressure in daily living during the past 4 weeks.

^e% reporting stress/pressure has affected health moderately, quite a lot, or extremely during the past 4 weeks.

Table 9.9

Percentage of Adults Reporting Problems in Health, Quality of Life, Pain-Related Interference, and Energy Level at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores Level	Health rated significantly below the mean ^a (1)		Overall quality of life rated as <i>fair</i> or <i>poor</i> ^b (2)		Pain interfered with enjoyment of life <i>very often</i> or <i>always</i> ^c (3)		Pain made one feel fed up and frustrated <i>very often</i> or <i>always</i> ^d (4)		Little or no energy ^e (5)		
	n	Mean	%	n	Mean	%	n	Mean	n	Mean	%
1	484	62.24	3.3	492	62.24	3.3	227	62.09	492	62.24	3.3
2	1,247	57.39	3.9	1,251	57.39	2.7	562	57.29	1,247	57.39	2.7
3	882	52.82	6.9	885	52.82	6.7	442	52.78	882	52.82	8.1
4	505	47.79	13.3	508	47.79	11.8	286	47.55	504	47.79	15.3
5	331	42.58	23.3	334	42.57	24.9	181	42.43	336	42.57	25.6
6	219	37.69	28.3	219	37.69	38.4	120	37.75	219	37.69	37.9
7	127	32.69	34.7	128	32.65	48.4	74	32.80	128	32.65	53.9
8	102	27.57	50.0	103	27.53	68.0	56	27.42	103	27.53	68.9
9	53	22.86	64.2	54	22.91	68.5	23	23.13	54	22.91	70.4
10	41	15.94	65.9	42	15.55	90.5	19	15.08	42	15.55	81.0

^a% rating health as being 1 SD or more below the general population mean 0–100 rating during the past 4 weeks (mean = 78.34, SD = 19.31).

^b% rating overall quality of life as *fair* or *poor*.

^c% reporting pain interfered with enjoyment of life *very often* or *always* during the past 4 weeks.

^d% reporting pain made one feel fed up and frustrated *very often* or *always* during the past 4 weeks.

^e% reporting amount of energy as *a little* or *none* during the past 4 weeks.

Table 9.10

Percentage of Adults Reporting Problems in Cognitive Functioning at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Difficulty reasoning and solving problems most or all of the time ^a			Difficulty doing activities involving concentration and thinking most or all of the time ^b			Becomes confused and starts several actions at a time most or all of the time ^c			Forgets things that recently happened most or all of the time ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	227	62.09	4.0	227	62.09	1.8	225	62.11	0.9	226	62.10	0.4
2	55–59.9	561	57.29	4.1	560	57.31	2.9	560	57.30	1.3	563	57.29	1.2
3	50–54.9	439	52.78	3.2	441	52.79	1.6	439	52.78	1.1	442	52.78	2.9
4	45–49.9	286	47.55	3.5	283	47.55	3.2	285	47.55	1.8	285	47.55	4.6
5	40–44.9	181	42.43	5.0	181	42.43	6.6	181	42.43	3.3	181	42.43	5.5
6	35–39.9	120	37.75	5.8	120	37.75	2.5	120	37.75	4.2	120	37.75	9.2
7	30–34.9	73	32.70	15.1	74	32.80	16.2	74	32.80	12.2	74	32.80	20.3
8	25–29.9	56	27.42	21.4	56	27.42	23.2	56	27.42	14.3	56	27.42	26.8
9	20–24.9	23	23.13	30.4	23	23.13	26.1	23	23.13	17.4	23	23.13	30.4
10	< 20	19	15.08	47.4	19	15.08	52.6	19	15.08	31.6	19	15.08	36.8

^a% reporting difficulty in reasoning and solving problems most or all of the time during the past 4 weeks.

^b% reporting difficulty in doing activities involving concentration and thinking most or all of the time during the past 4 weeks.

^c% reporting becoming confused and starting several actions at a time most or all of the time during the past 4 weeks.

^d% reporting forgetting things that recently happened most or all of the time during the past 4 weeks.

Table 9.11

Percentage of Adults Reporting Sleep Disturbance Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Number of minutes to fall asleep significantly above the mode ^a			Sleep not quiet most or all of the time ^b			Trouble falling asleep most or all of the time ^c			Awakened during sleep and trouble falling back to sleep most or all of the time ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	265	62.37	36.6	264	62.37	9.5	264	62.38	5.3	263	62.41	5.3
2	55–59.9	686	57.47	45.2	687	57.47	7.1	686	57.47	5.7	686	57.47	4.8
3	50–54.9	444	52.86	59.2	444	52.86	15.1	444	52.86	10.8	441	52.86	8.2
4	45–49.9	223	48.07	63.7	222	48.07	20.7	222	48.07	17.6	222	48.07	17.6
5	40–44.9	155	42.71	75.5	154	42.71	25.3	155	42.71	22.6	154	42.72	18.2
6	35–39.9	99	37.59	71.7	99	37.59	39.4	99	37.59	27.3	99	37.59	20.2
7	30–34.9	55	32.46	76.4	54	32.44	48.2	55	32.46	45.5	55	32.46	23.6
8	25–29.9	46	27.60	93.5	46	27.73	50.0	47	27.65	55.3	47	27.65	40.4
9	20–24.9	31	22.77	87.1	31	22.77	64.5	31	22.77	58.1	30	22.81	40.0
10	< 20	23	15.92	91.3	23	15.92	52.2	23	15.92	65.2	23	15.92	43.5

^a% reporting the number of minutes to fall asleep as being significantly above the mode (≥ 16 minutes) during the past 4 weeks.

^b% reporting sleep not being quiet most or all of the time during the past 4 weeks.

^c% reporting having trouble falling asleep most or all of the time during the past 4 weeks.

^d% reporting being awakened during sleep and having trouble falling back to sleep most or all of the time during the past 4 weeks.

percentages of respondents who reported a significant number of bed days due to illness or injury (Column 1), being disabled (Column 2), limitations in performing usual activities due to physical limitations (Column 3), pain frequently making simple tasks hard to complete (Column 4), and limitations in usual activities or enjoyment (Column 5) across the 8 RP score levels. The most interesting finding amongst this group of criterion variables was that only 53.6% of those scoring at the lowest

RP score level (Level 8, *T*-score range < 25) reported work status as being *disabled*.

Bodily Pain (BP)

Bodily pain and its effects. Table 9.17 shows the perfect or near perfect ordering of the increasing percentages of respondents reporting the pain-related limitations or effects assessed by each of the five criterion variables across the 9 levels of BP *T* scores. Such limitations and

Table 9.12

Percentage of Adults Reporting Sleep Somnolence Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Felt drowsy/sleepy during the day most or all of the time ^a			Trouble staying awake during the day most or all of the time ^b			Takes naps during the day most or all of the time ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	265	62.37	3.4	265	62.37	1.1	263	62.38	10.3
2	55–59.9	688	57.47	5.1	685	57.47	2.2	686	57.47	12.0
3	50–54.9	442	52.86	14.7	443	52.86	3.8	441	52.86	13.6
4	45–49.9	222	48.07	18.9	222	48.07	6.8	221	48.06	12.7
5	40–44.9	155	42.71	30.3	154	42.70	9.7	155	42.71	23.9
6	35–39.9	99	37.59	46.5	98	37.55	16.3	99	37.59	17.2
7	30–34.9	54	32.45	53.7	55	32.46	23.6	55	32.46	29.1
8	25–29.9	47	27.65	57.5	47	27.65	25.5	47	27.65	44.7
9	20–24.9	31	22.77	74.2	31	22.77	41.9	31	22.77	35.5
10	< 20	23	15.92	65.2	23	15.92	47.8	23	15.92	39.1

^a% reporting having felt drowsy/sleepy during the day most or all of the time during the past 4 weeks.

^b% reporting having trouble staying awake during the day most or all of the time during the past 4 weeks.

^c% reporting having to take naps during the day most or all of the time during the past 4 weeks.

Table 9.13

Percentage of Adults Reporting Problems With Sleep Quantity and Adequacy and Headaches or Shortness of Breath at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		More or less than average number of hours of sleep each night ^a			Got enough sleep to feel rested little or none of the time ^b			Awakened short of breath or with a headache most or all of the time ^c			Getting the needed amount of sleep little or none of the time ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	265	62.37	44.9	264	62.38	8.7	263	62.32	1.1	264	62.38	9.9
2	55–59.9	687	57.47	42.9	687	57.47	10.3	686	57.47	0.9	687	57.47	12.1
3	50–54.9	442	52.86	48.2	444	52.86	19.8	443	52.87	0.7	443	52.87	24.6
4	45–49.9	221	48.03	56.6	221	48.07	31.7	222	48.07	0.9	221	48.07	34.4
5	40–44.9	155	42.71	64.5	155	42.71	40.0	153	42.69	3.9	155	42.71	43.9
6	35–39.9	99	37.59	65.7	99	37.59	50.5	99	37.59	9.1	99	37.59	45.5
7	30–34.9	55	32.46	65.5	55	32.46	58.2	54	32.44	14.8	55	32.46	52.7
8	25–29.9	47	27.65	78.7	46	27.73	65.2	47	27.65	10.6	47	27.65	66.0
9	20–24.9	31	22.77	83.9	31	22.77	61.3	31	22.77	29.0	30	22.81	73.3
10	< 20	24	15.96	83.3	23	15.92	60.9	23	15.92	30.4	23	15.92	69.6

^a% reporting more or less than average number of hours of sleep each night (7–8 hours) during the past 4 weeks.

^b% reporting getting enough sleep to feel rested little or none of the time during the past 4 weeks.

^c% reporting awakening short of breath or with a headache most or all of the time during the past 4 weeks.

^d% reporting getting the needed amount of sleep little or none of the time during the past 4 weeks.

effects were minimal or not present in respondents at the higher BP score levels (Levels 1–3), while nearly half or more of those at the lower score levels (Levels 7–9) reported their presence.

Chronic conditions, treatment, and disability. Table 9.18 also demonstrates a perfect or near perfect ordering of the increasing percentages of respondents who reported the health-related problems or events assessed by the five criterion variables across the 9 BP T-score levels. Among the more notable findings, 9.0% of those who scored at

the highest BP score level (Level 1) reported having one or more chronic conditions that limited usual activities or enjoyment *moderately, quite a lot, or extremely* (Column 2). The percentage jumped to 17.7% at Level 3, which includes the BP mean score, and then more than tripled to reach 61.7% at Level 6. Also, among those scoring at the lowest BP level at baseline, only 19.1% reported significantly more hospitalizations (Column 4) and 51.2% reported significantly more outpatient visits (Column 3) than the general population averages for these variables.

Table 9.14

Percentage of Adults Reporting Future Mental Health Problems at 7 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Down/depressed/ hopeless several, more than half, or nearly every day ^a			Little interest/ pleasure in doing things several, more than half, or nearly every day ^b		
Level	Range	(1)		(2)			
		<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	60+	38	62.29	0.0	38	62.29	2.6
2	55–59.9	101	57.45	8.9	101	57.45	12.9
3	50–54.9	55	52.94	30.9	55	52.94	30.9
4	45–49.9	38	47.38	55.3	39	47.39	48.7
5	40–44.9	31	43.00	67.7	31	43.00	58.1
6	35–39.9	19	37.66	68.4	18	37.63	61.1
7	< 35	16	28.00	93.8	17	28.02	88.2

^a% reporting felt down, depressed, or hopeless several, more than half, or nearly every day during the 2 weeks preceding survey readministration.

^b% reporting experiencing little interest or pleasure in doing things several, more than half, or nearly every day during the 2 weeks preceding survey readministration.

General Health (GH)

Quality of life, general health, and disability.

As shown in Table 9.19, there is again a perfect or near perfect ordering of the increasing percentages of respondents reporting low health and HRQOL ratings and work-related disability across the 11 GH *T*-score levels. Notable is the fact that significant percentages of these problems did not appear until GH score Level 5 (*T*-score range = 45.0–49.9), representing the lower

half of the range of scores considered average when interpreting an individual respondent's SF-36v2 results. Also notable are the differences in percentages of reports of poor health based on a multiple-choice format (Column 2) and those using a Likert scale ranging from 0 to 100 (Column 3). In general, when answering the multiple-choice question in which the problematic health criterion was health rated as *poor* or *very poor*, a significant percentage reporting this problem (10.3%) did not appear until GH score Level 7, but quickly reached 100% at the lowest level (Level 11). Meanwhile, when using the 0–100 rating scale with a criterion of 1 *SD* or more below the population mean, a comparable percentage (11.9%) was identified at a higher GH score level (Level 5), but only reached 80.0% at the lowest score level.

Chronic conditions and treatment. Examining the associations of GH scores with number of recent outpatient visits (Column 1), number of chronic conditions (Column 2), and limitations imposed by chronic conditions (Column 3) in Table 9.20, one notes a general ordering of increasing percentages of those reporting problems with decreasing levels of GH scores. Regarding number of outpatient visits, the increase in the percentage of those reporting more visits than the average for the general population across the GH score levels is slow but steady up to Level 10 (38.7%). At Level 11 ($T < 20$), the percentage then doubles to 77.8%.

Vitality (VT)

Quality of life and energy level. Table 9.21 demonstrates a perfect ordering of increasing percentages

Table 9.15

Percentage of Adults Reporting Problems Related to Quality of Life and the Performance of Work and Other Activities at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Physical Functioning Scores, 2009 U.S. General Population

PF T Scores		Overall quality of life rated as fair or poor ^a			Could not do or had quite a lot of difficulty doing usual activities due to physical conditions ^b			Days in bed due to illness/injury significantly above the mean ^c			Current employment status is disabled ^d		
Level	Range	(1)		(2)		(3)		(4)		(4)			
		<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	55+	1,839	56.42	2.8	1,839	56.42	0.8	1,834	56.42	0.9	1,839	56.42	1.0
2	50–54.9	718	51.92	7.0	715	51.92	1.0	717	51.92	1.5	714	51.92	2.1
3	45–49.9	456	46.98	13.8	455	46.98	2.2	453	46.98	4.4	456	46.98	5.7
4	40–44.9	322	40.06	21.4	322	40.06	14.3	320	40.05	9.4	322	40.06	14.0
5	35–39.9	177	35.17	33.9	177	35.17	29.4	176	35.17	9.7	177	35.17	23.7
6	30–34.9	219	29.63	38.8	218	29.61	54.6	215	29.63	13.5	219	29.63	30.1
7	25–29.9	167	23.43	55.1	166	23.43	78.3	164	23.42	33.5	167	23.43	34.7
8	20–24.9	89	18.16	60.7	89	18.16	87.6	87	18.18	34.5	88	18.13	54.6
9	< 20	32	14.95	62.5	33	14.95	81.8	33	14.95	45.5	32	14.95	43.8

^a% rating overall quality of life as fair or poor.

^b% reporting could not do or were limited quite a lot in usual activities during the past 4 weeks.

^c% reporting number of days in bed due to illness or injury during the past 4 weeks as being 1 *SD* or more above the mean for the general population (mean = 1.04, *SD* = 3.60).

^d% reporting current work status as disabled.

Table 9.16

Percentage of Adults Reporting Work Performance Problems, Significant Illness or Injury, and Limitations Due to Pain and Physical Conditions at 8 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Physical Scores, 2009 U.S. General Population

Level	RP T Scores		Days in bed due to illness/injury significantly above the mean ^a			Current employment status is disabled ^b			Could not do or had quite a lot of difficulty doing usual activities due to physical conditions ^c			Pain made simple tasks hard to complete very often or always ^d			Chronic condition(s) limit usual activities/employment moderately, quite a lot, or extremely ^e		
	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	
1	55+	1,848	57.16	0.4	1,849	57.16	0.8	1,853	57.16	0.7	875	57.16	0.9	1,861	57.16	8.7	
2	50-54.9	751	52.86	1.2	752	52.86	1.2	752	52.86	1.7	394	53.14	0.5	754	52.86	22.4	
3	45-49.9	397	47.42	4.8	398	47.42	8.5	397	47.42	5.3	206	47.40	4.9	399	47.43	17.8	
4	40-44.9	230	42.58	8.3	232	42.58	9.5	231	42.58	16.5	112	42.45	10.7	232	42.58	52.2	
5	35-39.9	274	38.35	8.4	274	38.35	19.0	275	38.34	19.6	137	38.29	12.4	277	38.35	56.0	
6	30-34.9	260	32.13	21.2	263	32.14	29.3	264	32.12	53.0	143	32.21	37.1	264	32.12	73.5	
7	25-29.9	98	26.80	29.6	100	26.83	48.0	100	26.83	78.0	43	26.50	55.8	100	26.83	86.0	
8	< 25	137	21.82	44.5	140	21.86	53.6	139	21.86	90.7	69	21.80	68.1	140	21.86	90.7	

^a% reporting number of days in bed due to illness or injury during the past 4 weeks as being 1 SD or more above the mean for the general population (mean = 1.04, SD = 3.60).
^b% reporting current work status as disabled.

^c% reporting could not do or were limited quite a lot in usual activities during the past 4 weeks.

^d% reporting that pain made simple tasks hard to complete very often or always during the past 4 weeks.

^e% reporting one or more chronic condition(s) now has that limit usual activities/employment moderately, quite a lot, or extremely.

Table 9.17

Percentage of Adults Reporting Problems With Bodily Pain and Its Effects at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scores, 2009 U.S. General Population

Level	BP T Scores		Bodily pain rated moderate, severe, or very severe ^a			Pain interfered with enjoyment in life very often or always ^b			Pain made simple tasks hard to complete very often or always ^c			Leisure activities affected by pain very often or always ^d			Felt fed up and frustrated by pain very often or always ^e		
	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	
1	60+	809	62.00	0.0	365	62.00	1.1	364	62.00	0.6	363	62.00	1.1	363	62.00	0.8	
2	55-59.9	965	55.55	0.0	456	55.55	0.4	455	55.55	0.0	455	55.55	0.2	457	55.55	0.2	
3	50-54.9	616	51.28	0.0	333	51.24	0.3	331	51.24	0.3	332	51.24	0.6	332	51.24	1.5	
4	45-49.9	522	46.68	15.3	265	46.67	1.5	263	46.67	0.8	263	46.67	1.9	265	46.67	2.6	
5	40-44.9	332	42.36	75.3	187	42.36	6.4	185	42.36	3.8	184	42.36	7.1	187	42.36	7.5	
6	35-39.9	350	38.25	99.1	175	38.24	18.9	175	38.24	13.7	175	38.24	17.1	175	38.24	23.4	
7	30-34.9	312	32.46	100.0	148	32.83	60.1	147	32.77	55.1	148	32.83	56.8	147	32.82	46.9	
8	25-29.9	72	26.42	100.0	44	26.47	95.5	44	26.47	90.9	44	26.47	86.4	44	26.47	68.2	
9	< 25	45	21.68	100.0	17	21.68	94.1	17	21.68	94.1	17	21.68	94.1	17	21.68	82.4	

^a% reporting bodily pain as being moderate, severe, or very severe during the past 4 weeks.

^b% reporting that pain interfered with enjoyment in life very often or always during the past 4 weeks.

^c% reporting that pain made simple tasks hard to complete very often or always during the past 4 weeks.

^d% reporting that leisure activities were affected by pain very often or always during the past 4 weeks.

^e% reporting that pain made them feel fed up and frustrated very often or always during the past 4 weeks.

Table 9.18

Percentage of Adults Reporting Significant Chronic Conditions, Treatment, and Disability at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Bodily Pain Scores, 2009 U.S. General Population

BP T Scores Level	Range	Number of chronic conditions significantly above the mean ^a			Chronic condition(s) limit usual activities/employment moderately, quite a lot, or extremely ^b			Number of outpatient visits significantly above the mean ^c			Number of hospital stays significantly above the mean ^d			Current employment status is disabled ^e		
		n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	811	62.00	2.2	811	62.00	9.0	805	62.00	3.7	804	62.00	0.8	808	62.00	2.2
2	55-59.9	965	55.55	2.9	965	55.55	8.5	960	55.55	3.7	962	55.55	0.8	960	55.55	0.6
3	50-54.9	617	51.28	6.8	617	51.28	17.7	616	51.28	7.0	616	51.29	0.7	617	51.28	2.6
4	45-49.9	522	46.68	14.0	522	46.68	30.8	519	46.68	9.1	520	46.68	1.4	521	46.68	5.6
5	40-44.9	333	42.35	18.3	333	42.35	42.3	330	42.35	12.1	330	42.35	2.7	332	42.35	9.3
6	35-39.9	350	38.25	28.6	350	38.25	61.7	342	38.25	21.4	341	38.25	4.7	345	38.25	16.8
7	30-34.9	312	32.46	48.1	312	32.46	88.8	309	32.46	32.4	307	32.45	6.8	312	32.46	36.9
8	25-29.9	72	26.42	61.1	72	26.42	87.5	70	26.43	41.4	69	26.44	8.7	72	26.42	47.2
9	< 25	44	21.68	64.4	45	21.68	93.3	43	21.68	51.2	42	21.68	19.1	45	21.68	55.6

^a% reporting the number of chronic conditions now has as being 1 SD or more above the mean for the general population (mean = 2.10, SD = 2.20).

^b% reporting one or more chronic condition(s) now has that limit usual activities/employment moderately, quite a lot, or extremely.

^c% reporting number of outpatient visits during past 4 weeks as being 1 SD or more above the mean for the general population (mean = 0.82, SD = 1.57).

^d% reporting number of hospital stays during the past 12 months as being 1 SD or more above the mean for the general population (mean = 0.22, SD = 0.89).

^e% reporting current work status as disabled.

Table 9.19

Percentage of Adults Reporting Quality of Life and General Health Problems and Disability at 11 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scores, 2009 U.S. General Population

GH T Scores		Overall quality of life rated as fair or poor ^a (1)			Health rated as poor or very poor ^b (2)			Health rating significantly below the mean ^c (3)			Current employment status is disabled ^d (4)		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	65+	257	65.95	0.4	258	65.94	0.0	255	65.94	1.9	258	65.94	0.0
2	60–64.9	537	61.63	0.7	537	61.63	0.2	537	61.63	0.7	535	61.63	2.4
3	55–59.9	824	56.82	1.5	821	56.83	0.1	822	56.83	1.8	824	56.82	2.1
4	50–54.9	743	52.05	3.4	745	52.05	0.3	742	52.05	2.7	743	52.05	1.8
5	45–49.9	542	47.34	10.2	543	47.34	1.3	531	47.36	11.9	541	47.34	5.0
6	40–44.9	456	42.32	22.4	456	42.31	4.0	451	42.30	23.2	454	42.31	15.9
7	35–39.9	330	37.61	35.2	330	37.61	10.3	328	37.60	34.9	330	37.61	16.7
8	30–34.9	183	32.32	63.4	182	32.32	37.4	180	32.31	51.6	182	32.29	33.5
9	25–29.9	109	27.36	71.6	110	27.35	59.1	109	27.36	65.5	109	27.36	49.5
10	20–24.9	31	22.38	87.1	31	22.38	77.4	31	22.38	80.7	31	22.38	48.4
11	< 20	9	18.95	100.0	9	18.95	100.0	9	18.95	80.0	9	18.95	55.6

^a% rating overall quality of life as fair or poor.

^b% rating health as poor or very poor during the past 4 weeks.

^c% rating health 1 SD or more below the general population mean 0–100 rating during the past 4 weeks (mean = 78.34, SD = 19.31).

^d% reporting current work status as disabled.

Table 9.20

Percentage of Adults Reporting Significant Chronic Conditions and Treatment at 11 Levels of SF-36v2 Standard (4-Week Recall) Form General Health Scores, 2009 U.S. General Population

GH T Scores		Number of outpatient visits significantly above the mean ^a (1)			Number of chronic conditions significantly above the mean ^b (2)			Chronic condition(s) limit usual activities/enjoyment moderately, quite a lot, or extremely ^c (3)		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	65+	257	65.94	2.3	258	65.94	1.2	258	65.94	6.2
2	60–64.9	534	61.62	4.7	537	61.63	2.2	537	61.63	7.5
3	55–59.9	820	56.82	6.0	827	56.83	3.0	827	56.83	14.2
4	50–54.9	743	52.05	6.7	745	52.05	8.1	745	52.05	19.3
5	45–49.9	532	47.35	11.8	547	47.34	13.2	547	47.34	30.5
6	40–44.9	451	42.31	13.5	457	42.32	21.7	457	42.32	46.4
7	35–39.9	329	37.60	18.8	330	37.61	31.2	330	37.61	59.4
8	30–34.9	183	32.32	27.3	184	32.32	46.7	184	32.32	76.1
9	25–29.9	108	27.36	31.5	110	27.35	57.3	110	27.35	86.4
10	20–24.9	31	22.38	38.7	31	22.38	58.1	31	22.38	96.8
11	< 20	9	18.95	77.8	10	18.95	50.0	10	18.95	90.0

^a% reporting number of outpatient visits during past 4 weeks as being 1 SD or more above the mean for the general population (mean = 0.82, SD = 1.57).

^b% reporting the number of chronic conditions now has as being 1 SD or more above the mean for the general population (mean = 2.10, SD = 2.20).

^c% reporting one or more chronic condition(s) now has that limit usual activities/enjoyment moderately, quite a lot, or extremely.

of those who rated their HRQOL as fair or poor across the 10 VT score levels (Column 1). A similar but less perfect ordering was found for those reporting little or no energy (Column 2).

Social Functioning (SF)

Quality of life and limitations in social activities. As shown in Table 9.22, there is a perfect ordering across the 9 SF score levels of the increasing percentages of those rating their HRQOL as fair or poor (Column 1) as well as those that reported they could not do or were limited

quite a lot in usual social activities due to physical or emotional health problems (Column 2). Of interest, the percentage of those reporting limitations in social activities more than doubled from SF score Level 6 (37.6%) to score Level 7 (80.5%).

Role-Emotional (RE)

Quality of life, happiness, stress, and emotional problems. There is a perfect or near perfect ordering of increasing percentages of respondents who reported emotional and HRQOL problems with decreasing scores

Table 9.21

Percentage of Adults Reporting Quality of Life and Level of Energy Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Vitality Scores, 2009 U.S. General Population

VT T Scores		Overall quality of life rated as fair or poor ^a (1)			Little or no energy ^b (2)		
Level	Range	n	Mean	%	n	Mean	%
1	65+	167	68.84	0.0	167	68.84	1.2
2	60–64.9	443	62.56	0.5	441	62.56	0.0
3	55–59.9	995	57.25	1.2	991	57.25	0.4
4	50–54.9	463	52.59	3.2	463	52.59	1.7
5	45–49.9	856	48.13	10.2	856	48.12	5.3
6	40–44.9	489	42.34	20.5	489	42.34	23.1
7	35–39.9	172	37.75	34.3	173	37.75	48.6
8	30–34.9	263	33.39	54.4	261	33.39	65.5
9	25–29.9	127	27.41	73.2	126	27.40	90.5
10	< 25	44	22.89	75.0	43	22.89	90.7

^a% rating overall quality of life as fair or poor.

^b% reporting amount of energy as a little or none during the past 4 weeks.

Table 9.22

Percentage of Adults Reporting Quality of Life Problems and Limitations in Social Activities at 9 Levels of SF-36v2 Standard (4-Week Recall) Form Social Functioning Scores, 2009 U.S. General Population

SF T Scores		Overall quality of life rated as fair or poor ^a (1)			Could not do or was limited quite a lot in usual social activities due to physical health/emotional problems ^b (2)		
Level	Range	n	Mean	%	n	Mean	%
1	55+	2,184	57.34	1.6	2,180	57.34	0.2
2	50–54.9	434	52.33	5.1	434	52.33	0.7
3	45–49.9	484	47.31	14.7	484	47.31	1.2
4	40–44.9	267	42.30	27.0	267	42.30	4.9
5	35–39.9	296	37.29	35.1	297	37.29	12.1
6	30–34.9	125	32.27	55.2	125	32.27	37.6
7	25–29.9	118	27.26	73.7	118	27.26	80.5
8	20–24.9	63	22.25	74.6	63	22.25	88.9
9	< 20	48	17.23	77.1	48	17.23	89.6

^a% rating overall quality of life as fair or poor.

^b% reporting they could not do or were limited quite a lot in usual social activities due to physical health/emotional problems during the past 4 weeks.

over the 10 RE score levels, as indicated in Table 9.23. When comparing the lower and higher score levels, a greater percentage of respondents reported their HRQOL as fair or poor (Column 1), being only *sometimes fairly satisfied* or *generally dissatisfied* with their personal life (Column 2), experiencing a *good bit*, *quite a bit*, or a *great deal* of stress or pressure in their daily living (Column 3), and that they *could not do* or were limited *quite a lot* in usual social activities due to personal or emotional health

problems (Column 4) at the lower score levels. Notable is the pervasiveness of reports of significant stress in daily living, regardless of the RE score level. Even at the two highest levels (Levels 1 and 2), a significant percentage of respondents (23.0% and 33.5%, respectively) reported experiencing this level of stress. Also note that at the lowest score level, no more than 77.8% of the respondents reported experiencing any of these problems.

Cognitive functioning and health problems. Although Table 9.24 shows that increasing reports of difficulty in doing activities involving concentration and thinking (Column 1) were associated with decreasing RE score levels, this association was not significant until score Level 6 (*T*-score range = 30.0–34.9), at which 16.7% reported this problem. Even though this percentage more than doubled at the lowest RE score level, it was still relatively low (39.4%). Reports of one or more chronic conditions that limit usual activities or enjoyment *moderately*, *quite a lot*, or *extremely* (Column 3) were significant at all score levels, with respondents experiencing this problem reaching 15.2% at the highest level and increasing almost sixfold to 85.5% by the lowest level. The percentages of those rating their health as significantly below the general population mean (Column 2) progressively increased, from 4.7% at RE score Level 1 to 75.0% at Level 10.

Mental Health (MH)

Depression and anxiety. The clear association found between decreasing scores on the MH scale and increasing reports of psychological disorders and symptoms is shown in Table 9.25. Also revealed is how commonly these disorders and symptoms occur, even at the highest score levels. As expected, there is a perfect linear relationship between decreasing MH score levels and reports of the presence of both depression (Column 1) and anxiety (Column 4) as chronic conditions. Moreover, a near perfect ordering of increasing percentages of respondents who reported feeling down/depressed/hopeless *more than half the days* or *nearly every day* (Column 2) or taking little interest/pleasure in doing things *more than half the days* or *nearly every day* (Column 3) was evident. It is particularly important to note that even respondents experiencing average or high levels of mental health (Levels 1–4) reported frequently feeling down, depressed, or hopeless. In fact, nearly half (45.9%) of those scoring in the lower half of the average MH score range (Level 4, *T*-score range = 45.0–49.9) reported feeling this way at least half of the time.

Quality of life, emotional problems, happiness, and stress. Table 9.26 presents data that further demonstrate the connection between MH scores and emotional health

Table 9.23

Percentage of Adults Reporting Emotional Problems and Problems Related to Quality of Life, Happiness, and Stress at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scores, 2009 U.S. General Population

RE T Scores		Overall quality of life rated as <i>fair</i> or <i>poor</i> ^a			Happiness/satisfaction with personal life rated as <i>sometimes fairly satisfied</i> or <i>generally dissatisfied</i> ^b			Experienced a <i>good bit</i> , <i>quite a bit</i> , or a <i>great deal</i> of stress/pressure in daily living ^c			Could not do or was limited <i>quite a lot</i> in usual work, school, or other daily activities due to personal/emotional problems ^d		
Level	Range	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	55+	2,454	56.17	4.2	2,451	56.17	7.5	2,447	56.17	23.0	2,448	56.17	0.8
2	50–54.9	246	52.66	7.7	245	52.66	20.0	245	52.66	33.5	245	52.66	1.2
3	45–49.9	592	47.41	16.6	590	47.41	26.6	591	47.41	45.4	588	47.42	3.7
4	40–44.9	134	42.24	31.3	134	42.24	37.3	133	42.24	51.1	134	42.24	6.7
5	35–39.9	310	36.41	34.5	310	36.41	47.4	307	36.42	63.5	310	36.41	13.2
6	30–34.9	60	31.78	43.3	59	31.78	59.3	60	31.78	66.7	59	31.78	28.8
7	25–29.9	55	28.31	54.6	55	28.31	60.0	55	28.31	72.7	54	28.31	35.2
8	20–24.9	97	23.88	70.1	97	23.88	66.0	97	23.88	67.0	97	23.88	46.4
9	15–19.9	14	17.87	64.3	14	17.87	71.4	14	17.87	71.4	14	17.87	64.3
10	< 15	54	14.39	77.8	55	14.39	76.4	53	14.39	77.4	55	14.39	70.9

^a% rating overall quality of life as *fair* or *poor*.

^b% rating happiness/satisfaction with personal life as *sometimes fairly satisfied* or *generally dissatisfied* during the past 4 weeks.

^c% reporting having experienced a *good bit*, *quite a bit*, or a *great deal* of stress/pressure in daily living during the past 4 weeks.

^d% reporting they *could not do* or were limited *quite a lot* in usual work, school, or other daily activities due to personal/emotional problems.

Table 9.24

Percentage of Adults Reporting Cognitive Functioning and Health Problems at 10 Levels of SF-36v2 Standard (4-Week Recall) Form Role-Emotional Scores, 2009 U.S. General Population

RE T Scores		Difficulty doing activities involving concentration and thinking <i>most or all of the time</i> ^a			Health rating significantly below the mean ^b			Chronic condition(s) limit usual activities/enjoyment <i>moderately</i> , <i>quite a lot</i> , or <i>extremely</i> ^c		
Level	Range	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	55+	1,130	56.17	2.2	2,438	56.17	4.7	2,461	56.17	15.2
2	50–54.9	150	52.66	5.3	245	52.66	7.8	246	52.66	23.6
3	45–49.9	301	47.32	2.7	589	47.41	14.6	592	47.41	37.8
4	40–44.9	63	42.23	6.4	134	42.24	21.6	134	42.24	59.7
5	35–39.9	168	36.30	6.6	309	36.41	26.9	311	36.40	63.7
6	30–34.9	36	31.77	16.7	58	31.78	46.6	60	31.78	76.7
7	25–29.9	34	28.31	17.7	55	28.31	58.2	55	28.31	87.3
8	20–24.9	60	24.05	16.7	97	23.88	55.7	97	23.88	79.4
9	15–19.9	8	17.87	12.5	14	17.87	50.0	15	17.87	80.0
10	< 15	33	14.39	39.4	52	14.39	75.0	55	14.39	85.5

^a% reporting difficulty in doing activities involving concentration and thinking *most or all of the time* during the past 4 weeks.

^b% rating health as 1 *SD* or more below the general population mean 0–100 rating during the past 4 weeks (mean = 78.34, *SD* = 19.31).

^c% reporting one or more chronic condition(s) now has that limit usual activities/enjoyment *moderately*, *quite a lot*, or *extremely*.

and well-being. Ratings of quality of life being only *fair* or *poor* (Column 1) were found at even the highest MH score level (1.4%, *T*-score range = 60+) and progressively increased to the lowest level (92.9%, *T*-score range < 15). A similar progression of increasing percentages was seen for those reporting being bothered by emotional problems *moderately*, *quite a bit*, or *extremely* (Column 2) and/or rating their happiness or satisfaction

with their personal lives as *sometimes fairly satisfied* or *generally dissatisfied* (Column 3), with 100% of the respondents reporting these problems at Levels 10 and 11, respectively.

Meanwhile, the two stress-related criterion variables presented an interesting picture. On the one hand, the percentages those reporting that they experienced a *good bit*, *quite a bit*, or a *great deal* of stress or pressure in

Table 9.25

Percentage of Adults Reporting Problems With Depression and Anxiety at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scores, 2009 U.S. General Population

MH T Scores		Depression is a current chronic condition ^a			Feeling down/ depressed/hopeless <i>more than half the days or nearly every day</i> ^b			Felt little interest/ pleasure in doing things <i>more than half the days or nearly every day</i> ^c			Anxiety is a current chronic condition ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	444	62.50	1.1	442	62.52	6.3	444	62.52	1.8	444	62.50	1.4
2	55–59.9	1,312	57.54	2.6	1,316	57.54	16.2	1,314	57.54	2.0	1,310	57.54	3.0
3	50–54.9	868	52.35	7.7	873	52.35	30.9	873	52.35	3.6	863	52.35	10.0
4	45–49.9	492	47.13	14.6	494	47.12	45.9	493	47.11	9.1	491	47.13	19.8
5	40–44.9	332	41.89	33.7	333	41.89	58.4	332	41.90	23.2	333	41.89	34.8
6	35–39.9	268	36.86	43.7	268	36.85	70.6	266	36.86	36.1	268	36.86	39.9
7	30–34.9	84	32.52	53.6	83	32.51	85.4	83	32.51	36.1	83	32.52	49.4
8	25–29.9	98	28.72	70.4	100	28.72	90.9	98	28.72	66.3	98	28.72	66.3
9	20–24.9	61	23.81	78.7	61	23.81	95.1	61	23.81	77.1	61	23.81	68.9
10	15–19.9	21	18.26	90.5	21	18.26	95.2	21	18.26	95.2	21	18.26	85.7
11	< 15	14	12.82	92.9	14	12.82	85.7	14	12.82	85.7	14	12.82	85.7

^a% reporting depression as a current chronic condition.

^b% reporting feeling down/depressed/hopeless *more than half the days or nearly every day* during the past 2 weeks.

^c% reporting felt little interest/pleasure in doing things *more than half the days or nearly every day* during the past 2 weeks.

^d% reporting anxiety as a current chronic condition.

their daily lives (Column 4) ranged from 3.0% at MH score Levels 1 and 2 to 71.4% at Level 11. On the other hand, the percentages of those reporting that stress or pressure had affected their health *moderately, quite a bit, or extremely* (Column 5) ranged from 1.1% at Level 1, increased almost ninefold to 9.6% at Level 3, and finally reached 92.9% at the highest score level (Level 11).

It is notable that, for all but the variable addressing being bothered by emotional problems (Column 2), double-digit percentages were achieved by those scoring in the lower half of the average MH T-score range (Level 4, T-score range = 45.0–49.9). In particular, at this score level, 22.4% of respondents reported that stress or pressure had affected their health *moderately, quite a bit, or extremely* (Column 5).

Pain and treatment. Table 9.27 shows that pain interference was experienced throughout the range of MH scores, whether it be through interfering with enjoyment of life (Column 1) or making one feel fed up and frustrated (Column 2). The percentages of those respondents with more outpatient visits (Column 3) and more hospital stays (Column 4) than the general population generally increased with decreasing MH score levels. For hospital stays, a relatively slow increase in percentages was seen, with neither variable reaching 15.0% until MH score Level 10. For the outpatient and inpatient variables, it is notable that only a little more than one third (35.7%) of those at the lowest MH score level reported receiving more treatment than the general population.

Cognitive functioning. The association of MH scores with forgetting recent events (Column 1) and difficulty doing activities requiring concentration and thinking (Column 2) *most or all of the time* is illustrated in Table 9.28. For both variables, the percentages of those meeting each criterion generally increased as MH scores move from Level 1 to Level 11, at which point both peaked at 66.7%.

Criterion-Based Interpretation of the Acute Form Component Summary Measures

As with the standard form, criterion-based interpretation of the SF-36v2 acute (1-week recall) form PCS measure is facilitated through an examination of the percentage of respondents from the 2009 normative sample at each level of PCS T scores whose responses to background, validation, chronic condition, and health care utilization items included on Form C used in the 2009 norms study—thought to be conceptually related to the physical health dimension and likely to covary with changes in PCS scores—were indicative of problems or limitations imposed by the respondents' physical health status. Similarly, criterion-based interpretation of the MCS measure is facilitated through an examination of the percentage of respondents from the 2009 normative sample at each level of MCS T scores whose responses to background, validation, chronic condition, and health

Table 9.26

Percentage of Adults Reporting Quality of Life, Happiness, Emotional Problems, and Stress at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scores, 2009 U.S. General Population

Level	MH T Scores			Overall quality of life rated as fair or poor ^a			Bothered by emotional problems moderately, quite a lot, or extremely ^b			Happiness/satisfaction with personal life rated as sometimes fairly satisfied or generally dissatisfied ^c			Experienced a good bit, quite a bit, or a great deal of stress/pressure in daily living ^d			Stress/pressure has affected health moderately, quite a lot, or extremely ^e			
	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	446	62.52	1.4	443	62.53	0.0	446	62.52	0.5	443	62.52	3.0	443	62.52	1.1			
2	55-59.9	1,320	57.54	2.1	1,314	57.54	0.0	1,318	57.54	2.2	1,315	57.54	3.0	1,318	57.54	3.5			
3	50-54.9	874	52.35	7.1	872	52.35	0.5	874	52.35	10.6	874	52.35	6.8	873	52.35	9.6			
4	45-49.9	496	47.13	14.5	495	47.12	2.0	494	47.13	24.7	495	47.12	16.0	495	47.12	22.4			
5	40-44.9	335	41.90	23.6	334	41.89	11.7	333	41.89	39.6	332	41.89	21.8	335	41.90	34.0			
6	35-39.9	270	36.86	39.6	267	36.84	28.1	270	36.86	60.0	269	36.86	29.9	271	36.87	56.1			
7	30-34.9	83	32.51	54.2	84	32.52	46.4	83	32.51	73.5	82	32.51	38.6	84	32.52	64.3			
8	25-29.9	100	28.72	69.0	100	28.72	70.0	99	28.72	86.9	99	28.72	58.2	98	28.72	78.6			
9	20-24.9	61	23.81	73.8	61	23.81	91.8	61	23.81	82.0	61	23.81	55.0	61	23.81	90.2			
10	15-19.9	21	18.26	85.7	21	18.26	100.0	21	18.26	95.2	21	18.26	76.2	21	18.26	95.2			
11	< 15	14	12.82	92.9	14	12.82	92.9	14	12.82	100.0	14	12.82	71.4	14	12.82	92.9			

^a% rating overall quality of life as fair or poor.

^b% reporting bothered by emotional problems moderately, quite a lot, or extremely during the past 4 weeks.

^c% rating happiness/satisfaction with personal life as sometimes fairly satisfied or generally dissatisfied.

^d% reporting having experienced a good bit, quite a bit, or a great deal of stress/pressure in daily living during the past 4 weeks.

^e% reporting stress/pressure has affected health moderately, quite a lot, or extremely during the past 4 weeks.

Table 9.27

Percentage of Adults Reporting Problems Related to Pain and Treatment at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scores, 2009 U.S. General Population

MH T Scores		Pain interfered with enjoyment of life very often or always ^a (1)			Pain made one feel fed up and frustrated very often or always ^b (2)			Number of outpatient visits significantly above the mean ^c (3)			Number of hospital stays significantly above the mean ^d (4)		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	212	62.53	2.8	213	62.53	0.9	444	62.51	5.9	443	62.49	2.9
2	55–59.9	601	57.47	3.2	599	57.47	2.0	1315	57.54	7.8	1314	57.54	2.6
3	50–54.9	431	52.36	6.0	430	52.36	5.6	871	52.35	8.2	872	52.36	3.1
4	45–49.9	257	47.10	9.7	257	47.10	5.8	495	47.12	11.3	492	47.12	4.3
5	40–44.9	187	41.88	14.4	187	41.88	13.4	329	41.89	12.2	331	41.89	5.1
6	35–39.9	155	36.93	24.5	154	36.94	23.4	266	36.85	21.8	266	36.85	7.1
7	30–34.9	49	32.58	30.6	49	32.58	34.7	82	32.51	18.3	83	32.51	10.8
8	25–29.9	56	28.37	44.6	56	28.37	48.2	98	28.72	22.5	97	28.73	8.3
9	20–24.9	31	23.99	41.9	31	23.99	48.4	61	23.81	29.5	61	23.81	9.8
10	15–19.9	8	18.21	62.5	8	18.21	87.5	21	18.26	28.6	20	18.39	15.0
11	< 15	6	11.83	66.7	6	11.83	83.3	14	12.82	35.7	14	12.82	35.7

^a% reporting pain interfered with enjoyment of life very often or always during the past 4 weeks.

^b% reporting pain made one feel fed up and frustrated very often or always during the past 4 weeks.

^c% reporting number of outpatient visits during past 4 weeks as being 1 SD or more above the mean for the general population (mean = 0.82, SD = 1.57).

^d% reporting number of hospital stays during the past 12 months as being 1 SD or more above the mean for the general population (mean = 0.22, SD = 0.89).

Table 9.28

Percentage of Adults Reporting Cognitive Functioning Problems at 11 Levels of SF-36v2 Standard (4-Week Recall) Form Mental Health Scores, 2009 U.S. General Population

MH T Scores		Forgets things that recently happened most or all of the time ^a (1)			Difficulty doing activities involving concentration and thinking most or all of the time ^b (2)		
Level	Range	n	Mean	%	n	Mean	%
1	60+	213	62.53	1.4	213	62.53	3.3
2	55–59.9	600	57.47	1.7	597	57.47	2.2
3	50–54.9	431	52.36	2.1	429	52.36	2.3
4	45–49.9	257	47.10	5.1	257	47.10	2.3
5	40–44.9	186	41.88	7.0	186	41.88	4.8
6	35–39.9	155	36.93	9.0	156	36.93	7.7
7	30–34.9	49	32.58	24.5	49	32.58	16.3
8	25–29.9	56	28.37	14.3	56	28.37	16.1
9	20–24.9	31	23.99	38.7	31	23.99	38.7
10	15–19.9	8	18.21	25.0	8	18.21	25.0
11	< 15	6	11.83	66.7	6	11.83	66.7

^a% reporting forgetting things that recently happened most or all of the time during the past 4 weeks.

^b% reporting difficulty in doing activities involving concentration and thinking most or all of the time during the past 4 weeks.

care utilization items—thought to be conceptually related to the mental health dimension and likely to covary with changes in MCS scores—were indicative of problems or limitations imposed by the respondents' mental health status.

Physical Component Summary (PCS)

Tables 9.29 through 9.35 provide data for the criterion-based interpretations of SF-36v2 acute form PCS T scores relative to limitations in physical and role-functioning activities, pain interference, health care utilization, employment status, presence of chronic conditions, and ratings of quality of life, as well as ratings of general health, job performance, future health, and work-related problems.

General health, HRQOL, and PCS. Table 9.29 presents data related to the general health and quality of life criterion variables. When considering overall quality of life (Column 1) in relation to PCS scores, there was a moderately paced increase in the percentage of those reporting fair or poor quality of life from the highest to the lowest PCS score levels. While 60% reported this condition at Level 8, the percentage dropped to 44.4% at the lowest score level (Level 9). Generally, this variable was useful in interpreting PCS score differences at all levels, as were health ratings of poor or very poor (Column 2) and 0–100 scale health ratings that fell 1 SD below the population mean (Column 3).

In addition, Table 9.29 shows that the percentages of those having more chronic conditions than the mean for the U.S. general population (Column 4) increased

Table 9.29

Percentage of Adults Reporting General Health and Quality of Life Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

Level	PCS T Scores		Overall quality of life rated as <i>fair</i> or <i>poor</i> ^a (1)			Health rated as <i>poor</i> or <i>very poor</i> ^b (2)			Health rating significantly below the mean ^c (3)			Number of chronic conditions significantly above the mean ^d (4)			Chronic condition(s) limit usual activities/ enjoyment <i>moderately</i> , <i>quite a lot</i> , or <i>extremely</i> ^e (5)		
	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	
1	55+	839	58.39	4.4	836	58.4	0.4	836	58.39	2.9	838	58.40	1.7	841	58.40	6.4	
2	50-54.9	451	52.74	7.5	451	52.73	1.1	449	52.74	6.0	452	52.73	8.6	452	52.73	14.4	
3	45-49.9	228	47.77	15.8	229	47.77	2.6	227	47.77	12.8	227	47.75	12.3	229	47.77	24.0	
4	40-44.9	195	42.58	20.5	195	42.58	6.2	194	42.58	23.7	195	42.58	26.7	196	42.58	50.0	
5	35-39.9	116	37.55	21.6	116	37.55	7.8	115	37.55	19.1	116	37.55	35.3	116	37.55	60.3	
6	30-34.9	90	32.65	47.8	90	32.65	25.6	90	32.65	46.7	90	32.65	53.3	90	32.65	78.9	
7	25-29.9	69	27.96	47.8	69	27.96	40.6	67	27.93	52.2	69	27.96	53.6	69	27.96	87.0	
8	20-24.9	45	22.73	60.0	45	22.73	46.7	45	22.73	57.8	45	22.73	55.6	45	22.73	77.8	
9	<20	18	16.67	44.4	18	16.67	77.8	18	16.67	72.2	18	16.67	55.6	18	16.67	94.4	

^a% rating overall quality of life as *fair* or *poor*.

^b% rating health as *poor* or *very poor* during the past week.

^c% rating health 1 *SD* or more below the general population mean 0-100 rating during the past 4 weeks.

^d% reporting the number of chronic conditions ever been told he/she had as being 1 *SD* or more above the mean for the general population (mean = 2.27, *SD* = 2.23).

^e% reporting one or more chronic condition(s) ever been told he/she had that limit usual activities/enjoyment *moderately*, *quite a lot*, or *extremely*.

steadily from Level 1 (1.7%) to Levels 6 and 7 (53.3% and 53.6%, respectively) to Levels 8 and 9 (55.6%). Overall, this variable is useful in interpreting differences at the higher and middle PCS score levels. Among respondents with one or more chronic conditions, there was a near perfect linear increase in the percentage who reported being limited *moderately, quite a bit, or extremely* in usual activities or enjoyment (Column 5) from the highest score level (6.4% at Level 1) to the lowest score level (94.4% at Level 9), thus making this variable useful in interpreting score differences throughout the range of score levels.

Performance of work and other activities and PCS. With regard to the ability to work or engage in other activities (Table 9.30), there was a linear increase in the percentages of those reporting they *could not do* or had *quite a lot* of difficulty doing usual activities due to physical conditions (Column 2) and difficulty doing daily work (Column 3). For both variables, there was more than a threefold increase in the percentages observed from PCS score Level 4 (*T*-score range = 40.0–44.9, which includes the first scores below the average range) to Level 5 (*T*-score range = 35.0–39.9). This was followed by a twofold increase in percentages from Level 5 to Level 6 for both variables. As such, both of these variables are considered useful in interpreting score differences across all score levels.

Table 9.30 also reveals a near perfect linear increase across the PCS score levels in the percentages of respondents who reported being disabled (Column 1) and having missed more than the average number of workdays due to illness or injury (Column 4). While there was a fairly steady and significant increase in reports of being disabled through to the lowest score level (Level 9), no more than 16.7% reported a significant (i.e., more than 1 *SD* above the mean) number of missed workdays due to illness or injury. Overall, the former is useful for interpreting score differences at all PCS score levels, whereas the latter is considered useful at the higher and middle levels.

Health problems, treatment, and PCS. As PCS scores decreased, nearly perfect linear increases occurred in the percentages of respondents who reported significantly more outpatient visits (Column 2) and hospital stays (Column 3) than the general population mean, making both of these variables useful for the interpretation of score differences at all score levels (see Table 9.31). Note that in both cases, the percentages at the lowest PCS score level (50.0% and 38.9%, respectively, at Level 9) were lower than one might expect. Meanwhile, a similar pattern of increasing percentages was seen with those reporting days in bed due to illness or injury (Column 1) except at some of the lowest score

levels (Levels 6–8), thus limiting the variable's usefulness at these levels.

Sleep disturbance and PCS. Table 9.32 reveals a generally slow and linear increase in the percentages of those who reported sleep not being quiet (Column 1), trouble falling asleep (Column 2), and awakening with trouble falling back to sleep (Column 3) *most or all of the time*. Because the association between the percentages and scores seemed less direct at the lower PCS score levels, these three variables appear to be most useful for interpreting score differences at the highest and middle PCS score levels.

Sleep somnolence and PCS. Again, a generally slow and linear increase occurred in the percentage of respondents who reported various sleep-related problems with decreasing PCS score levels, as presented in Table 9.33. A significant percentage of respondents began reporting feeling drowsy or sleepy (Column 1) and taking naps (Column 3) during the day *most or all of the time* even at Level 2 (14.0% and 10.9%, respectively), which includes the upper half of the average *T*-score range (50.0–54.9) for PCS. On the other hand, trouble staying awake during the day does not become problematic until Level 6 (14.4%, *T*-score range = 30.0–34.9). Despite the differences, all criteria are considered to be useful throughout all PCS score levels.

Sleep quantity, sleep adequacy, headache/shortness of breath, and PCS. Table 9.34 presents the findings pertaining to PCS score levels and a variety of sleep-related problems, including getting enough sleep to feel rested (Column 1) and getting the needed amount of sleep (Column 2) *little or none of the time*, as well as awakening short of breath or with a headache *most or all of the time* (Column 3). The percentages of respondents who reported problems in these areas increased with decreasing PCS score levels in a perfect or near perfect linear fashion. It is interesting to note that reports of awakening short of breath or with a headache appeared to be an infrequent occurrence except at the lowest PCS score levels. Thus, this criterion is most useful in interpreting score differences at the lowest score levels whereas the other two variables demonstrate interpretive utility across the range of PCS scores.

Future health, work-related problems, and PCS. Table 9.35 offers a look at the relationship between a respondent's baseline PCS score level and the occurrence of health-related events assessed 3 to 4 months later. Generally, those scoring at the highest PCS score level (Level 1) were less likely to report one or more outpatient visits (Column 1) and/or not working at a paying job because of health (Column 2) at the time of reassessment than those scoring at the lowest PCS score levels.

Table 9.30

Percentage of Adults Reporting Problems in Work Performance and Other Activities at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Current employment status is disabled ^a			Could not do or had quite a lot of difficulty doing usual activities due to physical conditions ^b			Could not do or had quite a lot of difficulty doing daily work ^c			Days of missed work due to illness/injury significantly above the mean ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	838	58.40	0.8	836	58.41	1.0	839	58.39	0.0	546	58.40	1.5
2	50–54.9	449	52.74	2.5	451	52.73	1.6	450	52.74	1.1	264	52.90	3.0
3	45–49.9	228	47.77	34.0	228	47.77	2.6	227	47.78	2.6	116	47.79	8.6
4	40–44.9	195	42.58	11.8	194	42.58	7.7	193	42.61	5.7	75	42.59	9.3
5	35–39.9	116	37.55	22.4	115	37.55	25.2	116	37.55	19.8	31	38.14	12.9
6	30–34.9	90	32.65	32.2	90	32.65	52.2	90	32.65	42.2	6	32.90	16.7
7	25–29.9	67	27.93	41.8	69	27.96	73.9	68	27.94	67.7	8	27.51	12.5
8	20–24.9	45	22.73	51.1	45	22.73	95.6	45	22.73	77.8	6	22.98	16.7
9	< 20	18	16.67	61.1	18	16.67	100.0	18	16.67	88.9	0	—	—

^a% reporting current work status as *disabled*.

^b% reporting *could not do* or were limited *quite a lot* in usual activities due to physical health during the past week.

^c% reporting *could not do* or had *quite a lot* of difficulty doing daily work due to physical health during the past week.

^d% reporting number of days of work missed due to illness or injury as being 1 *SD* or more above the mean for the general population during the past 4 weeks (mean = 0.39, *SD* = 1.65).

Table 9.31

Percentage of Adults Reporting Health Problems and Treatment at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Days in bed due to illness/injury significantly above the mean ^a			Number of outpatient visits significantly above the mean ^b			Number of hospital stays significantly above the mean ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	833	58.38	0.1	837	58.39	2.9	836	58.40	1.3
2	50–54.9	449	52.72	1.6	448	52.73	4.2	448	52.73	1.8
3	45–49.9	224	47.77	4.5	228	47.77	10.5	227	47.83	4.0
4	40–44.9	194	42.58	6.2	195	42.58	12.8	193	42.59	4.7
5	35–39.9	116	37.55	7.8	116	37.55	22.4	116	37.55	9.5
6	30–34.9	88	32.65	26.1	90	32.65	36.7	90	32.65	12.2
7	25–29.9	67	27.92	16.4	68	27.94	35.3	68	27.94	11.8
8	20–24.9	45	22.73	22.2	45	22.73	44.4	45	22.73	11.1
9	< 20	17	16.63	29.4	18	16.67	50.0	18	16.67	38.9

^a% reporting number of days in bed due to illness or injury during the past 4 weeks as being 1 *SD* or more above the mean for the general population (mean = 1.00, *SD* = 3.35).

^b% reporting number of outpatient visits during past 4 weeks as being 1 *SD* or more above the mean for the general population (mean = 0.89, *SD* = 1.73).

^c% reporting number of hospital stays during the past 12 months as being 1 *SD* or more above the mean for the general population (mean = 0.23, *SD* = 0.89).

Mental Component Summary (MCS)

Tables 9.36 through 9.43 provide data for the criterion-based interpretations of SF-36v2 acute form MCS *T* scores relative to behavioral health; emotional, personal, and physical problems; interference of pain on functioning; ratings of quality of life, general health, and job performance; problems with sleep, cognitive functioning, and energy level; employment status; and future mental health and work-related problem.

Depression, anxiety, and MCS. Table 9.36 demonstrates a linear increase—from the highest to the lowest score level—in the percentages of respondents who

reported frequently feeling down, depressed, or hopeless (Column 1) or currently having depression (Column 2) and/or anxiety (Column 3). At Level 5, which includes the MCS *T*-score cutoff for depression screening (i.e., *T*-score ≤ 42), 18.2% of the respondents reported feeling down/depressed/hopeless *more than half* or *nearly every day*, 37.5% reported depression as a current condition, and 42.3% reported anxiety as a current condition. Overall, all three variables are useful in interpreting MCS score differences across all score levels.

Effects of personal, emotional, and physical problems and MCS. With a couple exceptions (see Column

Table 9.32

Percentage of Adults Reporting Sleep Disturbance Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Sleep not quiet most or all of the time ^a			Trouble falling asleep most or all of the time ^b			Awakened during sleep and trouble falling back to sleep most or all of the time ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	837	58.39	9.4	839	58.39	7.4	839	58.39	4.9
2	50–54.9	451	52.74	15.7	451	52.74	12.0	451	52.74	7.8
3	45–49.9	228	47.77	19.3	228	47.77	12.3	227	47.77	11.0
4	40–44.9	195	42.58	27.7	195	42.58	20.0	195	42.58	14.9
5	35–39.9	116	37.55	25.9	115	37.54	20.0	116	37.55	21.6
6	30–34.9	89	32.66	33.7	90	32.65	31.1	90	32.65	24.4
7	25–29.9	68	27.94	44.1	69	27.96	27.5	69	27.96	29.0
8	20–24.9	45	22.73	40.0	45	22.73	37.8	45	22.73	26.7
9	< 20	18	16.67	50.0	18	16.67	33.3	18	16.67	33.3

^a% reporting sleep not being quiet *most or all of the time* during the past week.

^b% reporting having trouble falling asleep *most or all of the time* during the past week.

^c% reporting being awakened during sleep and having trouble falling back to sleep *most or all of the time* during the past week.

Table 9.33

Percentage of Adults Reporting Sleep Somnolence Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Felt drowsy/ sleepy during the day most or all of the time ^a			Trouble staying awake during the day most or all of the time ^b			Take naps during the day most or all of the time ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	839	58.39	9.3	840	58.39	2.3	839	58.39	4.8
2	50–54.9	450	52.74	14.0	450	52.73	3.8	450	52.74	10.9
3	45–49.9	227	47.76	19.8	228	47.77	7.0	228	47.77	11.4
4	40–44.9	195	42.58	28.7	194	42.58	8.8	195	42.58	17.4
5	35–39.9	116	37.55	26.7	115	37.54	7.0	115	37.53	27.8
6	30–34.9	90	32.65	31.1	90	32.65	14.4	90	32.65	28.9
7	25–29.9	68	27.94	33.8	69	27.96	18.8	68	27.94	19.1
8	20–24.9	45	22.73	55.6	45	22.73	24.4	45	22.73	37.8
9	< 20	18	16.67	55.6	18	16.67	50.0	18	16.67	50.0

^a% reporting having felt drowsy/sleepy during the day *most or all of the time* during the past week.

^b% reporting having trouble staying awake during the day *most or all of the time* during the past week.

^c% reporting having to take naps during the day *most or all of the time* during the past week.

2), Table 9.37 demonstrates the perfect ordering across all 10 MCS score levels of increasing percentages of respondents being significantly limited in usual social activities due to physical health/emotional problems (Column 1), bothered at least *moderately* by emotional problems (Column 2), only *fairly satisfied* or *generally dissatisfied* with their personal lives (Column 3), and feeling little interest or pleasure in doing things (Column 4). Overall, being bothered by emotional problems is useful in interpreting MCS score differences at the higher and middle levels, while three remaining variables are useful throughout all MCS score levels.

Job performance and the effects of stress and MCS.

Table 9.38 reveals a near perfect relationship between decreasing MCS score levels and increasing percentages of respondents who reported that they *could not do* or were *kept quite a lot* from doing usual activities due to personal, emotional, or physical problems (Column 1). These percentages slowly increased through the higher and middle score levels, with only 13.2% reporting the problem at Level 6 (*T*-score range = 35.0–39.9) and rapidly increasing thereafter. Somewhat different and more inconsistent patterns of reporting were seen for ratings of overall job performance being significantly below

Table 9.34

Percentage of Adults Reporting Problems With Sleep Quantity and Adequacy and Headaches or Shortness of Breath at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Got enough sleep to feel rested little or none of the time ^a (1)			Awakened short of breath or with a headache most or all of the time ^b (2)			Getting the needed amount of sleep little or none of the time ^c (3)		
Level	Range	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	55+	838	58.39	17.5	839	58.39	0.7	837	58.39	19.4
2	50–54.9	449	52.74	21.6	449	52.74	1.6	451	52.74	25.1
3	45–49.9	227	47.75	24.7	227	47.75	3.5	227	47.77	24.7
4	40–44.9	195	42.58	31.8	195	42.58	5.1	195	42.58	33.9
5	35–39.9	116	37.55	34.5	116	37.55	6.0	116	37.55	33.6
6	30–34.9	90	32.65	46.7	90	32.65	7.8	90	32.65	41.1
7	25–29.9	66	27.91	48.5	68	27.94	16.2	69	27.96	46.4
8	20–24.9	45	22.73	48.9	45	22.73	20.0	44	22.73	54.6
9	< 20	18	16.67	66.7	18	16.67	16.7	18	16.67	55.6

^a% reporting getting enough sleep to feel rested little or none of the time during the past week.

^b% reporting awakening short of breath or with a headache most or all of the time during the past week.

^c% reporting getting the needed amount of sleep little or none of the time during the past week.

Table 9.35

Percentage of Adults Reporting Future Health and Work-Related Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Component Summary Measure Scores, 2009 U.S. General Population

PCS T Scores		Outpatient visit with health professional ^a (1)			Not working because of health ^b (2)		
Level	Range	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	55+	80	58.73	23.8	80	58.73	26.3
2	50–54.9	34	52.33	41.2	34	52.33	32.4
3	45–49.9	21	47.33	42.9	21	47.33	52.4
4	40–44.9	13	42.92	76.9	13	42.92	84.6
5	35–39.9	10	37.13	40.0	10	37.13	90.0
6	30–34.9	5	32.06	60.0	5	32.06	100.0
7	< 30	8	22.54	87.5	8	22.54	75.0

^a% reporting one or more outpatient visits with a health professional during the 4 weeks preceding survey readministration.

^b% reporting not working at a paying job because of health at the time of survey readministration.

the population mean (Column 2) and being disabled (Column 3). Compared to those reporting problems in doing usual activities, the percentages reporting job performance and disability problems increased faster at the higher and middle MCS score levels and progressed more slowly at the lowest levels. Overall, the three variables appear to be most useful for interpreting score differences at the middle and lower levels.

Table 9.38 also demonstrates both the commonness of reported stress or pressure in daily living (Column 4) and its adverse effect on health (Column 5), as well as their similar pattern of decreasing MCS scores. Even at the two MCS score levels that encompass the average

score range for individual respondents (Levels 3 and 4), the percentages of respondents who reported a *good bit*, *quite a bit*, or a *great deal* of stress or pressure (34.4% and 55.8%, respectively) were significant, as were the percentages of those who reported that stress or pressure affected their health *moderately*, *quite a lot*, or *extremely* (10.0% and 17.8%, respectively). Overall, both of these variables are most useful in interpreting score differences at all MCS score levels.

Health, quality of life, energy level, and MCS. Table 9.39 reveals a pattern of decreasing MCS scores with increasing percentages of respondents who rated their health as significantly below the general population mean (Column 1), their quality of life as *fair* or *poor* (Column 2), and their amount of energy as *little* or *none* (Column 3). The extent of these problems was apparent in the percentages reported for each (15.6%, 17.2%, and 18.8%, respectively) at MCS score Level 4, which represents the lower half of the average range of MCS scores (*T*-score range = 45.0–49.9). In general, these three variables appear most useful in interpreting score differences at the higher and middle MCS score levels.

Sleep disturbance and MCS. Table 9.40 shows that reports of various aspects of sleep disturbance were relatively common in the 2009 normative sample. In general, reports of sleep disturbances tended to increase as MCS scores began to fall through to the lower levels, with significant percentages of respondents reporting such problems even as high as MCS score Level 3, which includes the mean MCS *T* score of 50. At this score level, 51.7% of respondents reported needing a longer time to fall asleep than the modal time for the general population (Column 1). At Level 4, which represents the lower

Table 9.36

Percentage of Adults Reporting Depression and Anxiety at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Feeling down/depressed/ hopeless more than half the days or nearly every day ^a			Depression is a current chronic condition ^b			Anxiety is a current chronic condition ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	238	61.94	0.0	4	60.51	1.7	9	62.25	3.9
2	55–59.9	716	57.20	0.3	22	56.87	3.1	28	56.87	3.9
3	50–54.9	418	52.80	0.5	35	52.83	8.4	50	52.29	12.1
4	45–49.9	230	47.68	5.7	39	47.75	17.0	43	47.46	18.8
5	40–44.9	176	42.52	18.2	66	42.56	37.5	74	42.54	42.3
6	35–39.9	106	37.65	34.9	53	37.87	50.5	47	37.63	44.8
7	30–34.9	63	32.71	39.7	33	32.31	54.1	35	32.66	56.5
8	25–29.9	36	28.11	80.6	27	28.21	77.1	21	28.23	60.0
9	20–24.9	33	21.91	84.9	27	21.90	81.8	28	21.88	84.9
10	< 20	24	15.54	91.7	21	15.39	87.5	20	15.36	83.3

^a% reporting feeling down/depressed/hopeless more than half the days or nearly every day during the past 2 weeks.

^b% reporting depression as a current chronic condition.

^c% reporting anxiety as a current chronic condition.

Table 9.37

Percentage of Adults Reporting Negative Effects of Personal, Emotional, and Physical Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Could not do or was limited quite a lot in usual social activities due to physical health/emotional problems ^a			Bothered by emotional problems moderately, quite a lot, or extremely ^b			Happiness/satisfaction with personal life rated as sometimes fairly satisfied or generally dissatisfied ^c			Felt little interest/ pleasure in doing things more than half the days or nearly every day ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	238	61.94	0.8	236	61.96	0.4	237	61.94	1.3	237	61.95	1.7
2	55–59.9	716	57.21	1.0	717	57.20	0.4	717	57.21	3.2	718	57.21	0.8
3	50–54.9	418	52.80	3.4	417	52.81	3.8	419	52.8	13.1	418	52.79	4.3
4	45–49.9	234	47.69	6.0	231	47.69	18.2	233	47.69	36.5	232	47.69	12.5
5	40–44.9	178	42.54	14.6	179	42.54	36.3	176	42.52	46.6	178	42.52	28.7
6	35–39.9	107	37.64	17.8	107	37.64	63.6	105	37.67	55.2	106	37.65	41.5
7	30–34.9	63	32.71	28.6	63	32.71	77.8	63	32.71	65.1	63	32.71	47.6
8	25–29.9	36	28.11	38.9	36	28.11	77.8	36	28.11	77.8	36	28.11	72.2
9	20–24.9	33	21.91	69.7	33	21.91	66.7	33	21.91	90.9	33	21.91	75.8
10	< 20	24	15.54	87.5	24	15.54	54.2	24	15.54	91.7	24	15.54	95.8

^a% reporting they could not do or were limited quite a lot in usual social activities due to physical health/emotional problems during the past week.

^b% reporting bothered by emotional problems moderately, quite a lot, or extremely during the past week.

^c% rating happiness/satisfaction with personal life as sometimes fairly satisfied or generally dissatisfied during the past 4 weeks.

^d% reporting felt little interest/pleasure in doing things more than half the days or nearly every day during the past 2 weeks.

half of the average range of MCS scores (T -score range = 45.0–49.9), 61.6% reported needing a longer time to fall asleep, 24.6% indicated their sleep was not quiet (Column 2), 20.6% had trouble falling asleep (Column 3), and 13.7% were awakened during sleep and then had trouble falling back to sleep (Column 4) *most* or *all of the time*. Note that although a relationship of increasing percentages of reported problems with decreasing MCS scores generally exists, a perfect linear increase in percentages with decreasing MCS scores did not ex-

ist for any of these four aspects of sleep disturbance. In sum, the awakening and trouble falling back to sleep variables are useful in interpreting MCS score differences across the range of scores, while the other three variables discussed in Table 9.40 are most useful in the higher and/or lower levels.

Sleep somnolence and MCS. Table 9.41 reveals that reports of feeling drowsy or sleepy during the day (Column 1) and having trouble staying awake during the day (Column 2) *most* or *all of the time* increased with

Table 9.38

Percentage of Adults Reporting Job Performance Problems and the Effects of Stress at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores Level	Could not do or were kept quite a lot from doing usual work/ school/other activities ^a			Rating of overall job performance significantly below the mean ^b			Employment status is disabled ^c			Experienced a good bit, quite a bit, or a great deal of stress/ pressure in daily living ^d			Stress/pressure has affected health moderately, quite a lot, or extremely ^e		
	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	239	61.95	0.4	81	61.55	7.4	239	61.95	7.1	237	61.95	9.3	238	61.95	4.2
2	717	57.21	0.1	371	57.05	3.0	720	57.21	2.6	718	57.20	15.7	717	57.23	4.2
3	419	52.80	1.7	262	52.91	3.8	418	52.80	6.0	419	52.80	34.4	419	52.80	10.0
4	233	47.69	2.6	139	47.64	10.8	231	47.71	6.9	233	47.69	55.8	231	47.71	17.8
5	177	42.55	7.3	91	42.37	14.3	177	42.55	13.6	177	42.52	71.8	178	42.52	43.8
6	106	37.63	13.2	46	37.79	28.3	105	37.68	19.1	107	37.64	76.6	107	37.64	51.4
7	63	32.71	23.8	26	32.76	15.4	63	32.71	20.6	63	32.71	85.7	63	32.71	58.7
8	34	28.09	44.1	17	28.58	35.3	36	28.11	19.4	35	28.10	85.7	35	28.10	77.1
9	33	21.91	66.7	11	21.85	45.5	33	21.91	48.5	33	21.91	90.9	33	21.91	84.9
10	24	15.54	87.5	6	13.49	50.0	24	15.54	41.7	24	15.54	95.8	23	15.60	78.3

^a% reporting they *could not do* or were kept *quite a lot* from doing usual work/school/other activities during the past week due to personal/emotional/physical problems.

^b% rating overall job performance as 1 *SD* or more below the general population mean 0–10 rating during the past 4 weeks (mean = 8.33, SD = 1.50).

^c% reporting employment status as *disabled*.

^d% reporting having experienced a *good bit*, *quite a bit*, or a *great deal* of stress/pressure in daily living during the past 4 weeks.

^e% reporting stress/pressure has affected health *moderately*, *quite a lot*, or *extremely* during the past 4 weeks.

Table 9.39

Percentage of Adults Reporting Problems in Health, Quality of Life, and Energy Level at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Health rating significantly below the mean ^a (1)			Overall quality of life rated as fair or poor ^b (2)			Little or no energy ^c (3)		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	239	61.95	8.0	239	61.95	4.2	239	61.95	5.0
2	55–59.9	715	57.21	4.3	719	57.21	2.8	721	57.21	2.6
3	50–54.9	418	52.80	8.1	419	52.80	7.6	416	52.79	10.8
4	45–49.9	231	47.69	15.6	233	47.69	17.2	234	47.69	18.8
5	40–44.9	178	42.52	27.0	178	42.52	29.8	178	42.55	35.4
6	35–39.9	107	37.64	25.2	107	37.64	32.7	107	37.64	43.9
7	30–34.9	63	32.71	27.0	63	32.71	44.4	63	32.71	38.1
8	25–29.9	34	28.10	52.9	36	28.11	63.9	36	28.11	66.7
9	20–24.9	33	21.91	60.6	33	21.91	69.7	32	21.83	87.5
10	< 20	23	15.60	60.9	24	15.54	79.2	24	15.54	83.3

^a% rating health 1 SD or more below the general population mean 0–100 rating during the past 4 weeks (mean = 78.17, SD = 19.30).

^b% rating overall quality of life as fair or poor.

^c% reporting amount of energy as a little or none during the past week.

Table 9.40

Percentage of Adults Reporting Sleep Disturbance Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Number of minutes to fall asleep significantly above the mode ^a (1)			Sleep not quiet most or all of the time ^b (2)			Trouble falling asleep most or all of the time ^c (3)			Awakened during sleep and trouble falling back to sleep most or all of the time ^d (4)		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	239	61.95	38.9	239	61.95	11.3	239	61.95	6.3	239	61.95	5.4
2	55–59.9	717	57.22	39.6	717	57.22	6.6	718	57.20	4.0	720	57.21	2.8
3	50–54.9	418	52.81	51.7	419	52.8	13.4	419	52.80	8.1	419	52.8	6.9
4	45–49.9	232	47.69	61.6	232	47.69	24.6	233	47.69	20.6	233	47.69	13.7
5	40–44.9	178	42.52	77.5	177	42.51	35.6	178	42.52	27.0	178	42.52	19.1
6	35–39.9	107	37.64	76.6	107	37.64	32.7	107	37.64	22.4	106	37.63	21.7
7	30–34.9	63	32.71	74.6	63	32.71	34.9	63	32.71	41.3	62	32.73	24.2
8	25–29.9	36	28.11	86.1	36	28.11	58.3	36	28.11	38.9	36	28.11	47.2
9	20–24.9	33	21.91	90.9	33	21.91	57.6	33	21.91	63.6	33	21.91	48.5
10	< 20	24	15.54	87.5	24	15.54	75.0	24	15.54	70.8	24	15.54	66.7

^a% reporting the number of minutes to fall asleep as being significantly above the mode (≥ 16 minutes) during the past week.

^b% reporting sleep not being quiet most or all of the time during the past week.

^c% reporting having trouble falling asleep most or all of the time during the past week.

^d% reporting being awakened during sleep and having trouble falling back to sleep most or all of the time during the past week.

decreasing MCS scores; however, the later issue was not as problematic as the former. At Level 9 (*T*-score range = 20.0–24.9), 72.7% reported feeling drowsy or sleepy while only 39.4% reported having trouble staying awake during the day. Overall, both variables seem to be most useful in interpreting score differences at the middle MCS score levels.

Sleep quantity and adequacy and headache/shortness of breath and MCS. Table 9.42 presents the findings regarding a variety of other sleep-related problems and their impact on MCS. Both getting enough sleep to feel

rested (Column 1) and getting the needed amount of sleep (Column 3) little or none of the time appeared to assess similar aspects of sleep problems, revealed very similar patterns of increasing percentages with decreasing MCS scores, and were useful in interpreting score differences across the entire range of MCS scores. The percentages of respondents who reported awakening short of breath or with a headache (Column 2) most or all of the time also increased with decreasing MCS scores; in contrast, however, these problems were reported much less frequently, especially at the lowest score levels. As a result,

Table 9.41

Percentage of Adults Reporting Sleep Somnolence Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Felt drowsy/sleepy during the day most or all of the time ^a			Trouble staying awake during the day most or all of the time ^b		
Level	Range	n	Mean	%	n	Mean	%
1	60+	239	61.95	5.4	239	61.95	2.5
2	55–59.9	719	57.21	5.3	719	57.21	1.3
3	50–54.9	419	52.8	12.4	419	52.8	3.3
4	45–49.9	233	47.69	23.6	232	47.7	8.6
5	40–44.9	177	42.52	37.9	178	42.52	10.1
6	35–39.9	107	37.64	42.1	107	37.64	15.0
7	30–34.9	62	32.73	43.6	63	32.71	19.1
8	25–29.9	35	28.16	60.0	35	28.13	17.1
9	20–24.9	33	21.91	72.7	33	21.91	39.4
10	< 20	24	15.54	70.8	24	15.54	37.5

^a% reporting having felt drowsy/sleepy during the day most or all of the time during the past week.

^b% reporting having trouble staying awake during the day most or all of the time during the past week.

this criterion variable is most useful in interpreting score differences at the lower MCS score levels.

Future mental health problems and MCS. Table 9.43 offers a look at the relationship between a respondent's baseline MCS score level and the occurrence of health-related events assessed 3 to 4 months later. A near perfect linear relationship existed between decreasing MCS scores at baseline and reassessment reports of having had problems with feeling down, depressed, or

hopeless (Column 1) and having little interest or pleasure in doing things (Column 2) during the preceding 2 weeks. Note that significant increases in the percentage of respondents who reported either or both of these problems were realized when moving from score Level 4 (21.1% and 30.0%, respectively) to score Level 5 (82.4% and 55.6%, respectively).

Criterion-Based Interpretation of the Acute Form Health Domain Scales

Tables 9.44 through 9.54 present the findings from the 2009 normative study regarding reported problems on relevant criterion variables at each of the health domain T-score levels.

Physical Functioning (PF)

Quality of life and performance of work and other activities. As shown in Table 9.44, a clear association exists between decreasing PF scale scores and increasing percentages of respondents who reported quality of life as *fair* or *poor* (Column 1), they *could not do* or had *quite a lot* of difficulty doing usual activities due to physical conditions (Column 2), more than an average number of bed days due to illness or injury (Column 3), and being disabled (Column 4). Notably, the difficulty performing usual activities variable saw percentages more than double from Level 4 to Level 5 (11.1% and 26.7%, respectively) and again from Level 5 to Level 6 (26.7% and 54.7%, respectively). A linear increase is apparent in the percentage of those who reported diffi-

Table 9.42

Percentage of Adults Reporting Problems With Sleep Quantity and Adequacy and Headaches or Shortness of Breath at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Got enough sleep to feel rested little or none of the time ^a			Awakened short of breath or with a headache most or all of the time ^b			Getting the needed amount of sleep little or none of the time ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	238	61.96	8.4	239	61.95	1.3	237	61.94	9.3
2	55–59.9	718	57.21	8.5	719	57.21	0.6	718	57.21	11.1
3	50–54.9	418	52.81	22.5	418	52.8	2.6	419	52.8	23.2
4	45–49.9	232	47.69	41.8	233	47.69	3.4	233	47.69	43.4
5	40–44.9	176	42.51	46.6	177	42.52	5.7	177	42.52	48.6
6	35–39.9	107	37.64	45.8	107	37.64	6.5	107	37.64	47.7
7	30–34.9	62	32.69	54.8	62	32.69	8.1	63	32.71	58.7
8	25–29.9	36	28.11	66.7	35	28.11	14.3	36	28.11	63.9
9	20–24.9	33	21.91	90.9	33	21.91	24.2	33	21.91	69.7
10	< 20	24	15.54	79.2	24	15.54	29.2	24	15.54	79.2

^a% reporting getting enough sleep to feel rested little or none of the time during the past week.

^b% reporting awakening short of breath or with a headache most or all of the time during the past week.

^c% reporting getting the needed amount of sleep little or none of the time during the past week.

Table 9.43

Percentage of Adults Reporting Future Mental Health Problems at 7 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Component Summary Measure Scores, 2009 U.S. General Population

MCS T Scores		Down/depressed/hopeless <i>several, more than half, or nearly every day</i> ^a (1)			Little interest/pleasure in doing things <i>several, more than half, or nearly every day</i> ^b (2)		
Level	Range	n	Mean	%	n	Mean	%
1	60+	21	61.40	4.8	21	61.40	0.0
2	55–59.9	40	57.41	12.5	37	57.32	13.5
3	50–54.9	39	53.10	18.0	39	53.10	28.2
4	45–49.9	19	47.29	21.1	20	47.26	30.0
5	40–44.9	17	42.36	82.4	18	42.42	55.6
6	35–39.9	12	37.65	75.0	12	37.65	91.7
7	< 35	21	27.79	95.2	21	27.79	81.0

^a% reporting feeling down/depressed/hopeless *several, more than half, or nearly every day* during the 2 weeks preceding survey readministration.

^b% reporting experiencing little interest or pleasure in doing things *several, more than half, or nearly every day* during the 2 weeks preceding survey readministration.

culty doing usual activities and being disabled. Overall, all variables demonstrate usefulness in interpreting score differences across all PF score levels.

Role-Physical (RP)

Table 9.45 highlights the perfect linear relationship between decreasing RP score levels and increasing percentages of respondents who reported having dif-

ficulty doing usual activities due to physical conditions (Column 3) and having one or more chronic conditions that *moderately, quite a bit, or extremely* limit usual activities or enjoyment (Column 4). For both variables, 86.8% report such problems at the lowest score level (Level 8); however, whereas the percentage reporting the former problem slowly increased, significant percentages reported chronic condition-related limitations even at the higher RP score levels that represent the average score range: 18.6% at Level 2 and 35.6% at Level 3. A slower, more steady progression of increasing percentages with decreasing RP scores was seen for bed days due to injury or illness (Column 1) and being disabled (Column 2). In general, these four variables are useful for interpreting score differences throughout the range of RP score levels.

Bodily Pain (BP)

Table 9.46 reveals a general ordering of decreasing of BP scale scores with increasing percentages of respondents who reported being disabled (Column 3) and having more than the average number of chronic conditions (Column 1) and hospital stays (Column 2). Note that even at the lowest score level (Level 9, *T*-score range < 25), none of the reported percentages were greater than 70%. In sum, these BP variables are most useful in the mid-range score levels for chronic conditions, the middle and lowest score levels for hospital stays, and all score levels for the disabled status.

Table 9.44

Percentage of Adults Reporting Problems Related to Quality of Life and the Performance of Work and Other Activities at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Physical Functioning Scores, 2009 U.S. General Population

PF T Scores		Overall quality of life rated as <i>fair or poor</i> ^a (1)			<i>Could not do or had quite a lot of difficulty doing usual activities due to physical conditions</i> ^b (2)			Days in bed due to illness/injury significantly above the mean ^c (3)			Current employment status is <i>disabled</i> ^d (4)		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	966	57.02	4.1	961	57.02	1.3	958	57.02	3.1	965	57.02	0.5
2	50–54.9	354	52.90	8.8	354	52.89	1.4	351	52.90	6.6	352	52.88	3.4
3	45–49.9	262	48.18	19.9	262	48.18	3.1	261	48.18	14.9	262	48.18	9.2
4	40–44.9	163	42.37	20.3	162	42.36	11.1	161	42.34	14.3	163	42.37	9.8
5	35–39.9	91	37.72	24.2	90	37.65	26.7	90	37.65	23.3	90	37.65	21.1
6	30–34.9	86	32.53	34.9	86	32.53	54.7	84	32.52	29.8	86	32.53	29.1
7	25–29.9	39	27.85	46.2	39	27.85	66.7	38	27.89	29.0	38	27.83	36.8
8	20–24.9	77	22.83	63.6	77	22.83	90.9	75	22.83	42.7	76	22.83	54.0
9	< 20	15	19.03	53.3	15	19.03	93.3	15	19.03	66.7	15	19.03	73.3

^a% rating overall quality of life as *fair* or *poor*.

^b% reporting *could not do* or were limited *quite a lot* in usual activities during the past week.

^c% reporting number of days in bed due to illness or injury during the past 4 weeks as being 1 *SD* or more above the mean for the general population (mean = 1.00, *SD* = 3.35).

^d% reporting current work status as *disabled*.

Table 9.45

Percentage of Adults Reporting Significant Illness or Injury and Limitations Due to Physical Conditions at 8 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Physical Scores, 2009 U.S. General Population

RP T Scores		Days in bed due to illness/ injury significantly above the mean ^a			Current employment status is <i>disabled</i> ^b			Could not do or had quite a lot of difficulty doing usual activities due to physical conditions ^c			Chronic condition(s) limit usual activities/ employment <i>moderately</i> , <i>quite a lot</i> , or <i>extremely</i> ^d		
Level	Range	<i>n</i>	(1) Mean	%	<i>n</i>	(2) Mean	%	<i>n</i>	(3) Mean	%	<i>n</i>	(4) Mean	%
1	55+	1,017	57.12	2.8	1,019	57.12	0.8	1,020	57.12	0.9	1,025	57.12	6.5
2	50–54.9	356	52.85	6.2	361	52.85	2.8	359	52.84	2.0	361	52.85	18.6
3	45–49.9	204	47.52	11.8	205	47.53	5.9	203	47.54	5.9	205	47.53	35.6
4	40–44.9	97	42.87	13.4	100	42.83	10.0	101	42.83	11.9	102	42.83	46.1
5	35–39.9	163	37.95	22.1	165	37.94	18.2	166	37.94	25.3	166	37.94	65.7
6	30–34.9	96	31.50	45.8	96	31.50	43.8	96	31.50	64.6	96	31.50	79.2
7	25–29.9	48	27.37	50.0	49	27.36	38.8	48	27.38	70.8	49	27.36	81.6
8	< 25	52	22.36	44.2	52	22.41	69.2	53	22.40	86.8	53	22.40	86.8

^a% reporting number of days in bed due to illness or injury during the past 4 weeks as being 1 *SD* or more above the mean for the general population (mean = 1.00, *SD* = 3.35).

^b% reporting current work status as *disabled*.

^c% reporting *could not do* or were limited *quite a lot* in usual activities due to physical conditions during the past week.

^d% reporting one or more chronic condition(s) ever been told he/she had that limit usual activities/enjoyment *moderately*, *quite a lot*, or *extremely*.

Table 9.46

Percentage of Adults Reporting Significant Chronic Conditions, Treatment, and Disability at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Bodily Pain Scores, 2009 U.S. General Population

BP T Scores		Number of chronic conditions significantly above the mean ^a			Number of hospital stays significantly above the mean ^b			Current employment status is <i>disabled</i> ^c		
Level	Range	<i>n</i>	(1) Mean	%	<i>n</i>	(2) Mean	%	<i>n</i>	(3) Mean	%
1	60+	541	60.87	2.4	540	60.87	1.7	540	60.87	0.7
2	55–59.9	1	56.13	0.0	1	56.13	0.0	1	56.13	0.0
3	50–54.9	664	53.43	7.1	661	53.42	2.4	666	53.43	2.7
4	45–49.9	351	47.22	14.0	347	47.23	1.7	349	47.22	5.7
5	40–44.9	146	41.60	28.8	146	41.60	4.8	146	41.60	11.6
6	35–39.9	150	37.58	35.3	151	37.58	7.3	151	37.58	20.5
7	30–34.9	153	31.83	46.4	153	31.83	12.4	152	31.83	34.9
8	25–29.9	23	26.36	43.5	22	26.38	18.2	22	26.38	50.0
9	< 25	20	21.39	45.0	19	21.39	36.8	20	21.39	70.0

^a% reporting the number of chronic conditions ever been told he/she had as being 1 *SD* or more above the mean for the general population (mean = 2.27, *SD* = 2.23).

^b% reporting number of hospital stays during the past 12 months as being 1 *SD* or more above the mean for the general population (mean = 0.23, *SD* = 0.89).

^c% reporting current work status as *disabled*.

General Health (GH)

Quality of life, general health, and disability. As shown in Table 9.47, a perfect or near perfect ordering of decreasing GH scale scores with increasing percentages of respondents who rated their overall quality of life as *fair* or *poor* (Column 1), rated their health as *poor* or *very poor* (Column 2) and as being significantly below the general population mean on a 0–100 scale (Column 3), and reported being disabled (Column 4). It is interesting to note that among those with a GH Level 4 score (*T*-score range = 50.0–54.9), only 0.9% rated their health as *fair* or *poor* whereas

25.3% indicated a numerical health rating that was at least 1 *SD* below the general population mean. At Level 5 (*T*-score range = 45.0–49.9), these percentages increased to 2.4% and 42.2%, respectively, and finally became equivalent (94.1% and 93.8%, respectively) at Level 10 (*T*-score range < 25). With the exception of health rated as *poor* or *very poor* at the highest score levels, all criteria are useful in interpreting score differences at all GH score levels.

Chronic conditions, missed workdays, and treatment. When examining the associations of GH scores with number of recent outpatient visits (Column 1),

Table 9.47

Percentage of Adults Reporting Quality of Life and General Health Problems and Disability at 10 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scores, 2009 U.S. General Population

GH T Scores		Overall quality of life rated as <i>fair</i> or <i>poor</i> ^a (1)			Health rated as <i>poor</i> or <i>very poor</i> ^b (2)			Health rating significantly below the mean ^c (3)			Current employment status is <i>disabled</i> ^d (4)		
Level	Range	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	65+	84	65.40	0.0	82	65.40	0.0	84	65.40	1.2	84	65.40	0.0
2	60–64.9	211	62.23	0.5	211	62.23	0.0	211	62.23	2.4	211	62.23	1.0
3	55–59.9	585	57.24	1.2	582	57.23	0.0	583	57.24	8.2	586	57.25	0.9
4	50–54.9	346	52.15	5.2	346	52.15	0.9	344	52.15	25.3	344	52.17	2.0
5	45–49.9	254	47.62	13.4	255	47.61	2.4	251	47.60	42.2	252	47.60	4.8
6	40–44.9	209	43.10	21.1	210	43.09	4.8	207	43.10	66.2	208	43.10	13.9
7	35–39.9	183	38.09	33.9	183	38.09	9.8	182	38.09	84.1	181	38.09	17.1
8	30–34.9	128	32.23	58.6	128	32.23	39.1	127	32.23	91.3	128	32.23	40.6
9	25–29.9	37	26.92	78.4	37	26.92	51.4	37	26.92	97.3	37	26.92	51.4
10	< 25	17	22.49	82.4	17	22.49	94.1	16	22.53	93.8	17	22.49	64.7

^a% rating overall quality of life as *fair* or *poor*.

^b% rating health as *poor* or *very poor* during the past week.

^c% rating health 1 *SD* or more below the general population mean 0–100 rating during the past 4 weeks (mean = 78.17, *SD* = 19.30).

^d% reporting current work status as *disabled*.

Table 9.48

Percentage of Adults Reporting Significant Chronic Conditions, Missed Workdays, and Treatment at 10 Levels of SF-36v2 Acute (1-Week Recall) Form General Health Scores, 2009 U.S. General Population

GH T Scores		Number of outpatient visits significantly above the mean ^a (1)			Missed workday due to illness/injury ^b (2)			Number of chronic conditions significantly above the mean ^c (3)			Chronic condition(s) limits usual activities/enjoyment <i>moderately, quite a lot, or extremely</i> ^d (4)		
Level	Range	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	65+	84	65.40	7.1	54	65.40	0.0	84	65.40	0.0	84	65.40	0.0
2	60–64.9	211	62.23	10.0	127	62.20	3.2	211	62.23	1.4	211	62.23	4.3
3	55–59.9	582	57.24	13.8	332	57.26	3.6	585	57.24	3.9	587	57.25	9.5
4	50–54.9	346	52.15	14.5	211	52.12	6.2	345	52.17	11.0	346	52.15	13.0
5	45–49.9	252	47.60	21.4	131	47.63	6.9	255	47.61	14.1	258	47.61	25.2
6	40–44.9	208	43.11	29.8	101	43.19	9.9	207	43.09	22.7	210	43.09	41.9
7	35–39.9	181	38.10	32.6	65	38.15	13.9	183	38.09	25.1	183	38.09	60.7
8	30–34.9	128	32.23	45.3	25	31.86	28.0	128	32.23	51.6	128	32.23	80.5
9	25–29.9	37	26.92	56.8	4	27.91	0.0	37	26.92	67.6	37	26.92	91.9
10	< 25	17	22.49	58.8	4	21.92	50.0	17	22.49	58.8	17	22.49	82.4

^a% reporting number of outpatient visits during past 4 weeks as being 1 *SD* or more above the mean for the general population (mean = 0.89, *SD* = 1.73).

^b% reporting one or more days of missed work because of illness or injury during past 4 weeks.

^c% reporting the number of chronic conditions ever been told he/she had as being 1 *SD* or more above the mean for the general population (mean = 2.27, *SD* = 2.23).

^d% reporting one or more chronic condition(s) ever been told he/she had that limit usual activities/enjoyment *moderately, quite a lot, or extremely*.

number of missed workdays due to illness or injury (Column 2), number of chronic conditions (Column 3), and limitations imposed by chronic conditions (Column 4) presented in Table 9.48, a general ordering of increasing percentages of those reporting problems with decreasing levels of GH scores is apparent. Note that the unexpected finding that none of the respondents at score Level 9 reported missed workdays due to illness or injury is likely due to there being only four respondents at this GH score level. Overall, all four variables addressed in

Table 9.48 are useful in interpreting score differences across all GH score levels.

Vitality (VT)

Table 9.49 shows a perfect or near perfect linear increase in the percentage of respondents who reported *little* or *no* energy (Column 2) and who rated their overall quality of life as *fair* or *poor* (Column 1) with decreasing VT scale score levels. Overall, both criterion variables are useful in interpreting score differences at the middle

Table 9.49

Percentage of Adults Reporting Quality of Life and Level of Energy Problems at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Vitality Scores, 2009 U.S. General Population

VT T Scores		Overall quality of life rated as fair or poor ^a (1)			Little or no energy ^b (2)		
Level	Range	n	Mean	%	n	Mean	%
1	65+	93	67.93	0.0	93	67.93	0.0
2	60–64.9	238	61.77	1.3	238	61.77	0.4
3	55–59.9	519	57.16	1.0	520	57.15	0.6
4	50–54.9	430	51.38	4.9	430	51.38	3.7
5	45–49.9	178	47.38	16.3	176	47.38	10.8
6	40–44.9	249	43.52	22.1	250	43.51	23.2
7	35–39.9	193	38.00	38.3	193	38.00	49.7
8	30–34.9	97	32.45	59.8	96	32.43	85.4
9	< 30	55	26.80	69.1	55	26.80	92.7

^a% rating overall quality of life as fair or poor.

^b% reporting amount of energy as a little or none during the past week.

and lower score levels.

Social Functioning (SF)

Similar to the data found in the previous table, Table 9.50 reveals a perfect of near perfect linear increase in respondents who reported they *could not do* or were limited *quite a lot* in usual social activities due to physical health or emotional problems (Column 2) and ratings of overall quality of life as fair or poor (Column 1) with decreasing SF scale score levels. Overall, both criterion

Table 9.50

Percentage of Adults Reporting Quality of Life Problems and Limitations in Social Activities at 9 Levels of SF-36v2 Acute (1-Week Recall) Form Social Functioning Scores, 2009 U.S. General Population

SF T Scores		Overall quality of life rated as fair or poor ^a (1)			Could not do or was limited quite a lot in usual social activities due to physical health/emotional problems ^b (2)		
Level	Range	n	Mean	%	n	Mean	%
1	55+	1,229	56.74	3.2	1,227	56.74	0.2
2	50–54.9	211	51.79	11.4	211	51.79	0.5
3	45–49.9	192	46.85	14.1	190	46.85	2.6
4	40–44.9	138	41.91	37.0	138	41.91	7.3
5	35–39.9	122	36.97	34.4	122	36.97	25.4
6	30–34.9	58	32.03	53.5	58	32.03	37.9
7	25–29.9	53	27.08	62.3	53	27.08	77.4
8	20–24.9	29	22.14	69.0	29	22.14	89.7
9	< 20	20	17.20	80.0	20	17.20	100.0

^a% rating overall quality of life as fair or poor.

^b% reporting they *could not do* or were limited *quite a lot* in usual social activities due to physical health/emotional problems during the past week.

variables are useful in interpreting score differences at all score levels.

Role-Emotional (RE)

Quality of life, happiness, stress, and emotional problems. As shown in Table 9.51, a near perfect ordering exists of increasing percentages of respondents reporting emotional and HRQOL problems with decreasing scores over the 10 RE score levels. Particularly notable amongst the findings is the pervasiveness of experiencing a *good bit, quite a bit, or a great deal* of stress or pressure in daily living (Column 3), even at the highest RE score levels. For example, 23.9% of the respondents at Level 1 and 52.9% at Level 2 reported this problem. Feeling only *sometimes fairly satisfied* or *generally dissatisfied* with one's personal life (Column 2) was similarly problematic, with 8.5% and 30.4% reporting these feelings at RE score Levels 1 and 2, respectively. With the exception of the stress variable (Column 3) at the lowest score levels, all criteria are useful in interpreting score differences at all RE score levels.

Job performance and health problems. From the highest to the lowest RE score levels, Table 9.52 shows a high and increasing percentage of respondents who reported relatively low ratings of job performance (Column 1) and health (Column 2) and being *moderately, quite a lot, or extremely* limited in usual activities or enjoyment due to a chronic condition (Column 3). At RE score Levels 1 and 2, respectively, 14.5% and 15.0% reported relatively low job performance ratings, 21.6% and 44.6% reported relatively low health ratings, and 14.6% and 31.4% indicated significant limitations in activities or enjoyment due to chronic conditions. Generally, the health rating and chronic condition limitations criteria are useful in interpreting score differences across all score levels, whereas the job performance variable is useful at all but the lowest score level.

Mental Health (MH)

Depression and anxiety. As expected, Table 9.53 reveals a strong, while not perfect, linear relationship between decreasing MH scale score levels and reports of depression (Column 1) and its common comorbidity, anxiety (Column 4), as current chronic conditions. Similar findings are noted for those respondents who reported feeling down, depressed, or hopeless (Column 2) and those who reported having felt little interest or pleasure in doing things (Column 3) *more than half the days* or *nearly every day* during the weeks preceding reassessment.

Note that the relatively low percentage (33.3%) of those at the lowest MH score level who reported experiencing depression and/or anxiety (Column 1 and

Table 9.51

Percentage of Adults Reporting Emotional Problems and Problems Related to Quality of Life, Happiness, and Stress at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scores, 2009 U.S. General Population

RE T Scores		Overall quality of life rated as fair or poor ^a			Happiness/satisfaction with personal life rated as sometimes fairly satisfied or generally dissatisfied ^b			Experienced a good bit, quite a bit, or a great deal of stress/pressure in daily living ^c			Could not do or was limited quite a lot in usual work, school, or other daily activities due to personal/emotional problems ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	1,384	55.64	5.1	1,380	55.64	8.5	1,379	55.64	23.9	1,382	55.64	0.9
2	50–54.9	102	51.82	19.6	102	51.82	30.4	102	51.82	52.9	102	51.82	4.9
3	45–49.9	166	48.01	22.9	165	48.01	40.6	166	48.01	55.4	166	48.01	3.6
4	40–44.9	179	42.92	24.0	179	42.92	43.0	179	42.92	63.7	179	42.92	7.3
5	35–39.9	44	36.55	38.6	44	36.55	50.0	43	36.55	69.8	44	36.55	13.6
6	30–34.9	80	32.74	33.8	78	32.74	52.6	80	32.74	73.8	78	32.74	16.7
7	25–29.9	37	26.98	64.9	37	26.98	67.6	37	26.98	67.6	35	26.86	48.6
8	20–24.9	29	21.29	65.5	29	21.29	75.9	28	21.29	85.7	29	21.29	65.5
9	15–19.9	6	17.47	83.3	6	17.47	83.3	6	17.47	83.3	6	17.47	66.7
10	< 15	25	11.27	80.0	24	11.30	79.2	25	11.27	84.0	25	11.27	80.0

^a% rating overall quality of life as fair or poor.

^b% rating happiness/satisfaction with personal life as sometimes fairly satisfied or generally dissatisfied during the past 4 weeks.

^c% reporting having experienced a good bit, quite a bit, or a great deal of stress/pressure in daily living during the past 4 weeks.

^d% reporting they could not do or were limited quite a lot in usual work, school, or other daily activities due to personal/emotional problems.

Table 9.52

Percentage of Adults Reporting Poor Job Performance and Health Problems at 10 Levels of SF-36v2 Acute (1-Week Recall) Form Role-Emotional Scores, 2009 U.S. General Population

RE T Scores		Rating of overall job performance significantly below the mean ^a			Health rating significantly below the mean ^b			Chronic condition(s) limit usual activities/enjoyment moderately, quite a lot, or extremely ^c		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%
1	55+	766	55.64	14.5	1,379	55.64	21.6	1,388	55.64	14.6
2	50–54.9	60	51.82	15.0	101	51.82	44.6	102	51.82	31.4
3	45–49.9	73	48.01	27.4	166	48.01	48.2	166	48.01	35.5
4	40–44.9	86	42.74	38.4	178	42.92	60.7	179	42.92	43.0
5	35–39.9	17	36.55	64.7	44	36.55	68.2	44	36.55	63.6
6	30–34.9	27	32.74	37.0	79	32.74	74.7	81	32.74	55.6
7	25–29.9	13	26.93	92.3	36	26.97	83.3	37	26.98	75.7
8	20–24.9	6	21.29	66.7	27	21.29	88.9	29	21.29	86.2
9	15–19.9	1	17.47	100.0	6	17.47	100.0	6	17.47	83.3
10	< 15	1	9.84	100.0	25	11.27	92.0	25	11.27	92.0

^a% rating overall job performance as 1 SD or more below the mean 0–10 rating for the general population during the past 4 weeks (mean = 8.33, SD = 1.50).

^b% rating health 1 SD or more below the general population mean 0–100 rating during the past 4 weeks (mean = 78.17, SD = 19.30).

^c% reporting one or more chronic condition(s) ever been told he/she had that limit usual activities/enjoyment moderately, quite a lot, or extremely.

Table 9.53

Percentage of Adults Reporting Problems With Depression and Anxiety at 11 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scores, 2009 U.S. General Population

MH T Scores		Depression is a current chronic condition ^a			Feeling down/ depressed/ hopeless more than half the days or nearly every day ^b			Felt little interest/ pleasure in doing things more than half the days or nearly every day ^c			Anxiety is a current chronic condition ^d		
Level	Range	n	Mean	%	n	Mean	%	n	Mean	%	n	Mean	%
1	60+	4	60.93	1.5	277	61.45	0.0	276	61.44	0.7	2	61.42	0.7
2	55–59.9	30	56.25	3.9	763	56.74	0.1	767	56.74	1.3	39	56.59	5.1
3	50–54.9	41	51.70	12.1	340	51.66	1.8	338	51.68	6.8	47	51.83	13.9
4	45–49.9	42	46.45	17.7	237	46.65	5.5	238	46.65	8.8	59	46.67	24.9
5	40–44.9	64	41.54	37.2	173	41.64	16.8	176	41.65	32.4	68	41.59	39.3
6	35–39.9	51	36.94	47.7	112	36.93	36.6	112	36.93	45.5	51	36.94	47.7
7	30–34.9	35	31.70	53.0	66	31.86	53.0	66	31.86	56.1	30	31.73	46.2
8	25–29.9	36	26.96	81.8	44	27.09	86.4	44	27.09	75.0	34	26.97	77.3
9	20–24.9	16	21.41	88.9	18	21.76	88.9	18	21.76	66.7	17	21.78	94.4
10	15–19.9	7	17.31	87.5	8	17.40	100.0	8	17.40	87.5	7	17.31	87.5
11	< 15	3	13.12	33.3	3	13.12	100.0	3	13.12	100.0	3	13.12	33.3

^a% reporting depression as a current chronic condition.

^b% reporting feeling down/depressed/hopeless more than half the days or nearly every day during the past 2 weeks.

^c% reporting felt little interest/pleasure in doing things more than half the days or nearly every day during the past 2 weeks.

^d% reporting anxiety as a current chronic condition.

4) were unexpected but likely due to there being only three respondents at this MH score level.

Quality of life, happiness, and stress. Continuing to demonstrate the connection between MH scores and emotional health and well-being, Table 9.54 shows the perfect or near perfect ordering of decreasing MH score levels with increasing percentages of respondents rating their overall quality of life as *fair* or *poor* (Column 1), being *sometimes fairly satisfied* or *generally dissatisfied* with their personal lives (Column 2), experiencing significant stress in daily living (Column 3), and having stress significantly affect their health (Column 4). In general, all of these problems were experienced by most, if not all, of those respondents scoring at the lowest three MH score levels. As such, all the criterion variables are considered most useful for interpreting score differences at the highest and middle score levels.

Interpolation of Score-Related Percentages

Only the score ranges and the means within those ranges for the component summary measure and health domain scale scores are reported in Tables 9.1 through 9.54. Therefore, users must calculate ratios of differences and interpolate to estimate the percentage that is associated with a specific score within a given score range and to determine the percentage difference between two scores. The processes for estimating the percentages associated with specific scores and for determining differences within and across levels are the same as those used with the content-based interpretation tables and can be found in Chapter 8 of this manual.

Table 9.54

Percentage of Adults Reporting Quality of Life, Happiness, and Stress at 11 Levels of SF-36v2 Acute (1-Week Recall) Form Mental Health Scores, 2009 U.S. General Population

MH T Scores		Overall quality of life rated as <i>fair</i> or <i>poor</i> ^a			Happiness/satisfaction with personal life rated as <i>sometimes fairly satisfied</i> or <i>generally dissatisfied</i> ^b			Experienced <i>a good bit, quite a bit, or a great deal</i> of stress/pressure in daily living ^c			Stress/pressure has affected health <i>moderately, quite a lot, or extremely</i> ^d		
Level	Range	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%	<i>n</i>	Mean	%
1	60+	279	61.44	2.2	276	61.44	0.7	277	61.44	6.9	278	61.44	1.4
2	55–59.9	768	56.74	2.5	766	56.73	3.0	765	56.73	18.3	764	56.73	4.5
3	50–54.9	342	51.66	9.7	341	51.66	14.1	341	51.66	37.5	343	51.65	11.1
4	45–49.9	238	46.65	16.0	238	46.65	31.9	237	46.65	57.4	238	46.65	25.6
5	40–44.9	176	41.65	31.3	175	41.64	56.6	176	41.65	69.3	174	41.61	43.7
6	35–39.9	112	36.93	37.5	110	36.92	60.0	111	36.94	71.2	111	36.94	47.8
7	30–34.9	66	31.86	53.0	66	31.86	69.7	66	31.86	92.4	66	31.86	56.1
8	25–29.9	44	27.09	79.6	44	27.09	90.9	44	27.09	93.2	44	27.09	86.4
9	20–24.9	18	21.76	61.1	18	21.76	88.9	18	21.76	100.0	17	21.71	88.2
10	15–19.9	8	17.40	100.0	8	17.40	100.0	8	17.40	100.0	8	17.40	100.0
11	< 15	3	13.12	66.7	3	13.12	100.0	3	13.12	100.0	3	13.12	100.0

^a% rating overall quality of life as *fair* or *poor*.

^b% reporting happiness/satisfaction with personal life rated as *sometimes fairly satisfied* or *generally dissatisfied* during the past 4 weeks.

^c% reporting having experienced *a good bit, quite a bit, or a great deal* of stress/pressure in daily living during the past 4 weeks.

^d% reporting stress/pressure has affected health *moderately, quite a lot, or extremely* during the past 4 weeks.



10

Determining Important Differences in Scores

Interpretation of scores and score differences has become a focus area in health status assessment. Researchers and clinicians have typically used the concept of *minimally important difference*, or *MID*, to differentiate between an important score difference and a trivial difference. While recent years have witnessed a rich literature on MID, a number of conceptual and empirical questions still remain: Does the MID concept apply to mean group differences or to differences in individual respondent scores? Should different MID standards apply to score differences at a single point in time and to changes in the score over time? Should different MID standards be applied to improvement and to decline? Are MIDs dependent on the score range? Are MIDs dependent on the disease group? What is the best method for determining MIDs? A thorough discussion of these issues is beyond the scope of this manual and their final resolution awaits further research. That said, this chapter summarizes the perspective taken by the authors of this manual.

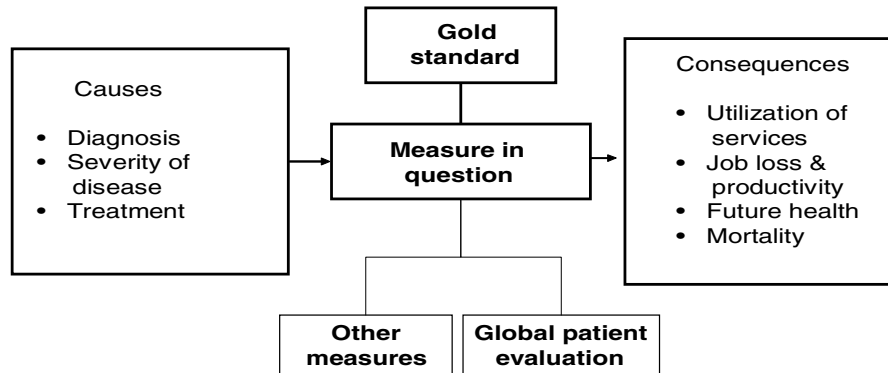
An important motivation underlying the MID concept is the proper design of clinical trials. Given a sufficiently large sample size, even trivial differences can show statistical significance. Therefore, the central question facing researchers when analyzing data is whether a particular difference is clinically significant. From this question came the concept of *minimal clinically important difference*. This term was shortened to *minimally important difference* in the context of patient-reported outcomes (PROs) to emphasize the perspective of the patient rather than the clinician and the importance of evidence of many types, including but not limited to clinical evidence. Thus, from the perspective of clinical studies, the focus is the minimal group difference that a given study has power to detect. For the purposes of this manual, this minimal mean group difference is referred to as *MID*.

For other endpoints, it is well recognized that the MID for a comparison of group mean scores is different

from the MID for individual respondent scores (Guyatt, Osoba, Wu, Wyrwich, & Norman, 2002). However, in much of the literature on PRO measures, this distinction is not made explicit, leading to some confusion regarding recommendations. Following suggestions from the Food and Drug Administration's (FDA's) guidance for the pharmaceutical industry regarding PRO measures (U.S. Department of Health and Human Services, 2009), this manual will use the concept of a *responder* in relation to an individual who has shown important change in his or her score over time. An MID is generally smaller than the responder definition for the same scale, partly because the magnitude of group mean differences is not affected by measurement precision (discussed in a later section) and partly because it is not reasonable to assume that everyone in a clinical trial will benefit from treatment. This chapter first reviews information that is relevant to the evaluation of MID (i.e., minimally important mean group differences in either cross-sectional or longitudinal analyses) and then follows with a review of information concerning the importance of score changes for individual respondents.

General Considerations for Determining Minimally Important Differences (MIDs)

Figure 10.1 presents the general logic by which the meaningfulness of differences in SF-36v2 scores has been derived. At the left of the figure are common causes or health events that can impact the observed results on measures that are sensitive to variations in health status. Understanding the impact of these variations comes not only from knowledge of the instrument being employed but also from its relationship with a "gold standard" and with other instruments that measure the same constructs

Figure 10.1 Model for the Analysis of the Meaningfulness of Differences in SF-36v2 Scores

and are sensitive to these same causes. At the right of Figure 10.1 are differences in the consequences of health status variations that are associated with differences in scores, such as utilization of health care services, job loss and productivity, future health, and mortality. These consequences thus provide another source of interpretive material (Keller & Ware, 1995).

Whenever possible, the importance of differences or changes in health status should be determined from a variety of perspectives, including but not limited to clinical judgments. Patient-based assessments of functional health, well-being, and other health status constructs are utilized in clinical research studies because of what they add to the understanding of patient outcomes beyond traditional clinical measures of disease severity and treatment response. Accordingly, the significance or importance of score differences should be defined in terms of their importance to the patient and to society, in addition to their importance from a more clinical perspective.

Several types of empirical evidence should be considered when judging the importance of a health status outcome. Thus, in the broader perspective regarding MID, a difference or change in health status is important when it:

- is associated with noteworthy differences in clinical markers,
- forecasts substantial changes in health-related events (e.g., disability, job loss, work productivity, hospitalization, death), and/or
- is associated with a change in patients' evaluations of their health.

In addition, it is also important to determine that the difference is unlikely to be due to chance or random error. In the comparison of group means, this issue is handled through statistical tests of mean differences. The precision of the scores will affect the standard error of

the mean group differences, but not the magnitude of the differences. Thus, measurement precision should not be considered in determining MID for clinical trial purposes because this precision will be taken into account through the standard error term. In contrast, measurement precision is an important consideration in defining the level beyond which an individual respondent is considered to have changed.

Criterion- or Anchor-Based Approaches to MID

Criterion-based techniques “examine the relationship between scores on the instrument whose interpretation is under question (the target instrument) and some independent measure (an anchor)” (Guyatt et al., 2002, p. 373). Such anchors, or criteria, can be specific diseases, clinical markers, and/or health-related events. By examining the score differences associated with particular differences in an established clinical criterion (e.g., comparing scores for patients with a specific disease to the scores of a health comparison group), researchers have been able to specify thresholds to demarcate the difference between an important change and a trivial change, thereby establishing MID. A useful anchor may consist of a given respondent's status on an easily understood measure, such as mobility defined as the difference between using a wheelchair and walking with an aid (see Ware & Keller, 1996).

Data on score differences as predictors of health-related events such as hospitalization, job loss, inability to work, and mortality can be found in Ware et al. (2007) and suggest that MID may in fact depend on the score level. For example, VT score differences above the mean score of 50 were not associated with increased 2-year mortality risks, whereas score differences below the mean were highly associated with mortality (Bjorner et al., 2007). In general, the largest risk differences are seen for low score levels, suggesting score differences

of a certain magnitude may be more important for the lower score levels. This dependence on score level may explain the occasional finding that MIDs for improvement are different from MIDs for deterioration. For example, if an improvement from mean of 40 to mean of 45 is compared to a deterioration from 40 to 35, two different score ranges are, in effect, being compared.

Once the score level is taken into consideration, large variations in MID by disease group have not been found. To illustrate, public-use data files containing SF-36 results from 519,035 respondents in the Medicare Health Outcomes Survey (HOS; Ware, Gandek, Sinclair, & Kosinski, 2004) were analyzed. Criterion-based analysis was separately performed for each of 14 different diseases to test the stability of MIDs across disease groups. Note that the very large sample sizes in the HOS made these very robust analyses. MID for the PCS measure was evaluated using logistic regression analyses with PCS as the independent variable and 2-year mortality as the dependent variable (i.e., the criterion). Furthermore, the PCS score differences associated with a 20% increase in mortality risk were calculated based on logistic regression results (see Table 10.1). With one exception (CHF), the results showed remarkable similarity across diseases, supporting an MID of about 3 points for PCS when a 20% increase in mortality risk is regarded as significant. Analyses using a 50% increase in mortality risk as the threshold for minimal importance demonstrated the same kind of stability across disease groups.

These results support the use of general MID criteria across disease groups. However, it may be relevant to

support MID recommendations for a particular patient group by including anchors that are specific to said group. Such supplementary information can often be attained through a systematic literature review (see Spiegel et al., 2005). When summarizing the literature, care must be taken to identify publications that provide MID recommendations using a 0–100 scoring metric. To be comparable to the recommendations found in this manual, such scores must first be transformed to the *T*-score metric (i.e., divide the 0–100 score by that metric's scale standard deviation, and multiply by 10). Approximate values for the 0–100 metric standard deviations (*SDs*) in the general population can be found in Ware et al. (2007).

Some anchors may involve concurrent measurement (cross-sectional data), while others may require measurement through time (longitudinal data). A popular and useful anchor method relies on patients' evaluations of change. From this perspective, changes in health should be considered important when they are large enough to change patients' own evaluations of their health status. When using this method, the obtained change in the target measure that is associated with patients' own evaluations of change is quantified to define MID (see Angst, Aeschlimann, & Stucki, 2001; Carreon, Glassman, Campbell, & Anderson, 2010; Colangelo, Pope, & Peschken, 2009; Guyatt, Berman, Townsend, Pugsley, & Chambers, 1987; Guyatt et al., 2002; Kosinski, Zhao, Dedhiya, Osterhaus, & Ware, 2000; Lauridsen, Hartvigsen, Manniche, Korsholm, & Grunnet-Nilsson, 2006; Sekhon, Pope, & Baron, 2010; Ware, Snow, Kosinski, & Gandek, 1993).

Table 10.1

Physical Component Summary Measure Score Differences as Predictors of Mortality at 2-Year Follow-Up, Medicare Health Outcomes Survey (N = 519,035)

Disease Group	<i>n</i>	(Deaths)	Beta	MID for PCS	
				Based on Increase in Mortality 20% Increase	50% Increase
Depression	37,618	(5,499)	−0.062	2.9	6.5
Acute myocardial infarction	48,206	(7,759)	−0.054	3.4	7.5
Angina or coronary artery disease	72,153	(10,033)	−0.057	3.2	7.2
Any cancer	59,477	(9,781)	−0.062	2.9	6.5
Arthritis, hand or wrist	161,596	(14,437)	−0.060	3.0	6.7
Arthritis, hip or knee	185,084	(16,438)	−0.059	3.1	6.9
Congestive heart failure (CHF)	31,325	(8,478)	−0.044	4.1	9.2
Chronic obstructive pulmonary disease (COPD)	59,733	(9,200)	−0.056	3.2	7.2
Diabetes	81,439	(10,033)	−0.053	3.4	7.6
Gastrointestinal problems	25,364	(2,688)	−0.055	3.3	7.4
High blood pressure	255,713	(23,675)	−0.056	3.2	7.2
Other heart conditions	97,831	(12,746)	−0.059	3.1	6.9
Sciatica	109,436	(9,118)	−0.057	3.2	7.1
Stroke	39,049	(7,453)	−0.051	3.6	8.0

In some cases, the SF-36v2 Self-Evaluated Transition (SET) item (Item 2) can serve as an indicator of self-perceived change. On the standard form, this item asks, “Compared to 1 year ago, how would you rate your health in general now?” For the SF-36v2 acute form, the recall period is 1 week for the same question. For both forms, the five-level rating scale for this item ranges from *much better* to *much worse*. Because the SET item is not used to score any of the health domain scales or component summary measures, the information it provides can easily be overlooked. However, it can serve as an important source of additional information regarding a given respondent’s self-perceived change in health status. For example, on the standard form, a response of either *much better* or *somewhat better* would generally be expected from a respondent with a chronic illness whose health has improved. Similarly, on the acute form, these same responses would be expected when a respondent has an acute condition with a typically rapid recovery time. Alternately, the SET item can be used as a template for developing an item that is more specific to a given respondent and his or her circumstances, which would then be administered *in addition to* the standardized SF-36v2 SET item. This anchor method is critically dependent on the validity of a patient’s rating of change (Guyatt et al., 2002). Specifically, this strategy assumes that patients can judge whether they are the same, better, or worse after a specified time period. Further, estimation of MID can be biased if the evaluation of change is affected by other factors than the scale in question. It is therefore recommended that inferences made from MIDs based on this approach be corroborated by other methods (Guyatt et al., 2002).

The literature is divided on whether an MID for cross-sectional data is equivalent to an MID for change over time (sometimes referred to as *minimally important change*, or *MIC*). If, for a particular scale, the MID for change over time were different from the MID for cross-sectional differences, then cross-sectional data would not be useful for establishing the former. However, it does not appear that a compelling case has yet been made for MID generally being different between cross-sectional and longitudinal analyses. Thus, the two will be treated together here.

Regardless of whether researchers rely on a single anchor or multiple anchors or whether they use cross-sectional or longitudinal analyses, all anchor-based approaches have two requirements: (a) the anchor must be interpretable, and (b) a notable association must exist between the target and the anchor (Guyatt et al., 2002). The first requirement demonstrates the need for an anchor to present face-valid interpretation or to be

articulated in terms of interpretation guidelines. The second requirement underscores the importance of anchor selection. To the extent that an anchor is not related to the target measure, it will not offer inferences about the interpretation of the chosen target and will likely produce misleading results. In general, the stronger the association between the target and the anchor, the more fruitful the resulting interpretations of the target measure will be. This becomes even more salient when using a single anchor; to generate convincing results, a higher degree of association between the target and the anchor is necessary than when multiple anchors are used (Guyatt et al., 2002).

Distribution-Based Approaches To MID

Distribution-based techniques define MID based on the distribution of the scores themselves. These methods interpret results in terms of the relationship between the magnitude of difference and some measure or measures of variability. The differences can be defined as (a) within-group longitudinal comparisons before and after treatment or as (b) the difference in mean scores between two groups. The measure of variability relevant for group-based MIDs is between-patient variability, such as the *SD* of respondents at baseline. For the individual responder definition (see following section), another relevant difference measure is the within-patient change over time and additional measures of variability: within-patient variability (the *SD* of change that respondents experienced during a study) and the standard error of measurement (*SEM*; see following discussion).

An early and often cited criterion for important group mean changes was based on *effect size*: the mean change divided by the baseline *SD* (Cohen, 1988). It has been suggested that a small effect size was 0.2 (equivalent to 2 *T*-score points on the SF-36v2 health domain scale and component summary measure scores in the general population), a median effect size was 0.5 (equivalent to 5 *T*-score points), and a large effect size was 0.8 (equivalent to 8 *T*-score points) in the context of comparing group averages (Cohen, 1988). This approach would lead to a suggested MID of 2 *T*-score points for all SF-36v2 scale and summary measure scores. However, this approach has been criticized in that it seems arbitrary (Guyatt et al., 2002). Other researchers have found a convergence of evidence suggesting 0.5 *SD* as a relevant threshold for importance (Norman, Sloan, & Wyrwich, 2003), but their discussion suggests that this rule was intended for differences in individual respondent scores (referred to here as a *responder definition* and discussed later in this chapter). Thus, there is an emerging consensus that a criterion of 0.5 *SD* (equivalent

to 5 *T*-score points) is too high an MID for group mean comparisons.

MID Criteria in Relation to Individual SF-36v2 Measures and Scales

This section will focus on the implications of an MID of 3 *T*-score points for the SF-36v2 component and scale scores in relation to various criteria/anchors. The discussion will elaborate on the criteria presented in Chapter 9 and assumes that the baseline score for the group to be evaluated is lower than the general population average (e.g., in the 30–40 *T*-score range). An MID of 3 *T*-score points was chosen as a starting point for this discussion because it represents a compromise between the results from different distribution-based approaches to MID (from 2 *T*-score points [Cohen, 1988] to 5 *T*-score points [Norman et al., 2003]).

Physical Component Summary (PCS)

As a summary measure of physical health, PCS is associated with a wide range of conditions and outcomes. For example, a 3-points lower PCS *T*-score is associated with an odds ratio (OR) of 1.43 for being unable to work (i.e., approximately 40% higher risk), an OR of 1.25 for job loss in the following year (for the employed population; i.e., approximately 25% higher risk), and an OR of 1.15 of being hospitalized in the subsequent year (i.e., approximately 15% higher risk). Among the Medicare population, a 3-points lower *T* score implies approximately 20% higher 1-year mortality risk (OR = 1.19–1.22 in the 25–50 *T*-score range, with higher OR for the low scoring groups). Using the 2009 general population data and self-reported diseases, a 3-point threshold for importance would imply that the unique disease burdens (controlled for other diseases) of diabetes, congestive heart failure, chronic obstructive pulmonary disease (COPD), arthritis, back pain, stroke, and limited use of arms or legs are significant for PCS, while the unique impact of conditions like anemia, asthma, migraine headaches, and depression would not be minimally important for PCS. Further, other conditions such as HIV, ulcers, and myocardial infarction (within the last year) do not have a unique important impact on PCS when using a 3-point threshold, but would have a unique impact if a 2-point threshold were used. The seemingly low impact of these conditions is probably due to the heterogeneity of disease in these self-identified groups and to the control for comorbidity. Finally, data from clinical trials of effective treatments also point to an MID of 2 *T*-score points. For example,

in six randomized controlled trials of patients with intermittent claudication due to peripheral arterial disease, cilostazol had a significant impact on treadmill walking distance, on Walking Impairment Questionnaire scores, and on SF-36 PCS scale scores (2-point mean PCS score change; Regensteiner et al., 2002). Thus, *an MID of 2 T-score points seems reasonable for PCS.*

Mental Component Summary (MCS)

As a summary measure of mental health, the pattern of associations is very different for MCS than for the PCS measure. As such, a 3-points lower MCS *T* score is associated with an OR of 1.13 for being unable to work and an OR of 1.16 for 1-year job loss. Risk of hospitalization is not noticeably increased with a 3-points lower score, but the probability of using mental health services is increased by approximately 30% (OR = 1.31). Among the Medicare population, a 3-points lower *T* score implies approximately 10% higher 1-year mortality risk (OR = 1.10–1.13 in the 25–50 *T*-score range). Depression and anxiety are associated with highly significant MCS decrements, with no other diseases having a unique burden exceeding 3 *T*-score points on the MCS scale. For example, chronic fatigue syndrome/fibromyalgia has a disease impact of 2.8 *T*-score points. When used as a predictor of clinically diagnosed depression, an MCS *T*-score difference of 3 points means an approximately 30% increased risk of depression (OR = 1.34). While there currently are data for fewer criteria/anchor variables of relevance for mental health than for physical health, *an MID of 3 T-score points seems reasonable for MCS.*

Physical Functioning (PF)

The patterns and strengths of associations for PF resemble those found for PCS. To wit, a 3-points lower PF *T* score is associated with an OR of 1.38 for being unable to work, an OR of 1.22 for job loss in the following year (for the employed population), and an OR of 1.13 for hospitalization in the subsequent year. Among the Medicare population, a 3-points lower *T* score implies a higher 1-year mortality risk, with the OR increasing particularly in the low scoring group (OR = 1.08–1.31 in the 25–50 *T*-score range, with higher OR for the low scoring groups). A 3-point threshold for importance would imply that the unique disease burdens (controlled for other diseases) of congestive heart failure, myocardial infarction within the last year, stroke, limited use of arms or legs, arthritis, COPD, and diabetes are important for PF, while the unique impact of conditions like anemia, asthma, and kidney disease would not be minimally important for PF. Further, conditions like liver disease, ulcers, sleep apnea, and clinical depression would not

have an important unique impact on PF using a 3-point threshold, but would have an important unique impact if a 2-point threshold were used. In conclusion, *an MID of 3 T-score points seems well chosen for PF.*

Role-Physical (RP)

Because the RP items were substantially revised during the development of the SF-36v2, data from the SF-36 cannot be used in the evaluation of MID for SF-36v2 data. That said, a 3-point threshold for importance would imply that the unique disease burdens (controlled for other diseases) of myocardial infarction within the last year, congestive heart failure, COPD, chronic fatigue syndrome/fibromyalgia, chronic back pain, clinical depression, arthritis, liver disease, stroke, and limited use of arms or legs are important for RP, while the unique impact of conditions like kidney disease, asthma, previous cancer, and migraine headaches would not be minimally important for RP. Conditions such as diabetes, sleep apnea, and ulcers would not have a unique important impact on RP using a 3-point threshold, but would have a unique impact if a 2-point threshold were used. Thus, *the data suggest an MID of 3 T-score points is reasonable for RP.*

Bodily Pain (BP)

As a measure of bodily pain, the BP scale has a fairly strong association with current inability to work. A 3-points lower BP *T* score is associated with an approximately 38% increased inability to work (OR = 1.38), an OR of 1.21 for 1-year job loss, an OR of 1.13 for 1-year hospitalization, and an OR of 1.08 to 1.29 for 1-year mortality (the higher OR comparing BP *T* scores of 25 and 22). Furthermore, a 3-point importance threshold would imply that the unique disease burdens (controlled for other diseases) of arthritis, back pain, chronic fatigue syndrome/fibromyalgia, diabetes, and limited use of arms or legs are important for BP, while the unique impact of conditions like sleep apnea, anxiety, congestive heart failure, and asthma would not be minimally important for BP. Conditions such as stroke, ulcers, and clinical depression would not have an important unique impact on BP using a 3-point threshold, but would have a unique impact if a 2-point threshold were used. The relatively low unique burden of migraine on bodily pain is surprising but likely due to lack of strict diagnostic criteria and to the control for comorbidities. In conclusion, *the data support an MID of 3 T-score points for BP.*

General Health (GH)

The pattern of associations for the GH scale demonstrates that a 3-points lower GH *T* score is associated

with approximately 46% increased inability to work (OR = 1.46), an OR of 1.25 for 1-year job loss, an OR of 1.15 for 1-year hospitalization, and an OR of 1.20 to 1.33 for 1-year mortality (the higher OR for a lower GH *T*-score range). Further, a 3-point importance threshold would imply that the unique disease burdens (controlled for other diseases) of congestive heart failure, COPD, diabetes, ulcers, clinical depression, chronic back problems, and limited use of arms or legs are important for GH, while the unique impact of conditions like asthma and migraine would not be minimally important for GH. Conditions such as stroke, myocardial infarction within the previous year, and rheumatoid arthritis would not have an important unique impact on GH using a 3-point threshold, but would have a unique impact if a 2-point threshold were used. In conclusion, *the data support an MID of 2 T-score points for GH.*

Vitality (VT)

While the VT items were revised during the development of the SF-36v2, the equivalence of scores (in *T* scores) from the two survey versions allows one to draw upon SF-36 data in evaluating SF-36v2 MIDs. A 3-points lower VT *T* score is associated with an approximately 38% increased inability to work (OR = 1.38), an OR of 1.21 for 1-year job loss, an OR of 1.12 for 1-year hospitalization, and an OR of 1.19 to 1.23 for 1-year mortality (the higher OR for a lower VT *T*-score range). Moreover, a 3-point importance threshold would imply that the unique disease burdens (controlled for other diseases) of COPD, clinical depression, anxiety, chronic fatigue syndrome/fibromyalgia, chronic back problems, and limitations in arms or legs are important for VT, while the unique impact of conditions like migraine headaches, irritable bowel syndrome, and ulcers would not be minimally important for VT. Conditions such as stroke, COPD, congestive heart failure, and diabetes would not have an important unique impact on VT using a 3-point threshold, but would have a unique impact if a 2-point threshold were used. Seeing that some of these diseases cause clinical fatigue, *these data suggest an MID of 2 T-score points for VT.*

Social Functioning (SF)

The pattern of associations for the SF scale demonstrates that a 3-points lower SF *T* score is associated with an approximately 30% increased inability to work (OR = 1.29), an OR of 1.16 for 1-year job loss, an OR of 1.10 for 1-year hospitalization, and an OR for 1-year mortality of 1.15 to 1.22 (the higher OR for a lower SF *T*-score score range). Further, a 3-point importance threshold would imply that the unique disease burdens

(controlled for other diseases) of clinical depression, anxiety, chronic fatigue syndrome/fibromyalgia, myocardial infarction within the last year, congestive heart failure, chronic back problems, and limited use of arms or legs are important for SF, while the unique impact of conditions like diabetes, migraine, asthma, and liver disease would not be minimally important for SF. Conditions such as rheumatoid arthritis, stroke, ulcers, and COPD would not have an important unique impact on SF using a 3-point threshold, but would have a unique impact if a 2-point threshold were used. With data on predictive validity suggesting a slightly higher MID and data on disease burden suggesting a slightly lower MID, the evidence is less clear than for most other scales but *an MID of 3 T-score points seems reasonable for SF.*

Role-Emotional (RE)

Because the RE items were substantially revised during the development of the SF-36v2, data from the SF-36 cannot be used in the evaluation of MID for the SF-36v2 data. That said, a 3-point threshold for importance would imply that the unique disease burdens (controlled for other diseases) of clinical depression, anxiety, rheumatoid arthritis, myocardial infarction within the last year, chronic fatigue syndrome/fibromyalgia, stroke, and limited use of arms or legs are important for RE, while the unique impact of conditions like diabetes, COPD, and sleep apnea would not be minimally important for RE. Conditions such as ulcers, congestive heart failure, and chronic back problems would not have a unique important impact on RE *T* scores using a 3-point threshold, but would have a unique impact if a 2-point threshold were used. Meanwhile, using a 4-point threshold would imply that depression, anxiety, and stroke have important unique impacts on RE. While there currently are fewer anchors of relevance for determining MID for the RE scale, *an MID of 4 T-score points seems reasonable for RE.*

Mental Health (MH)

While the MH items were revised during the development of the SF-36v2, the equivalence of scores (in *T* scores) from the two survey versions allows one to draw upon SF-36 data in evaluating SF-36v2 MIDs. The pattern of associations for MH resembles that found for MCS, with a 3-points lower MH *T* score being associated with an approximately 15% increased inability to work (OR = 1.15), an OR of 1.08 for 1-year job loss, and an OR of 1.10 to 1.13 for 1-year mortality. Furthermore, risk of hospitalization is not noticeably increased with a 3-points lower score, but the probability of using mental health services is increased by approximately 30% (OR

= 1.30). A 3-point importance threshold would imply that the unique disease burdens (controlled for other diseases) of clinical depression and anxiety are important for MH, with no other conditions having an important unique impact on MH *T* scores. A 14-point MH decrement was found in the MOS, using depression defined by clinical criteria (see Wells et al., 1989). When used as a predictor of clinically diagnosed depression, an MH *T*-score difference of 3 points means an approximately 35% increased risk of depression (OR = 1.36). While there currently are data for fewer criteria/anchor variables of relevance for mental health than for physical health, *an MID of 3 T-score points seems reasonable for MH.*

Responder Definition: Criteria for Minimally Important Difference in Individual Respondent Scores

In addition to the previously mentioned criteria for defining group-level MIDs (i.e., association with clinical criteria or self-evaluation of health and forecasting of life events), criteria for individual scores have to take measurement precision and reliability into account. Briefly, if a scale has low reliability, a larger score difference is required to signify that a change has occurred. Larger score differences are also necessary if a scale is crude, in the sense that it has a limited number of possible scores and has a large *step size* (i.e., the points difference for the smallest possible scale change, equivalent to a change from one item category to the next in a single item; see Table 10.2).

A difference of 1 standard error of measurement (*SEM*) has been proposed as the criteria for significant intraindividual change and found to agree well with patients' self-evaluation of significant change (Wyrwich, Nienaber, Tierney, & Wolinsky, 1999; Wyrwich, Tierney, & Wolinsky, 1999). From a statistical perspective, ± 1 *SEM* is equivalent to a 68% confidence interval around a single score (see Tables 7.2 and 10.2). However, the error around a change score is larger than the error for a single measurement point. Jacobson and Truax (1991) proposed a reliable change index (RCI) based on the change score error and using a 95% confidence interval. Thresholds based on this RCI are presented in Table 10.2. However, the RCI approach appears overly conservative because it assumes that the baseline and follow-up scores have uncorrelated error terms. Further, while a 95% confidence interval (equivalent to a 5% significance level) is the standard used in group-level analyses, this criterion seems overly narrow for analyses of individual

Table 10.2*Criteria for Significant Change Scores According to Different Proposals*

SF-36v2 Measures/Scales	Minimum Step Size	SEM ^a	RCI ^b	Proposed Values for Responder Definition		
				SC95 ^c	SC90 ^d	SC80 ^e
PCS	-	2.0	5.5	5.3	4.4	3.4
MCS	-	2.7	7.5	7.1	6.0	4.6
Physical Functioning	2.1	2.5	6.9	6.6	5.5	4.3
Role-Physical	2.5	2.0	5.5	5.3	4.4	3.4
Bodily Pain	4.7	3.6	10.0	9.5	7.9	6.2
General Health	1.4	4.2	11.6	11.0	9.3	7.2
Vitality	3.1	3.6	10.0	9.5	7.9	6.2
Social Functioning	5.5	4.0	11.1	10.5	8.8	6.9
Role-Emotional	3.9	2.6	7.2	6.8	5.7	4.5
Mental Health	2.8	3.6	10.0	9.5	7.9	6.2

Note. Values are based on reliability estimates and standard deviations for the eight health domain scales and the PCS and MCS measures in the 2009 U.S. general population sample.

^a1 SEM = 68% confidence interval around a single measurement point.

^bThreshold based on the reliable change index (RCI; Jacobson & Truax, 1991) assuming uncorrelated measures and using a 95% significance level.

^cSignificant change assuming baseline–follow-up correlation of .10 and using a 95% confidence interval.

^dSignificant change assuming baseline–follow-up correlation of .10 and using a 90% confidence interval.

^eSignificant change assuming baseline–follow-up correlation of .10 and using a 80% confidence interval.

respondents, where the risk of falsely identifying change must be balanced against the risk of overlooking true change. Therefore, assuming a baseline–follow-up error correlation of .10 and using an 80% confidence interval seems reasonable.

Responder definitions based on these assumptions are also displayed in Table 10.2. For example, using this principle, the responder definition for the PF scale would be 4.3 points. For a respondent whose true state is unchanged, this cut-point results in a 10% risk of incorrectly classifying the individual as a responder. However, as previously mentioned, this risk must be weighted against the risk of overlooking real change. Using the same cut-point, a respondent with a true score change of 8.6 points has a 10% chance of being classified as stable. Thus, the 80% confidence interval seems to represent a reasonable solution. For those wanting to take a more conservative approach to the evaluation of change in individual respondent scores, values for determining the 90% and 95% confidence intervals are also provided in Table 10.2. Note that all of these criteria are larger than the MIDs for group comparisons, for the reasons previously discussed in this chapter.

The responder definitions discussed here assume that the measurement error around a single score is constant throughout the measurement range. However, for the summative scoring approach used by the SF-36v2, this assumption is not realistic for scores close to or at the floor or ceiling (see Spratt, 2009). More realistic models assume smaller measurement error close to the floor or ceiling, suggesting that a smaller change may be

clinically significant in these score ranges (Spratt, 2009). Nevertheless, when evaluated empirically, responder definitions based on score-specific measurement error did not change study conclusions (Spratt, 2009), so the SF-36v2's developers recommend using a simple approach assuming constant measurement error.

Another approach to establishing thresholds for important individual change relies on expert consensus panels (Wyrwich, Fihn, et al., 2003; Wyrwich, Nelson, et al., 2003; Wyrwich et al., 2004). Such panels aim to reach consensus recommendations regarding MIDs for various health outcome scales within a particular disease area, based on descriptions of the measurement properties of each scale and the anchor information linked to individual respondent case histories. While the details of the information presented to such expert panels are not available, it seems to be of the same type as discussed in this chapter. However, the approach offered in this chapter differs in that a conceptual distinction between MID and a responder definition is made, separate MIDs by disease group are not generally believed to be necessary, and group-level anchors are considered more reliable than individual respondent case histories.

In addition to the anchors and statistical criteria considered so far, a third criterion for significant individual respondent change has been suggested (Jacobson & Truax, 1991). In relation to well-established norms, differences in scores that change from “normal” to “abnormal” should be considered important. Users may approach this evaluation of change using the SF-36v2 2009 U.S. general population normative data and any

relevant disease-specific norms (available from QualityMetric and its authorized resellers or from published studies). If the general population and the disease-specific population have overlapping score distributions (which is generally the case) and the *SDs* are roughly equal, then the threshold for moving out of the disease distribution can be calculated as the midpoint between the two population means (Jacobson & Truax, 1991). A respondent can be said to have experienced an important improvement if he or she had a baseline score below the threshold and a follow-up score above the threshold. Furthermore, to take measurement error into account, a dual criterion can be applied using the significant change criteria shown in Table 10.2. Note that the calculation of thresholds can be adjusted if the two populations have very different *SDs* (Jacobson & Truax, 1991). This approach differs from those previously discussed here, in that a respondent with a very low score would require a larger change to be considered a responder. Further, thresholds for significance will vary between disease groups. This approach is somewhat at odds with the predictive validity results showing that a certain score improvement generally has greater implications for low-scoring respondents. For this reason, establishing MID based on respondents moving in or out of the normal range is not recommended.

Summary

Minimally important difference is a research area with ongoing conceptual and methodological development. This chapter discussed an approach that builds on a theoretical distinction between MID as relating to mean group differences and to a responder definition that concerns change for individual respondents. While this distinction appears methodologically sound, the responder definition may be helpful in reporting results from clinical trials as well; for example, by comparing the proportion of individuals responding to treatment and

to placebo (i.e., improving by more than the responder definition cutoff; see Strand, 2005, for an example of this approach). If change scores have an approximately normal distribution, the odds ratio for responding will be fairly robust to changes in the threshold for response. However, since dichotomization into *responder* and *nonresponder* groups results in a loss of information and statistical power, such analyses are best used as auxiliary and not as the primary approach to hypothesis testing. Thus, the concept of MID is still core to the planning of clinical trials and to power analyses.

In short, based on available anchor data, the following MID values, in terms of *T*-score points, are proposed for SF-36v2 component and scale group mean scores: PCS, 2; MCS, 3; PF, 3; RP, 3; BP, 3; GH, 2; VT, 2; SF, 3; RE, 4; and MH, 3. These MID values are appropriate for groups with mean *T*-scores of 30 to 40. For higher *T*-score ranges, MID values tend to be higher.

When evaluating the treatment response of individual respondents, test precision must be considered in addition to the anchor-based approach used to establish MID. To this end, the SF-36v2's developers propose a *significant change* criteria that bears some similarities to the reliable change index (RCI) suggested by Jacobson and Truax (1991) but is more realistic and practical in the sense that it assumes a small correlation between baseline and follow-up assessments and weighs the risk of falsely classifying a person as a responder against the risk of overlooking a true treatment response. Based on this approach, the following responder definition values, in terms of *T*-score points, are proposed for SF-36v2 component and scale individual respondent scores: PCS, 3.8; MCS, 4.6; PF, 4.3; RP, 4.0; BP, 5.5; GH, 7.0; VT, 6.7; SF, 6.2; RE, 4.6; and MH, 6.7. While these suggestions represent the best estimates based on currently available evidence, note with caution that MID is the focus of many current research projects, the data from which may necessitate the modification of the MID and responder definition guidelines suggested in this manual.



11

Interpretation of Group Data

This chapter summarizes results from the reanalyses of three published studies that employed the SF-36v2. The purpose of these reanalyses is to illustrate the use of the interpretation guidelines, previously presented in this manual, by applying these guidelines to and interpreting the findings of independent researchers. Each of the three studies presented here evaluated the outcomes of a treatment or other intervention. For the purpose of this chapter, assume that all data quality indicators were within acceptable ranges.

Note that the publication of the SF-36v2's 2009 norms coincided with the publication of this third edition of the *User's Manual*; as a result, published studies employing the 2009 norms were not available for inclusion in this manual. Instead, the three case studies that appeared in the second edition of this manual (Ware et al., 2007) are presented here. Although the results of these case studies are based on the 1998 norms, the same approach would be taken when using the 2009 norms to analyze SF-36v2 group-level results.

Case 1

This randomized, controlled study reported mean changes in SF-36v2 scale scores for chronic hepatitis C (HCV) patients on combination therapy (interferon α and ribavirin or pegylated interferon α and ribavirin) who developed anemia (Afdhal et al., 2004). Patients were randomized to receive either epoetin alfa or placebo over an 8-week, double-blinded phase. Among the study's reported conclusions was that epoetin alfa improves quality of life in anemic HCV-infected patients receiving combination therapy, as evidenced by the significant differences in score changes between treated and placebo groups on seven of the eight health domain scales. Each scale's score changes were reported using the 0–100 metric, and no results were reported for the

PCS or MCS measures. In the reanalysis of this study, each scale's score changes were linearly transformed into T scores. Also, changes in T scores were estimated for the two component summary measures.

Figures 11.1 and 11.2 present the unstandardized (0–100 scores) and standardized (T -score) estimates, respectively, of average score changes in the eight health domain scales and the two summary measures for anemic HCV-infected patients randomized to epoetin alfa and placebo groups. Note that the PCS and MCS scores maintain the same relationship across the two figures because both are scored using T -score units. Because the standardization of change scores following treatment is a linear transformation, this standardization does not alter conclusions made from statistical tests of significance. However, standardization does change some of the conclusions about which health domain scale scores changed the most amongst patients treated with epoetin alfa. Figure 11.1 (unstandardized scores) suggests that the RP scale improved slightly more than the PF scale; however, as shown in Figure 11.2 (standardized scores), the PF scale improved slightly more than the RP scale. Similarly, Figure 11.1 suggests that the RE scale improved more than the MH scale, whereas Figure 11.2 reveal the reverse to be true when change scores are expressed in standardized T -score units.

Another important observation when comparing results between Figures 11.1 and 11.2 concerns the comparison of the eight health domain scales with the two component summary measures. Looking at Figure 11.1, one could erroneously conclude that the summary measures were less responsive to changes from treatment with epoetin alfa than were seven of the eight health domain scales. However, given the different scoring units (0–100 versus T scores) between the health domain scales and component summary measures represented in Figure 11.1, users cannot simply compare the results obtained from the scales against those from the summary

Figure 11.1 Unstandardized Changes in SF-36v2 Health Domain Scale and Component Summary Measure Scores Between Anemic HCV Patients Randomized to Epoetin Alfa and Placebo

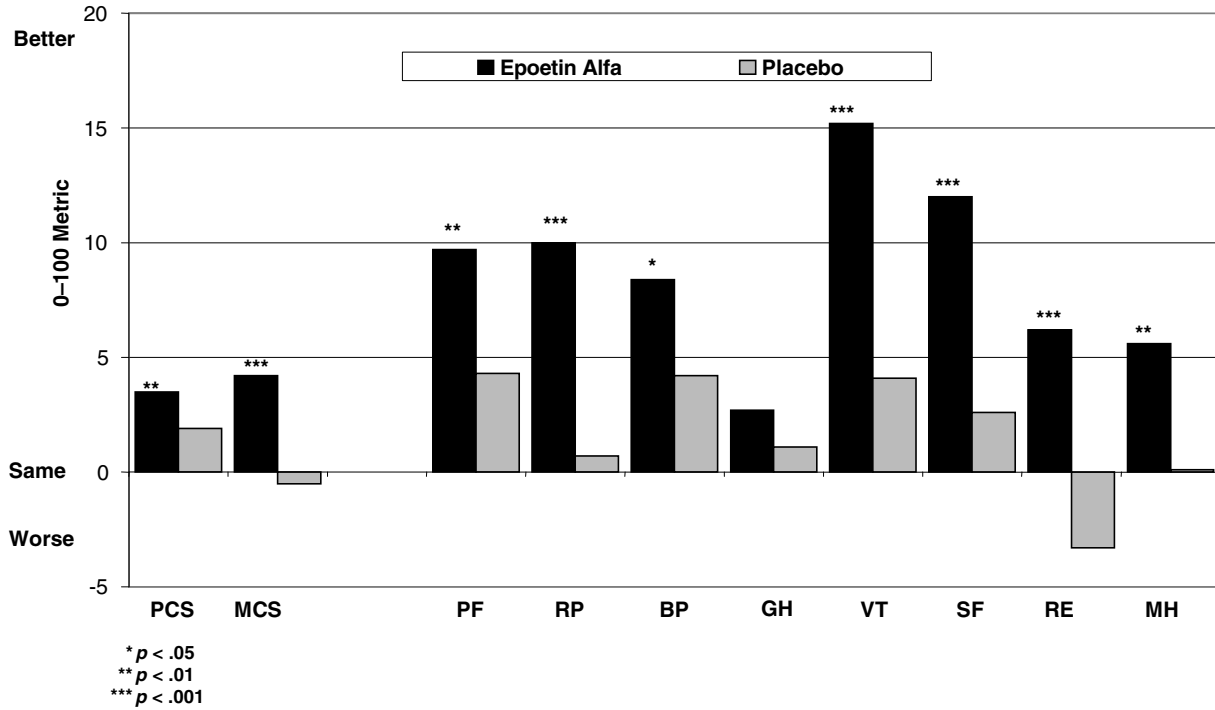
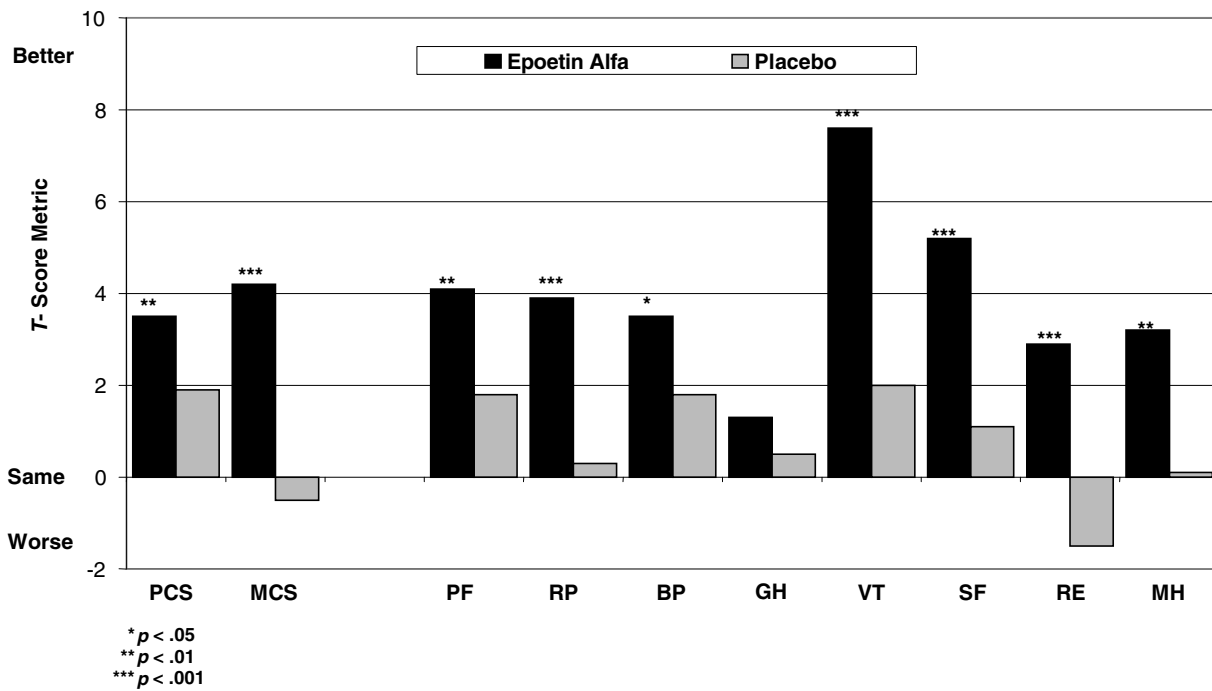


Figure 11.2 Standardized Changes in SF-36v2 Health Domain Scale and Component Summary Measure Scores Between Anemic HCV Patients Randomized to Epoetin Alfa and Placebo



measures. Furthermore, although the outcome differences between the epoetin alfa and placebo groups were statistically different in seven of the eight scales, comparison of change estimates in PCS and MCS T scores indicates that improvement was slightly better in the mental health measure, which is reflected in the profile by the large improvements in VT and SF scale scores.

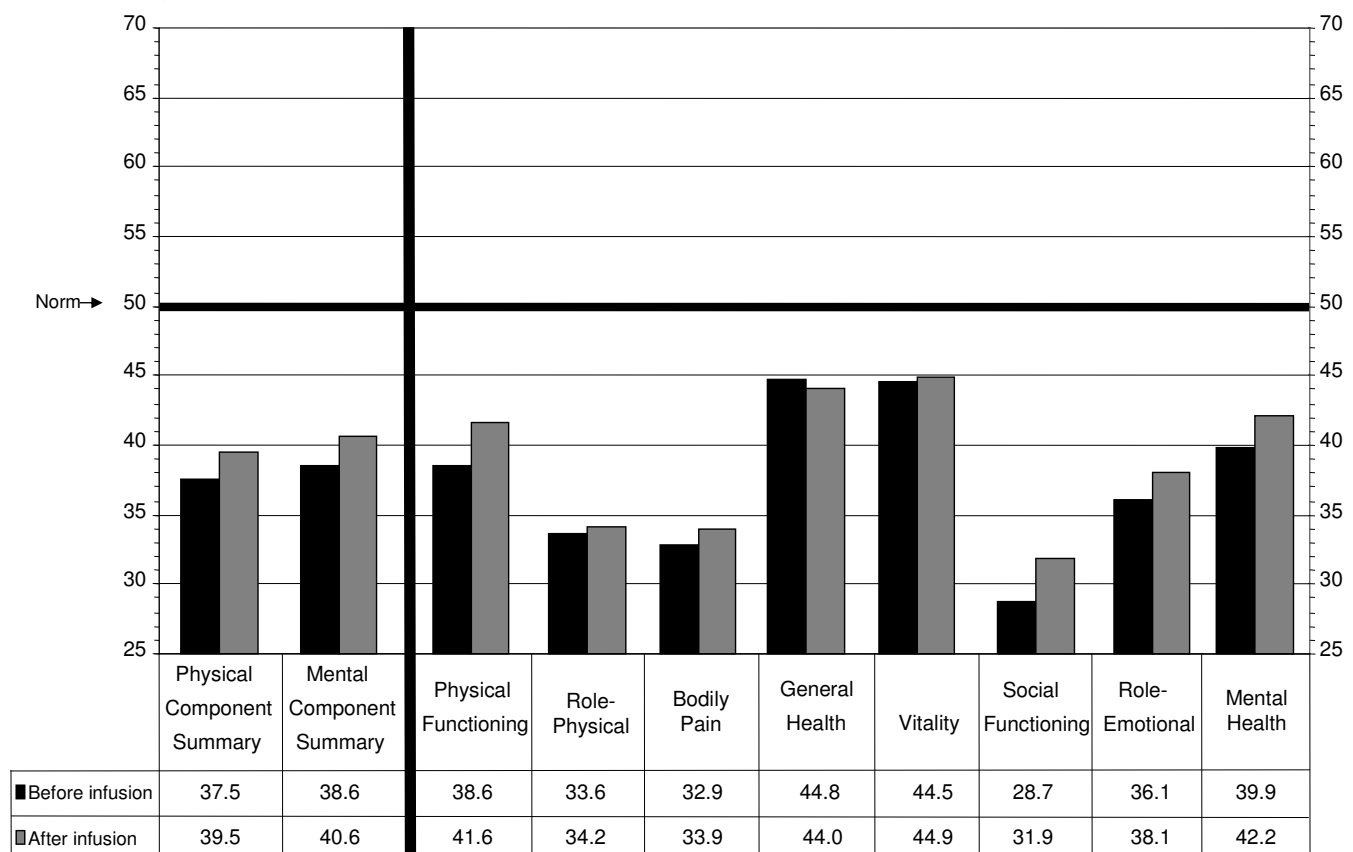
Case 2

This study reported SF-36v2 profiles for 20 patients with degenerative cervical disc disease, both before and after anterior cervical fusion (Lanman & Hopkins, 2004). At 3-months post-infusion, patients showed score improvement on all health domain scales and component summary measures except the GH scale. As in the Afdhal et al. (2004) study, results for the health domain scales were reported using the 0–100 metric, and no data were presented for the PCS and MCS measures. In this reanalysis, health domain scale scores at baseline and 3-months post-infusion were linearly transformed to T scores, and PCS and MCS scores were computed.

Figure 11.3 compares SF-36v2 health domain scale and component summary measure T scores before anterior cervical fusion with scores at 3-months post-infusion. Given that each scale and summary measure has a mean T score of 50 in the U.S. general population, it is clear that these patients were functioning well below the average range prior to surgery, scoring on average at least 1 SD (10 T -score points) below the U.S. general population norm on all health domain scales except GH and VT. The burden of this condition seems as much mental as it is physical, as demonstrated by the respondents' before-treatment mean scores on both the health domain scales and component summary measures. Three months after anterior cervical fusion, patient scores noticeably improved on the PF (3.0 points), SF (3.3 points), RE (2.0 points), and MH (2.3 points) scales, as well as on the PCS (2.0 points) and MCS (2.0 points) measures. Using the 10-point SD observed in the U.S. general population, the changes in scores are considered small effect size changes.

Meaningful interpretations of the score changes observed in this study go beyond effect size. Thus, the

Figure 11.3 Mean SF-36v2 Health Domain Scale and Component Summary Measure Scores Before and After Anterior Cervical Fusion ($N = 20$)



following paragraphs discuss the scales showing the largest changes from pre- to post-intervention to illustrate how to apply the interpretation guidelines to the outcomes of this study.

First, a norm-based (*T*-score) approach to interpretation reveals that average patient scores prior to the intervention were more than 1 *SD* below the mean scores of the U.S. general population on all but the GH and VT scales. Differences of more than 0.8 *SD* units are considered large effect-size differences (Cohen 1988). The change in the PF *T* score from 38.6 to 41.6 represents an improvement from the 18th to the 21st percentile score in the U.S. general population. The change in the SF scale *T* score from 28.7 to 31.9 represents an improvement from the 9th to the 12th percentile score in the U.S. general population.

Second, content-based interpretation reveals that the change in the PF score from 38.6 to 41.6 represents a reduction in the percentage of respondents reporting limitations in climbing one flight of stairs, from 65% to 43% (see Ware et al., 2007, Chapter 8). The change in the SF scale score from 28.7 to 31.9 represents a reduction in the percentage of respondents reporting that their physical or emotional health interferes with social activities *all or most of the time*, from 37% to 25% (see Ware et al., 2007, Chapter 8).

Third, the criterion-based approach to interpretation shows that the change in the PF *T* score from 38.6 to

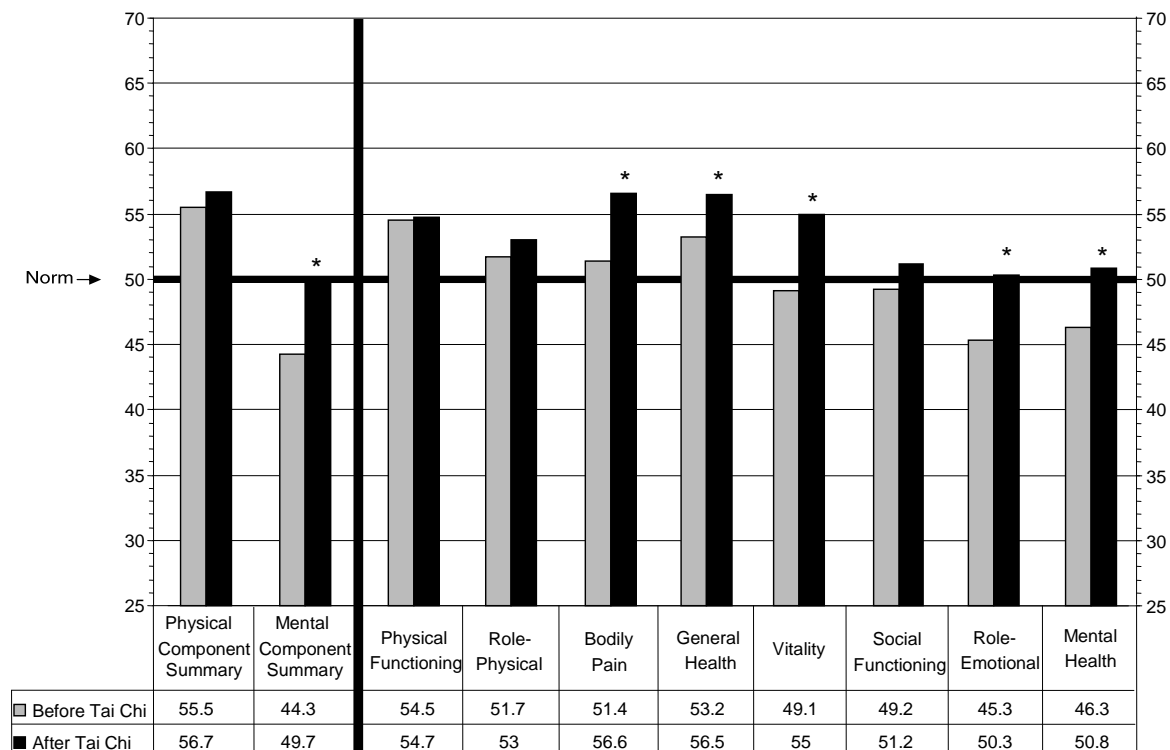
41.6 represents a reduction in the percentage of respondents reporting that they could not work a paying job because of their health, from 36% to 29% (see Ware et al., 2007, Chapter 9). The change in the SF *T* score from 28.7 to 31.9 represents a reduction in the percentage of respondents reporting participating in fewer social activities with groups of people because of their health, from 88% to 77% (see Ware et al., 2007, Chapter 9).

Case 3

In a nonclinical study, Wang Taylor, Pearl, and Chang (2004) reported SF-36v2 profiles for 31 college students who participated in tai chi exercise classes. The research design was a one-group, pretest, posttest design and the intervention consisted of a 3-month series of 1-hour tai chi exercise sessions performed twice weekly. Survey administration occurred before and after the intervention to determine the effects of tai chi exercise on the physical and mental health of college students.

Figure 11.4 compares SF-36v2 health domain scale and component summary measure *T* scores obtained before students began the tai chi exercise intervention with those scores obtained after 3 months of the intervention. It is clear from Figure 11.4 that prior to the start of the intervention the students, on average, were functioning

Figure 11.4 Effects of Tai Chi Exercise on the Physical and Mental Health of College Students ($N = 31$)



* $p < .05$

at or above the average range (T scores = 47–53) on all scales and summary measures except MCS, RE, and MH. Following the 3-month intervention, significant improvements were observed on the BP (5.2 points), GH (3.2 points), VT (5.9 points), RE (5 points), and MH (4.5 points) scales, as well as on the MCS measure (5.4 points). Using the SD (10 T -score points) observed in the U.S. general population, these changes in scores are considered small to moderate effect size changes and surpass minimally important difference (MID; see Chapter 10) thresholds established for group-level comparisons.

Norm-based interpretation of these results reveals that prior to the intervention, the average T scores for the MH and RE scales and the MCS measure were roughly 0.5 SD units below the general population norm. Differences in scores of 0.5 SD units are considered moderate effect-size differences (Cohen 1988). Also, according to the general interpretation guidelines presented in Chapter 7, the obtained RE and MH scores were below the average range for group-level data. Furthermore, average T scores for the BP, GH, and VT scales were at or above general population norm. To illustrate, the change in the MH scale score from 46.3 to 50.8 represents an improvement from the 28th to the 46th percentile score in the U.S. general population. The change in the RE scale score from 45.3 to 50.3 represents an improvement from the 26th to the 33rd percentile score, whereas the change in the MCS score from 44.3 to 49.7 represents an improvement from the 23rd to the 37th percentile score. The change in the BP score from 51.4 to 56.6 represents an improvement from the 57th to the 78th percentile score, and the change in the VT score from 49.1 to 55.0 represents an improvement from the 47th to the 70th percentile score.

A content-based interpretation illustrates that the change in the MH T score from 46.3 to 50.8 represents an increase in the percentage of respondents in the 1998 U.S. general population sample reporting being happy *all or most of the time*, from 60% to 79% (see Ware et al., 2007, Chapter 8). The change in the RE T score from 45.3 to 50.3 represents a decrease in the percentage of respondents reporting the need to cut down time spent at work *any of the time* during the past 4 weeks,

from 43% to 8%. The change in the MCS T score from 44.3 to 49.7 represents a decrease in the percentage of respondents reporting they accomplished less at work *all or most of the time*, from 7.9% to 1.9%; a decrease in the percentage of respondents reporting limitations in social activities *all or most of the time*, from 6.7% to 2.1%; and a decrease in the percentage of respondents reporting feeling tired *all or most of the time*, from 31% to 18%. The change in the BP T score from 51.4 to 56.6 represents a decrease in the percentage of respondents reporting *any* interference with normal work due to pain, from 37% to 0%, whereas the change in the GH T score from 53.2 to 56.5 represents an increase in the percentage of respondents reporting *excellent* health, from 9.4% to 14.8%. The change in the VT T score from 49.1 to 55.0 represents an increase in the percentage of respondents reporting feeling full of energy *all or most of the time*, from 21.2% to 75.2% (see Ware et al., 2007, Chapter 8, regarding results reported in this paragraph).

Use of criterion-based interpretation shows that the change in the MH T score from 46.3 to 50.8 represents a 35% reduction (from 26% to 17%) in the likelihood of receiving mental health specialty care in the next 6 months. The change in the RE T score from 45.3 to 50.3 represents a 32% reduction (from 13.4% to 9.1%) in the percentage of respondents reporting that their productivity at work, home, or school was reduced by one-half or more because of emotional problems (see Ware et al., 2007, Chapter 9). The change in the MCS T score from 44.3 to 49.7 represents a 36% reduction (from 12.8% to 8.2%) in the likelihood of having a diagnosis of depression (see Ware et al., 2007, Chapter 9). The change in the BP T score from 51.4 to 56.6 represents a 39% reduction in the percentage of respondents reporting one or more missed days of work due to pain, from 23.5% to 14.4%. The change in the GH T score from 53.2 to 56.5 represents a 12% reduction in the likelihood of being hospitalized overnight within 6 months, from 7.5% to 6.6%. Lastly, the change in the VT T score from 49.1 to 55.0 represents a 45% decrease in the likelihood of job loss due to health within 6 months, from 14.5% to 7.9% (see Ware et al., 2007, Chapter 9, regarding results reported in this paragraph).



12

Interpretation of Individual Respondent Data

Although the SF-36v2 was originally developed for administration to large population samples, it has increasingly been used to assess and monitor individual respondents being treated for a wide range of disorders in a variety of treatment settings. Whereas Chapter 11 specifically addresses considerations for the interpretation of results obtained from large group or population studies, the focus of this chapter is on the interpretation of SF-36v2 results obtained from individual respondents being evaluated and/or treated in clinical settings.

The usefulness of the SF-36 and SF-36v2 for monitoring treatment and improving health outcomes has previously been demonstrated. For example, one investigation included quarterly administrations of the SF-36 to expand the definition of the *adequacy*, or *quality*, of dialysis beyond traditional laboratory test values among respondents with end-stage renal disease (Kurtin, Davies, Meyer, DeGiacomo, & Kantz, 1992; Meyer et al., 1994). The reported results for these individual dialysis respondents illustrated both the feasibility and usefulness of periodic health assessments in managing a patient's progression from advanced renal failure to end-stage renal disease (Meyer et al., 1994). Ware and Kosinski (2001b) offer examples of the use of the SF-36 in monitoring a respondent with clinical depression and a respondent with congestive heart failure (CHF). Other examples are provided by Davies (2000) and Davies and Kram (2002).

Just as for aggregated group data, the general guidelines for interpreting SF-36v2 *T*-score results presented in Chapters 6 through 10 are also applicable to individual respondent data. However, there are additional considerations for the interpretation of such respondent data that the survey user should take into account: response consistency, item analysis, and situational considerations. This chapter briefly addresses these considerations and presents two case studies that illustrate the application of the general and additional interpretive considerations to SF-36v2 individual respondent data.

Considerations for Interpreting Individual Respondent Data

Normative data based on the findings from large samples provide the foundation for interpreting group or individual respondent SF-36v2 results. The observed *T* score on a given health domain scale provides an indication of a respondent's general or overall functioning in that domain. The ability to easily analyze item-level results provides a clinician or researcher with the opportunity to more fully use and integrate individual item data than is usually the case when interpreting aggregated group data.

Response Consistency

One method of evaluating the quality of the SF-36v2 data is by analyzing individual responses. The Response Consistency Index (RCI) can provide a means of evaluating the consistency of responses to 15 pairs of survey items (see Chapter 5). Recall that scoring the RCI consists of assigning a value of 1 to item pair responses that are inconsistent and a value of 0 to each item pair that has consistent responses. The final RCI score for an individual respondent is the sum of the 15 item-pair scores. Thus, the best RCI score is 0 and the worst score is 15.

Tables 6.2 and 6.3 present the frequency distributions of SF-36v2 RCI scores for the standard (4-week) and acute (1-week) formats, respectively, using 2009 U.S. general population norms. Approximately 97% of the respondents in each form's sample displayed inconsistent responses to no more than one of the item pairs. Based on these data, it is recommended that an RCI score of 2 or greater for an individual respondent be considered indicative of potential problems in understanding the survey items, understanding how to complete the survey, or having the motivation to respond honestly or carefully. For this reason, it is useful to follow-up

with any respondents who obtains RCI scores of 2 or greater. Giving respondents the opportunity to explain inconsistencies leads to greater understanding of the nature of the contradictory responses. In some cases, the responses may not be contradictory; rather, they may reflect unusual but real respondent circumstances. In most cases, however, the responses represent a lack of understanding regarding an item, carelessness in indicating a response, or some other factor that is unrelated to the respondent's perceived health status. Being faced with the apparent contradictions may lead respondents to change one of the responses in the problematic item pair, thus making it more indicative of their perceived health status.

It is not necessary for a respondent to complete all 15 item-pairs to compute the RCI. Pairs with missing or out-of-range data are not used in the calculation. However, if a response is missing to one or both items for *all* 15 item-pairs, then an RCI score cannot be calculated for that respondent. Note that cases in which several item responses are missing suggest the presence of another type of data quality problem (e.g., poor reading skills, poor understanding of items) and should be investigated. To request more information about scoring the RCI, please visit <http://www.qualitymetric.com>.

Item Analysis

Item-related information can be gleaned from SF-36v2 data using the content-based interpretation guidelines presented in Chapter 8 of this manual. These guidelines, developed primarily for use with group data, are also useful for interpreting data at the individual respondent level because they can provide greater understanding of the meaning of differences in health domain scale and component summary measure scores at various *T*-score levels. Unfortunately, this approach provides little specific information about variations in functioning within each domain.

More specific implications of health domain scale scores can be discerned through the examination of each individual item from each scale (see Table 2.1 in Chapter 2). Knowing a respondent's specific response to each health domain item provides the examiner with the opportunity to understand which areas of functioning are contributing to or account for the observed scale score. This step is particularly useful in determining the types of functional limitations that are present in cases where a scale score falls within the mid-range of potential health domain scale scores (e.g., $T = 30\text{--}45$). For example, it may be important to know that a respondent with a *T* score of 38 on the PF scale may not be able to walk more than a mile or engage in vigorous activities but

can carry groceries and climb several flights of stairs. Analysis of item responses is particularly important when a respondent's score falls into the "gray" area between impaired and unimpaired functioning (i.e., T score = 40–44). When this is the case, examination of responses to individual items in the health domain scale in question may enable the administrator to determine whether the score is more indicative of impaired or unimpaired functioning.

Situational Considerations

Interpretation of scale scores can be enhanced by considering additional information about the nature of a given problem. For example, examination of a respondent's item responses reveals that a *T* score of 42 on the RP scale can be attributed to difficulty and limitations in performing activities. Upon further investigation, the administrator learns that the respondent recently broke an arm, which is now in a cast. Thus, this score represents a temporary impairment that is the result of an injury, which an otherwise healthy respondent should quickly recover from, rather than an impairment stemming from a chronic disease or disorder. As a result, a relatively quick return to at least an average level of functioning can be expected.

Overall, consideration of individual item responses allows for a better interpretation of SF-36v2 findings and a better understanding of the respondent's functional strengths and weaknesses within a specific health domain. Such knowledge may have implications for respondent treatment.

Case Studies

The following sections present two case studies that demonstrate the use of the SF-36v2 in assessing, planning, and monitoring treatment over varying periods of time. Because publication of the SF-36v2's 2009 norms coincided with the publication of this third edition of the *User's Manual*, case studies employing the 2009 norms could not be developed in time for inclusion in this manual. Instead, two individual respondent case studies that appeared in the second edition of the SF-36v2 *User's Manual* (Ware et al., 2007) are presented here. Although the results of both these case studies are based on the 1998 norms, including their confidence intervals (CIs) and responder definitions, the same approach would be taken when using 2009 norms to assess individual respondent SF-36v2 results.

Note that the cases presented here are based on actual results obtained in a primary care setting. Aspects

of each case were changed to disguise the identity of the respondents. Because the tracking of these respondents began prior to the publication of the SF-36v2, some or all of each case's original data were obtained using the SF-36. In these instances, SF-36 data were converted to SF-36v2 data (based on 1998 norms) using the Quality-Metric Health Outcomes Scoring Software 2.0 (Saris-Baglana et al., 2007).

The approach to assessment taken here is different from the approach used with the group-data case studies presented in Chapter 11. The basis for this approach is the norm-based interpretation of SF-36v2 summary measure and domain scale *T* scores with consideration of individual item responses, data from other tests, and historic and situational data—an approach that is much more practical for analyzing individual respondent data than group data. Each point-in-time score is analyzed using an 80% CI. Furthermore, *T*-score responder definitions, which are based on an assumed baseline-to-follow-up correlation of .40 and an 80% confidence interval, are used to determine if a score change found in the trended results over time represents an important difference. The 1998 norm-based CIs and responder definitions used for both case studies are those from the original publication of these case studies (Ware et al., 2007) and are reproduced here in Table 12.1.

Case 1

Katherine D. is a 52-year-old divorced mother of two children who also cares for her father, who suffers

Table 12.1

Values for Determining Health Domain Scale and Component Summary Measure Confidence Intervals and Minimally Important Differences for Case 1 and Case 2

Measure/Scale	80% CI ^a	Responder Definition ^b
PCS	±2.8	3.1
MCS	±3.5	3.8
Physical Functioning	±3.2	3.5
Role-Physical	±2.9	3.2
Bodily Pain	±4.1	4.5
General Health	±5.3	5.7
Vitality	±5.0	5.5
Social Functioning	±4.6	5.0
Role-Emotional	±3.5	3.8
Mental Health	±5.0	5.5

Note. Estimates are based on reliability estimates and standard deviations for the eight health domain scales and the PCS and MCS measures in the 1998 U.S. general population.

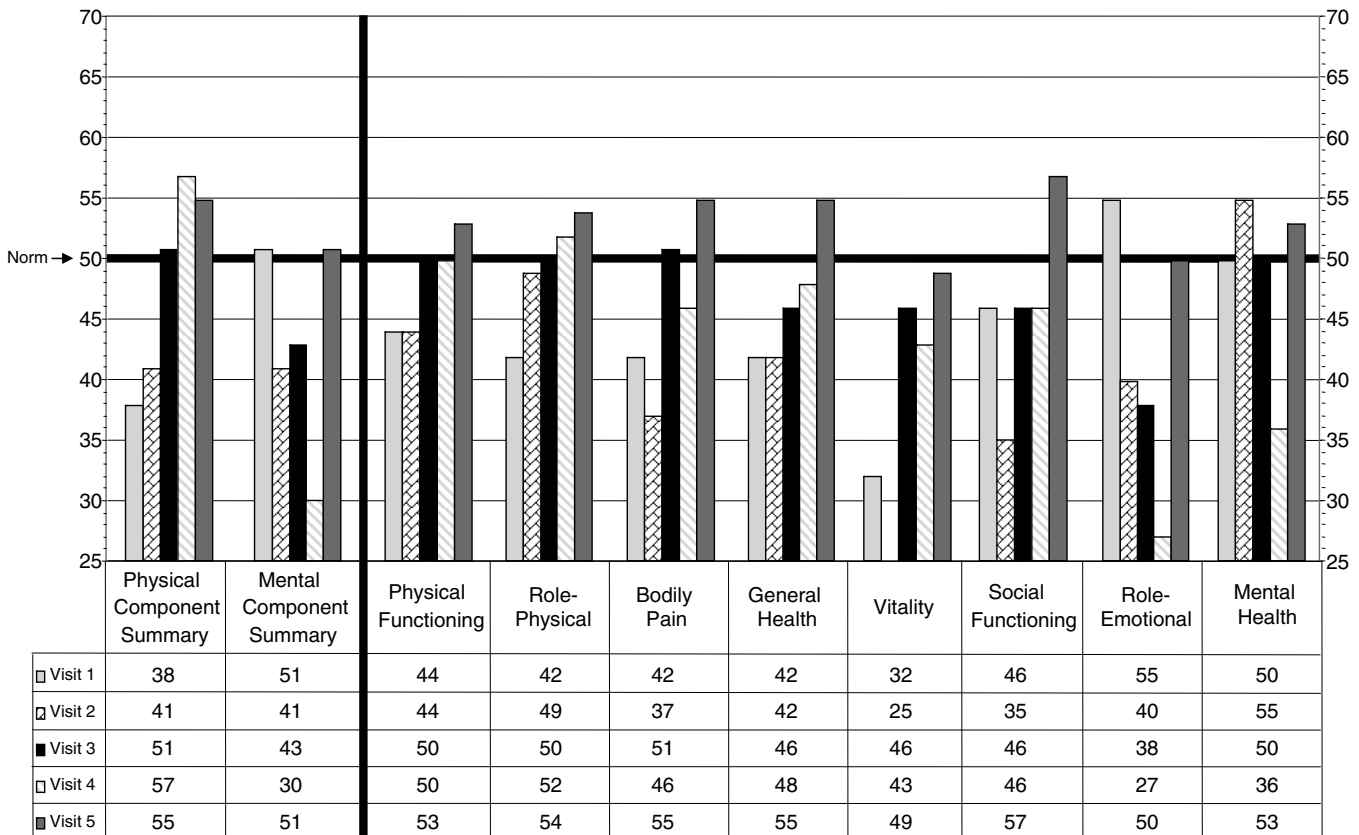
^a80% CI equals ±1.28 *SEMs*.

^bSignificant change assuming baseline–follow-up correlation of .40 and using an 80% confidence interval.

from diabetes. She was first seen in July 2001 (Visit 1) for a routine physical examination. At that time, she reported left knee pain, for which she was already taking medication, as well as fatigue, which she had been experiencing for about 1 year. With the exception of obesity, the results of the physical examination were normal. She was administered the SF-36 as part of the examination, with the results later being converted to SF-36v2 *T* scores. The Zung Self-Rating Depression Scale (SDS) was also administered during this visit. The results of the first and subsequent assessments of the SF-36 are profiled in Figure 12.1 as SF-36v2 *T* scores. There was nothing to suggest response inconsistency or any other data quality problems during any of the survey administrations.

The profile of Katherine's *T* scores from Visit 1 (see Figure 12.1) reveals that all of her physical health domain scores (PF, RP, BP, GH) fell into the borderline, or gray area, range (40–44), accounting for the moderate level of general physical impairment indicated by the PCS score (38). While the 80% CI for each of these scales extended the scores into the average range for individual respondents, the upper end of the CI for PCS (41) still suggested impairment in overall physical functioning. Upon investigation, the examining physician determined that, at least for the PF, RP, and BP scores, the respondent's knee pain contributed to these relatively low scores. With the exception of the VT scale, the MCS and the other mental health dimension domain scale scores were all within average range limits. The VT score (32) was consistent with her complaints of fatigue during the past year. Katherine's score on the Zung SDS (33) was also in the "normal" range. Visit 1 resulted in a treatment plan that called for her to continue on pain medication.

Katherine was seen again in December 2001 (Visit 2) for complaints of respiratory problems and a urinary tract infection (UTI). Sleep difficulties were also noted. During this visit, she was once again administered SF-36 and the Zung SDS and was also administered the Beck Anxiety Inventory (BAI) and the Epworth Sleepiness Scale (ESS). The SF-36v2–converted results included a *T* score of 41 on both the PCS and MCS measures, indicating a rise in the PCS score into the low-end of the borderline area and a significant deterioration in the MCS score from the average range during the previous 5 months. When investigating the reason for the increased PCS *T* score, it was noted that her GH and PF *T* scores did not change from the previous assessment. Her RP *T* score, however, was found to have significantly improved (49), moving into the average range, while her BP *T* score (37) had dropped significantly, falling into the impaired range as a result of increased interference

Figure 12.1 SF-36v2 Profile of Scores for Case 1

of pain in completing her work. Katherine attributed this lowered score to knee and neck pain, as well as to headaches.

A review of the mental health domain scales indicated that the significant drop in the MCS *T* score could be attributed to significant drops of 7, 11, and 15 points in the VT, SF, and RE scales, respectively. Upon questioning, sleep problems resulting in increased fatigue were found to account for the decrease in her VT score. Her ESS score of 14 suggested a high probability of sleep apnea. The 15-point drop in the RE *T* score reflected that she was not accomplishing all that she would like, which she attributed to conflicting feelings about caring for her diabetic father. Caring for her father, along with time spent working and caring for her children, allowed Katherine little time to be with friends and accounted for the drop in the SF *T* score. Although the rise in the MH *T* score (55) should be considered significant, it still remained in the average range. At the same time, her scores of 23 on the SDS and 37 on the BAI indicated the presence of severe anxiety without any significant depressive symptomatology. Thus, in this case, the MCS measure and VT, SF, and RE scales appear to have been more sensitive than the MH scale to the effects of Katherine's particular anxious symptomatology. Visit 2

ended with Katherine being diagnosed with bronchitis, cystitis, probable sleep apnea, and caregiver stress. She was continued on her pain medication, started on medication for the UTI, and referred to a sleep specialist to facilitate getting her on a continuous positive airway pressure (CPAP) device. She was also referred for assistance in caring for her father.

Katherine was again treated for UTIs in August 2002 and May 2003, but an SF-36 was not administered on either occasion. She was next seen in June 2003 (Visit 3) for a physical examination. At that time she complained of shoulder and leg aches and pain, noting that she had discontinued the previously prescribed pain medication due to intolerance. She also reported that she could not tolerate the CPAP intervention. She completed the SF-36v2, SDS, and ESS during this visit. For the first time, and despite her stated complaints to the physician, all of Katherine's physical health dimension (PCS, PF, RP, BP, GH) *T* scores fell in the average range, with all but the GH scale score representing a significant improvement from baseline.

In addition, while her MCS and RE *T* scores remained essentially unchanged at Visit 3, significant improvements were noted on the VT and SF scales, which rose into the average range. Although the ESS

score (13) again suggested sleep apnea, her VT *T* score showed improvement. This increase in her VT score was consistent with her report of sleeping better and having gotten through the seasonal stress related to her job. The RE scale *T* score was attributed to frustration over her son's legal and drug treatment problems and her father's deteriorating condition, which led to his undergoing dialysis, a second amputation, and placement in a nursing home. Again, the MH *T* score (50) was found to be in the average range, despite a significant drop from the previous assessment. Her SDS score (35) also remained in the normal range. At the end of Visit 3, the physician made two referrals for Katherine: one to an internist for intervention for the apnea, the other to an orthopedist for cervical disk disease evaluation related to left shoulder pain and right hand tingling.

During her next physical examination in August 2004 (Visit 4), Katherine complained of hot flashes and short-term memory loss. Since the last physical (Visit 3), her father had died, her 28-year-old son had had a drug relapse and had been sent to jail for theft, and she had begun psychotherapy to help her cope with the stress of her son's situation. Also, a glucose tolerance test had resulted in a diagnosis of type II diabetes. Her SF-36v2 PCS *T* score of 57 was significantly higher than her score from 14 months earlier, reflecting slight increases in the RP and GH *T* scores to 52 and 48, respectively. Although still in the average range, a significant drop in her BP *T* score (46) resulted from a reported increase in body pain over the previous 4 weeks.

Conversely, the MCS *T* score had dropped significantly into the impaired range (30), reflecting significant drops of 11 and 14 points in the RE and MH *T* scores, respectively. While the RE score change reflected accomplishing less work and doing so less carefully, the MH score was indicative of increased feelings of anxiety and depression. Katherine attributed the RE *T*-score change to being "freaked" over problems with short-term memory and concentration, as well as to her father's death and her son's situation. Also, her VT *T* score dropped slightly into the borderline area of interpretation (43). While her SDS score (40) indicated an increased presence of depressive symptomatology from the previous assessment, it still remained in the normal range. Similarly, her BAI score (32) remained in the severe anxiety range, representing some improvement over her previous BAI score that was obtained at the end of 2001 (Visit 2). Unlike previous assessments, her ESS score (6) indicated a low probability of sleep apnea. Katherine was maintained on her then-current treatment regimen and was also referred to a class to assist with learning to control her diabetes.

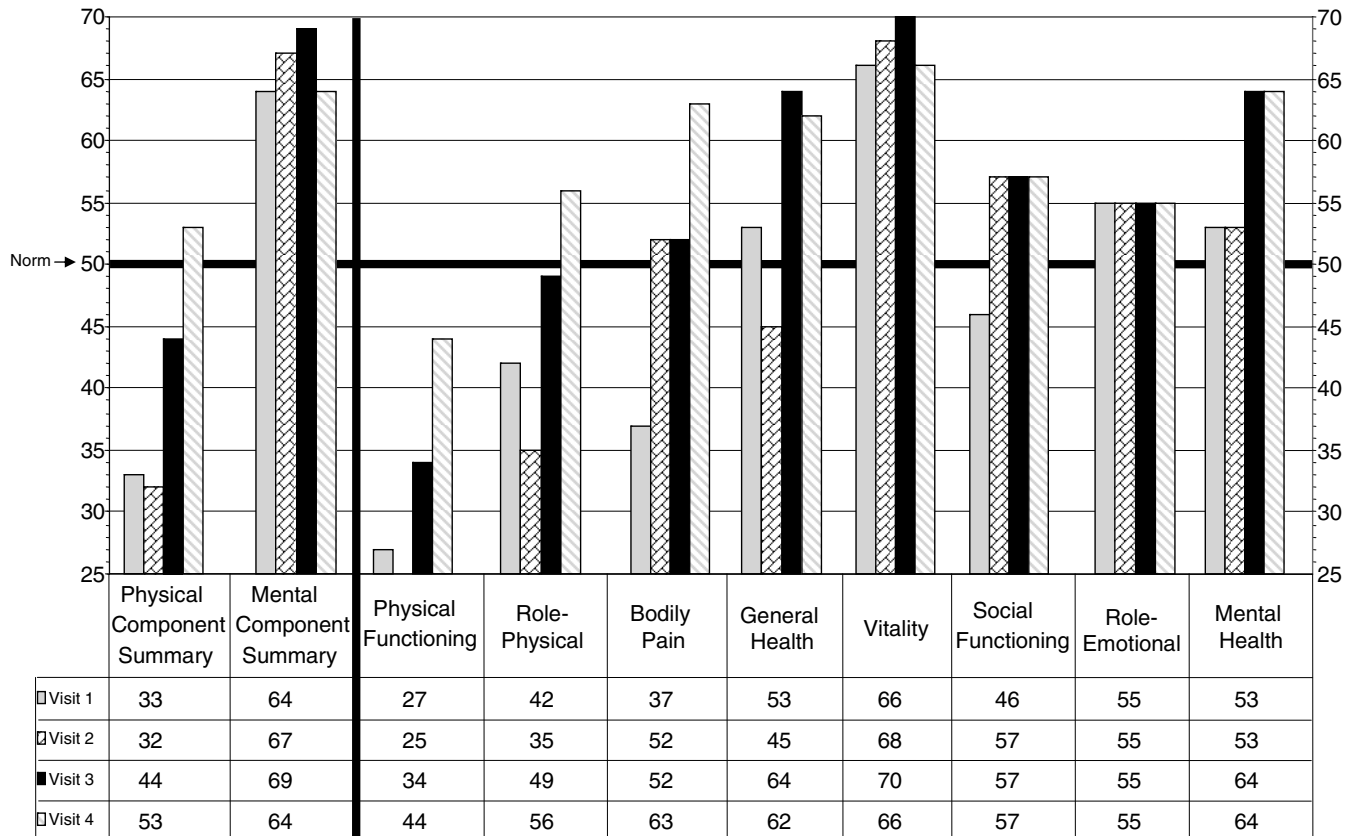
During the next 14 months, Katherine was seen on three occasions for follow-up visits regarding her diabetes and her blood pressure and cholesterol problems. She continued to have problems coping with her son's behavior, which resulted in a referral to a psychiatrist. During a follow-up appointment for her diabetes and short-term memory concerns in June 2005 (Visit 5), the SF-36v2 was administered once again. At this time, scores on all component summary measures and health domain scales were found to be in the average range or higher, with significant improvement indicated on all mental health dimension measures (MCS, VT, SF, RE, MH), as well as on the BP and GH scales. The physician attributed Katherine's improvement to her efforts to better care for herself and to better manage the situation with her son.

Over the 4 years this case was followed, Katherine displayed significant improvement on the PCS measure, all physical health domains, and the VT and SF scales. No significant improvement was noted on the MCS measure or the MH scale; however, *T* scores from both were within the average range at baseline. Although Katherine's RE *T* score significantly decreased over the 4-year period of measurement—at one point, by almost 3 standard deviations—it had returned to the average range (50) during the last measurement period.

Case 2

Abigail C. is a 69-year-old married homemaker who was initially seen by her family physician in May 2004 for a bleeding ulcer and depression. The ulcer had resulted from the use of nonsteroidal medication for hip pain secondary to severe osteoarthritis. Treatment for the ulcer was initiated and a selective serotonin reuptake inhibitor (SSRI) was prescribed for the treatment of her depression. She was seen 6 weeks later in July 2004 (Visit 1), at which time she completed the SF-36, the SDS, and the BAI. The results of this and three subsequent administrations of the SF-36, which were converted to SF-36v2 *T* scores, are presented in Figure 12.2. Overall, the findings are indicative of the effect of pain on Abigail's ability to carry out physical activities prior to her hip replacement surgery in September 2004 and of the improvement in physical functioning and amelioration of pain subsequent to the surgery.

Of the scores obtained during Visit 1 (see Figure 12.2), the relatively impaired physical functioning associated with Abigail's SF-36v2 scores on the physical health component measure and most physical health domain scales (PCS, PF, RP, and BP) stands out. This

Figure 12.2 SF-36v2 Profile of Scores for Case 2

holds true even when the CIs for the obtained *T* scores are taken into consideration, contrasting with her average to above-average *T* scores on the mental health component measure and scales (MCS, VT, SF, RE, MH). The results of this July 2004 assessment battery also included an SDS score of 38, indicating no significant depressive symptomatology, and a BAI score of 35, which is indicative of severe anxiety. Although none of the SDS, MCS, or MH *T* scores indicated the presence of clinical depression, the SSRI medication was continued to treat her anxiety symptoms.

Abigail was seen again in September 2004 (Visit 2) for her hip replacement surgery's required preoperative physical. The SF-36, SDS, and BAI were again administered during this visit. The SF-36v2 mental health *T* scores from this second administration remained essentially the same, with the exception of a significant increase (11 points) in Abigail's SF scale score. This 11-point increase was evidenced by the fact that physical health and/or emotional problems ceased interfering with her social activities. Also improved were her SDS and BAI scores (both 30); however, the BAI score was still indicative of severe anxiety. In terms of physical health, Abigail's BP *T* score increased significantly into the average range (52); at the same time, significant

declines were noted in the RP and GH *T* scores due to her inability to accomplish as much as she would have liked and her perception of herself as being less healthy than others. Note that Abigail's GH *T* score remained in the average range.

Abigail again completed the SF-36 survey during a postsurgical follow-up visit in December 2004 (Visit 3). Significant increases were noted in her PCS, PF, RP, and GH *T* scores; however, the PF score remained in the impaired range. Item analysis revealed that she had improved health perceptions and outlook, was less limited in most of the measured physical activities, and was able to work and accomplish more. Also noted was a significant increase in the MH *T* score (64) that reflected an increase in feelings of happiness accompanied by a decrease in feelings of nervousness and depression.

Abigail completed the SF-36 for a fourth time at her March 2005 office visit (Visit 4). With the exception of the PF scale, all component summary measure and health domain scale *T* scores fell within the average range or above, with significant increases over the previous assessment being noted for RP, BP, and PCS. The PF *T* score (44) rose significantly into the borderline area as a result of decreasing limitations in vigorous activities and increasing ability to climb several flights

of stairs, walk several blocks, and bend, kneel, or stoop. Although a significant decrease occurred in the MCS *T* score (64) from the previous visit, the score was still well above the average range of *T* scores for individual respondents.

This case illustrates the type of SF-36v2 results one might expect with an individual who is suffering primarily from a physically painful and limiting medical condition that is typically responsive to standard and effective means of treatment. The profile of scores

(see Figure 12.2) provides clear documentation of the positive effects of surgery on this woman's functioning, with clinically significant increases from baseline being evident in all but the MCS, VT, and RE *T* scores, which were each in the average to above average range to begin with. The treatment that Abigail received for her anxiety, as well as the support that was provided to her by the family practice staff, probably accounts for her ability to maintain an adequate level of emotional functioning throughout this episode of care.



**PART IV:
DEVELOPMENT AND
PSYCHOMETRIC
EVALUATION**

13

Development of the SF-36v2

The SF-36v2 is a revised and improved version of what has been one of the leading measures of health status for almost two decades—the SF-36. The “developmental” version of the SF-36 was first introduced in 1988 and was later followed by the “standard” SF-36 in 1990. Eventually published research and feedback from experienced users indicated that changes and enhancements would be highly desirable, necessary even, to make the instrument even more psychometrically sound and user friendly, worldwide. The result of these improvements was the SF-36v2, formally introduced in 2000. To understand the development of the SF-36v2, one must first understand the background and development of the original instrument, as it provides the context, rationale, and foundation upon which the revised instrument was developed.

This chapter begins with a brief overview of some of the well-accepted published standards that guided the development of all Short Form instruments. Then, an overview of the selection and development of the SF-36 items, health domain scales, and component summary measures, as well as the changes to the original instrument that were incorporated into the SF-36v2, are presented. Other advances that accompanied the development of the SF-36v2, such as the application of norm-based *T* scores and missing score estimation procedures, are also discussed. Note that Chapter 14 also addresses the advances made with the SF-36v2 via a detailed description of 2009 norming study, which provided the data for the current norms. Furthermore, evidence of the SF-36v2’s reliability and validity when using the 2009 data is provided in Chapters 15 and 16, respectively.

Published Standards for Psychometric Measures

The quality of health status assessment instruments began to receive much attention during the last half of

the 20th century. As such, a number of organizations, researchers, and clinicians have published standards that they believe should guide the development, evaluation, and use of psychometric measures. Many of these published sets of standards address the same issues (e.g., validity, reliability) but provide little in terms of detailed criteria by which to judge the adequacy of a specific instrument for a specific purpose. However, there are exceptions, among which the standards developed by the American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), and the Medical Outcomes Trust (MOT) are the most notable and relevant to health status and quality of life assessment instruments. Also included in this group are the guidelines recently developed by the U.S. Food and Drug Administration.

Standards for Educational and Psychological Testing

In 1999, the AERA, APA, and NCME published the latest version of the *Standards for Educational and Psychological Testing (Standards)*, the sixth in a series of such publications (the first of which was published in 1954) whose intent is to guide the development and use of tests (AERA, APA, & NCME, 1999). For 50 years, these standards have guided developers and users of tests, surveys, and other psychometric measures by providing “criteria for the evaluation of tests, testing practices, and the effects of test use,” thus acting as a “frame of reference to assure that relevant issues are addressed” (AERA, APA, & NCME, 1999, p. 2). This comprehensive, detailed set of standards guided the development of all Short Form instruments and their accompanying user’s manuals, not only to conform to the expectations of the scientific and clinical communities but also to ensure that these instruments provide meaningful contributions to the field of health status and

quality of life measurement. Note that the *Standards* was in the process of being reviewed and revised at the time of this manual's publication.

Medical Outcomes Trust Instrument Review Criteria

In 1995, the Scientific Advisory Committee (SAC) of the Medical Outcomes Trust (MOT) published a set of criteria for evaluating instruments submitted to the MOT for inclusion in its library of approved measures (Scientific Advisory Committee of the Medical Outcomes Trust, 1995, 1996). The criteria identified by the SAC included the following eight attributes, against which instruments would be judged: a conceptual and measurement model, reliability, validity, responsiveness to change, interpretability, respondent and administrative burden, alternative forms, and cultural and language adaptations. The SAC noted that the importance of each criterion would depend on the instrument's intended use or application. The MOT criteria were subsequently revised, making them easier for developers to apply to their specific circumstances and more applicable to instruments developed using modern test theory principles and methods (SAC of the MOT, 2002). Although not as detailed as the *Standards for Educational and Psychological Testing*, the MOT criteria can serve as clear and concise guides for those seeking to identify psychometrically sound instruments for research or clinical applications.

U.S. Food and Drug Administration Guidelines

In 2006, the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (DHHS) circulated for comment a draft version of its recommendations for the development and use of patient-reported outcomes (PRO) measures, recommendations that would be used to support claims made in describing and summarizing the use, safety, and effectiveness of medical products (U.S. Department of Health and Human Services, 2006). Like the other sets of test development and use guidelines, the FDA's draft document drew attention to the importance and indicators of an instrument's reliability, validity, interpretability, and ability to detect change. Just as vital, according to the FDA, is a solid conceptual framework with identified concepts and domains that are important to patients, as well as are the hypothesized relationships between these concepts and domains. The FDA's document, which also provides justifications and considerations for modifying existing PRO instruments and for developing instruments for specific populations, was finalized and published in 2009 (U.S. DHHS, 2009).

These three sets of guidelines reflect the classical test development theory that guided the development of the SF-36v2, and it is against these standards that the survey will be evaluated in this and subsequent chapters of this manual. Furthermore, as the development of Short Form instrumentation proceeds currently and in the future, techniques of modern test theory will continue to be employed, and it is against these techniques and accompanying standards that the psychometric integrity of these instruments will then be evaluated.

Background

Interest in the use of brief health surveys became a need during the Health Insurance Experiment (HIE; see Chapter 1) when some of the study's participants refused to complete traditional, lengthy health surveys (Ware, Brook, et al., 1980). To ensure that these individuals would not be lost to follow-up, a very short survey was developed that could be administered by telephone in approximately 5 minutes. This strategy to gain the participants' cooperation worked well and yielded preliminary data supporting the use of short-form scales. Subsequently, several short-form scales were successfully used in various studies (Brook et al., 1987; Davies & Ware, 1981; Fowler et al., 1988; Lurie, Ward, Shapiro, & Brook, 1984; Nelson et al., 1983; Nelson & Berwick, 1989; Read, Quinn, & Hoefler, 1987).

Other analyses of HIE data demonstrated the value of a compromise in the search for brevity. For example, a well-constructed, multi-item scale, even with only 5 to 10 items, achieved better validity in predicting subsequent medical expenditures than a single-item measure. Those analyses also demonstrated that longer scales and more comprehensive questionnaires achieved higher levels of validity in predicting subsequent medical expenditures than relatively short, multi-item scales (Manning, Newhouse, & Ware, 1982). These findings underscored the trade-offs involved when choosing between short and long scales.

In 1984, an attempt was made to construct a comprehensive, short-form health survey that consisted of 18 items measuring physical functioning, role limitations due to poor health, general mental health, and current health perceptions. Constructed for a 1984 national survey fielded by Louis Harris and Associates (Montgomery & Paranjpe, 1985), this survey was developed from items that had been successfully used in previous studies (see Ware, Sherbourne, & Davies, 1992). In 1986, two items measuring social functioning and bodily pain were added to the 18-item survey, creating a

20-item short form that would later be referred to as the SF-20 (Ware, Sherbourne, & Davies, 1992). In the Medical Outcomes Study (MOS), the SF-20 was administered to 11,336 participants who were sampled from 523 medical practices in Boston, Chicago, and Los Angeles. The resulting cross-sectional data sets were used to perform psychometric evaluations, develop preliminary norms, and test the usefulness of the SF-20 scales in detecting differences in functional status and well-being amongst patients with chronic medical and psychiatric conditions (Stewart et al., 1989; Stewart, Hays, & Ware, 1988; Ware, Sherbourne, & Davies, 1992; Wells et al., 1989).

In the MOS, the SF-20 was used for screening purposes and for norm-based comparisons made at baseline. Its usefulness, however, was limited. With the exception of the 5-item version of the Mental Health Inventory (MHI-5; Veit & Ware, 1983) that was included in the SF-20, noteworthy problems were evident in the instrument's scales and several important health domains were not being assessed. These and other findings provided considerable experience with balancing the trade-offs between the breadth of constructs represented and the depth of measurement for each construct when developing short-form health status surveys. Moreover, since the 18-item and 20-item short forms were first used, strategies had been identified for improving the precision of short-form scales in measuring health-related constructs. Although problematic in several aspects, some items from the SF-20 served as bases for the items that were developed for the SF-36, while other items were dropped (discussed in later sections of this chapter; see also Ware & Sherbourne, 1992).

Prior to the final selection of concepts and specific items for the SF-36, versions of candidate items were embedded throughout the 149-item Patient Assessment Questionnaire (PAQ) that was longitudinally administered in the MOS (Stewart & Ware, 1992). The SF-36 items differ from the original MOS PAQ items in a number of important respects. First, the original MOS PAQ versions of the 36 items were much longer than the SF-36 versions (844 vs. 677 words, or about 25% longer). The wording of SF-36 directions and items was shortened by adopting a more efficient response grid format and by not repeating instructions to respondents as often as was done in the MOS PAQ. For example, field tests confirmed that questions 9a through 9i could be substantially shortened without a loss of data quality.

Following the selection of concepts and scales and the editing of original MOS PAQ items and directions, a developmental (prepublication) version of the SF-36 was made available for testing in late 1988. For example,

InterStudy utilized this version for testing under its Outcomes Management System (OMS) program, as did numerous other investigators and projects (Ware, 1988). In the fall of 1990, Ware and his colleagues at The Health Institute at Tufts-New England Medical Center finalized the content and format of the standard version of the SF-36. Changes from the developmental version included referring to *the past four weeks* rather than *the past month*, a change made in response to ambiguity reported by field test respondents regarding whether the recall period was the number of days that had elapsed in the current month (e.g., the previous nine days if taken on the tenth day of a month) or a period of time 4 weeks in duration. Numerous other changes were adopted based on respondents' comments gathered during the 2-year testing period, including the underlining or boldfacing of key words in the instructions, questions, and response choices. In addition, many other individuals and organizations contributed useful suggestions for improvement (see Acknowledgements section of this manual). Subsequent to the SF-36's publication, the need for changes to the instrument resulting from work conducted as part of the International Quality of Life Assessment (IQOLA) Project (see Chapter 1) led to the development of the second version of the survey, the SF-36v2. These changes are discussed in detail in later sections of this chapter.

Conceptual Framework

To be certain, excluding important health constructs can shorten a health status survey. However, minimum standards of comprehensiveness (i.e., content validity in relation to accepted definitions of health) argue for representation of both physical and mental health constructs and for multiple manifestations of functioning and well-being for each concept (Ware, 1987, 1990a). Based on these standards and empirical work to date, multiple categories of operational definitions were chosen to measure each health construct: (a) behavioral functioning, (b) perceived well-being, (c) social and role functioning, and (d) personal evaluations (perceptions) of health in general. Table 13.1 presents the physical and mental health phenomena represented by each scale of both versions of the SF-36.

Self-reports of behavioral functioning are widely used to measure limitations due to poor health and/or bodily pain in physical, social, and role activities. These indicators often focus on observable and tangible standards external to the individual, such as walking a specific distance or performing customary self-care behaviors. Perceived well-being is more subjective and refers to how an individual feels. Well-being is a

Table 13.1*Summary of Health Phenomena Captured by the SF-36 and SF-36v2 Health Domain Scales*

Scale	Physical Health Phenomena				Mental Health Phenomena			
	Functioning	Well-Being	Disability	Personal Health Evaluation	Functioning	Well-Being	Disability	Personal Health Evaluation
Physical Functioning	•							
Role-Physical			•					
Bodily Pain		•	•					
General Health				•				•
Vitality		•				•		
Social Functioning			•				•	
Role-Emotional							•	
Mental Health					•	•		

psychological state that cannot be completely inferred from observable behavior (Ware, 1987, 1990a). For both versions of the SF-36, perceived well-being was defined in terms of well-proven self-reports of the frequency and intensity of feeling states, including general mental health (psychological distress and psychological well-being), bodily pain, and vitality (energy and fatigue).

A comprehensive and valid health survey must also reflect the values or preferences of the individual respondent, for who else is more qualified to evaluate current health status or expectations for health in the future? As such, perceptions of health in general (i.e., personal evaluation of current health status, susceptibility to illness, and health outlook) were also included in the SF-36 surveys. It is well documented that such evaluations provide good summaries of health status and reflect the impact of specific symptoms and other health states that are not explicitly captured by measures found in the other three categories (Davies & Ware, 1981).

Selection and Origin of Items

The content of the items contained in both versions of the survey will seem very familiar to those who follow the literature on health assessment. Many of the selected items have their content roots in instruments that have been in use for more than 30 years. In addition to the questionnaires referenced in Table 13.2, the content of other historical instruments is described in a number of documents, including the development of the MOS measures detailed in Stewart & Ware (1992). Other useful publications discuss measures of limitations in physical, social, and role functioning (Donald & Ware, 1984; Stewart, Ware, & Brook, 1981; Stewart, Ware, Brook, & Davies-Avery, 1978); general mental health (Veit & Ware, 1983; Ware, Johnston, Davies-Avery, & Brook, 1979); and general health perceptions (Davies & Ware, 1981; Ware, 1976a; Ware & Karmos, 1976a, 1976b).

The most difficult task in developing the SF-36 was the selection of a subset of eight health constructs from the more than 40 constructs and scales studied in the MOS. Among those seriously considered, but not chosen, were measures of health distress, cognitive functioning, sexual functioning, family functioning, and sleep adequacy.

Health Domain Scales

As previously noted, the SF-36v2 represents a new generation of instruments utilizing a measurement approach that has proven valuable to researchers and clinicians for more than a decade. The SF-36v2 includes one scale measuring each of eight health domains: physical functioning, role limitations due to physical health problems, bodily pain, general health, vitality (energy/fatigue), social functioning, role limitations due to emotional health problems, and mental health (psychological distress and well-being). The conceptual origins of these eight scales are presented in Table 13.2.

A major problem in the field has been the absence of agreed-upon criteria for constructing and validating health scales. As such, when selecting items for each health domain scale, the corresponding full-length MOS scale was used as the criterion. Each SF-36 health domain scale's items were selected to reproduce that "parent" scale as much as possible. Note that other psychometric standards were considered as well. Specific strategies for selecting the SF-36 items, which varied across concepts, are summarized in the following sections. Aside from the item modifications discussed in this chapter, the two versions of the SF-36 contain the same basic items.

Physical Functioning (PF)

Because of the importance of distinct aspects of physical functioning and the necessity of sampling a

Table 13.2*Conceptual Origins of Short Form Survey Content*

Scale	Based on	References
Physical Functioning	Canadian Sickness Survey Index of Well-Being Functional Status Index Functional Limitations Index Functional Status Assessment Duke-UNC Profile	Cameron (1954); Hatcher (1956) Patrick, Bush, & Chen (1973) Jette (1980, 1987) Berdit & Williamson (1973) Deniston & Jette (1980) Parkerson et al. (1981)
Role-Physical	Sickness Impact Profile	Bergner, Bobbitt, Carter, & Gilson (1981)
Bodily Pain	Wisconsin Brief Pain Questionnaire	Daut, Cleeland, & Flannery (1983)
General Health	National Health Interview Survey Health Perceptions Questionnaire	National Center for Health Statistics (1976) Davies & Ware (1981); Ware (1976b)
Vitality	General Well-Being Schedule	Dupuy (1973)
Social Functioning	MOS-FSWBP	Donald, Ware, Brook, & Davies-Avery (1978)
Role-Emotional	Sickness Impact Profile	Bergner et al. (1981)
Mental Health	General Well-Being Schedule Mental Health Inventory	Dupuy (1973) Ware, Johnston, Davies-Avery, & Brook (1979)

range of severe and minor physical limitations, the full-length (10-item) MOS Physical Functioning scale was adopted without modification. This scale reflects two important improvements over previous health status questionnaires such as the SF-20. First, more items were utilized to better represent levels and types of limitations between the extremes, including lifting and carrying groceries; climbing stairs; bending, kneeling, and stooping; and walking moderate distances. As with the SF-20 and HIE versions of this scale, only one self-care item was included to represent limitations in such activities and to define the floor of the scale. Although limitations in self-care activities are very important and can be measured in considerable detail (Katz, Downs, Cash, & Grotz, 1970; Katz, Ford, Moskowitz, Jackson, & Jaffe, 1963), they are relatively rare in both general and patient populations (Stewart et al., 1988; Stewart, Ware, & Brook, 1981; Stewart et al., 1978; Stewart, Ware, & Brook, 1982a, 1982b; Ware, Sherbourne, & Davies, 1992). For example, of the 11,336 patients screened in doctors' offices for the MOS, only 7.4% reported any limitations in self-care activities. Thus, the routine administration of a lengthy battery of self-care items was deemed inefficient for the purposes of a general health survey.

Second, standardized response choices were revised to estimate the severity of each limitation, thereby increasing the precision of scores. The HIE and the SF-20 measured the duration (more or less than 3 months) of any reported limitation. However, because the great majority of physical limitations are chronic, measures of duration proved to be of little value in data analysis and

had been ignored in the scoring of HIE items for over 10 years (Stewart, Ware, & Brook, 1981). Methodological comparisons revealed that the distinction between those who are able to perform physical activities with and without difficulty is more useful in increasing precision (Stewart & Kamberg, 1992). Some performance-based measures ignore this distinction (Kaplan & Anderson, 1988), while others do not (Jette et al., 1986). The SF-36v2 items capture both the presence and extent of physical limitations using a three-level response continuum. With this three-level response scale, the number of scale levels defined by the 10 survey items was doubled (relative to the number achieved with dichotomous items) and the precision of hypothesis testing was increased, all without adding to respondent burden.

Role-Physical (RP)

Both versions of the SF-36 include a subset of the 11 role-functioning items found on MOS long forms. The items selected differ from other widely used surveys in two important respects (Stewart & Ware, 1992). First, they cover a rich array of role limitations, including (a) limitations in kind of work or other usual activities, (b) reduction in the amount of time spent doing work or other usual activities, and (c) difficulty performing work or other usual activities. Thus, in addition to defining more levels of role limitations due to health problems, the Role-Physical scale is more applicable than other surveys to retired individuals and those with more than one usual role. Second, the RP items discriminate between role limitations due to physical health and those due to mental health, thus allowing the RP scale to measure

role limitations due to physical problems with improved precision in discriminating amongst groups known to differ in medical conditions (Hays & Stewart, 1990). Note that the SF-36v2 RP scale has an advantage over its SF-36 counterpart due to the inclusion of five-level response choices for each of its four items in place of dichotomous (*yes or no*) response choices.

Bodily Pain (BP)

Both versions of the SF-36 contain an item about the intensity of bodily pain or discomfort (taken from the SF-20) and a second item measuring the functional impact of pain in terms of the extent of its interference with normal activities. The latter item was chosen because it was the best predictor ($r = .84$) of the total score for the Behavioral Effects of Pain scale used in the MOS (Stewart & Ware, 1992). The result is a gain in content validity, scale reliability, and precision (i.e., an 11-level scale vs. a 6-level scale) relative to a single pain item (McHorney, Ware, Rogers, Raczek, & Lu, 1992).

Note that the two Bodily Pain health domain scale items offer unequal numbers of response choices (six for Item 7 and five for Item 8). As a result, their variances are not equal, as required for a summated rating scale. Further, in all MOS studies published to date, Item 8 was administered (following a skip pattern) only to those respondents reporting at least some pain. Although the MOS skip pattern was dropped to make the SF-36 easier to administer, this dependence between responses must be taken into account when comparing results from new studies with previously published data.

In studies conducted during the HIE, Davies and Ware (1981) reported that recalibration of the pain severity rating was necessary to satisfy the equal interval assumption. MOS studies have confirmed that the relationship between Item 7 and criterion measures of pain significantly departs from a linear association. Criterion pain measures used in these tests include visual analogue scales measuring pain severity and categorical ratings of pain frequency and duration. Final response values for Item 7 were derived from the mean values of a summary MOS criterion pain measure computed for respondents who chose each of the six levels defined by this item, using methods much like those used for Item 1 in the General Health scale (discussed in the following section).

The scoring rules recommended for the BP scale are based on three considerations: (a) the items offer both different amounts and different content of response choices, (b) administration of Item 8 in the MOS depended on the response to an item similar to Item 7, and (c) empirical studies have indicated that recalibration of

Item 7 is necessary to achieve a linear fit with the scale score and with other measures of functioning with pain. The recommended recoding of the first response choice for Item 8 on the basis of the response to Item 7 solves two problems. First, it converts Item 8 to a six-level item of roughly equal variance to Item 7. This is achieved by splitting those respondents who report being free of role interference due to pain into two different groups: (a) free of interference and free of pain (the best level) and (b) free of interference but with at least some pain (the next best level). Second, it approximates the dependence between the two items found in published MOS studies (McHorney et al., 1992; McHorney, Ware, & Raczek, 1993).

General Health (GH)

Both versions of the SF-36 combine the widely used single-item rating of health (*excellent to poor*) with four items from the Health Perceptions Questionnaire (HPQ; Davies & Ware, 1981; Ware, 1976a). As a result, the General Health scale (a) achieves an adequate sample of the content of the HPQ (current health, resistance to illness, and health outlook) and (b) correlates highly ($r = .96$) with the 22-item General Health Rating Index (GHRI; Davies & Ware, 1981; Ware, Davies-Avery, & Donald, 1978) constructed from the HPQ. Further, the GH scale strikes a good balance between favorably and unfavorably worded items, which controls for response-set effects.

Substantial empirical evidence of validity has accumulated for the GHRI (Davies & Ware, 1981; Stewart & Ware, 1992). Specifically, the pattern of correlations between the scale's summary score and other health measures has been quite consistent with hypotheses (Davies & Ware, 1981), and the GHRI differentiates the impact of serious and minor acute symptoms (Shapiro, Ware, & Sherbourne, 1986). In addition, it is a good predictor of medical care expenditures (Manning et al., 1982) and return to work after a heart attack (Smith, Monson, & Ray, 1986). The GHRI also proved useful in detecting health outcomes in the HIE (Ware et al., 1986; also see Chapter 1).

In terms of the SF-36 surveys, the *very good* and *good* responses to Item 1 have been recalibrated to achieve a better linear fit with the general health evaluation concept measured by the GH scale. Empirical studies during the HIE were among the first to document that the intervals between response choices for this item were not equal (Davies & Ware, 1981). Subsequent studies of this item, using the Thurstone method of equal-appearing intervals (Thurstone & Chave, 1929) and other empirical methods, have consistently shown that the interval

between the *excellent* and *very good* response choices is about half the size of the interval between *fair* and *good* (Ware, Nelson, Sherbourne, & Stewart, 1992). These results have been confirmed in studies of SF-36 translations from 10 countries participating in the IQOLA project (Keller et al., 1998; see also Chapter 1). Finally, in all studies known to date, mean values for a criterion general health scale for respondents who choose each of the five levels defined by Item 1 departed significantly from linearity.

Table 13.3 summarizes the MOS results that served as the basis for the recommended recalibration of Item 1. As shown in Table 13.3 and discussed elsewhere (Ware, Nelson, et al., 1992), the mean criterion scores were remarkably similar for those who chose the same category of Item 1 across the screening ($N = 18,573$) and longitudinal ($N = 3,054$) samples. Also note that intervals between adjacent response categories were unequal, as was observed in the HIE (Davies & Ware, 1981). For these reasons, item scale values were transformed, using specific results from the screening sample. The result was a very high correlation of .70 with the sum of the other four items in the GH scale.

Vitality (VT)

The four items of the Vitality scale (assessing energy level and fatigue) have an impressive track record in terms of empirical validity, striking a balance between favorably and unfavorably worded items to control for response-set effects. These VT items were adapted from the MHI (Veit & Ware, 1983), which was fielded in the HIE. The MHI was derived from the 1976 HANES survey by the National Center for Health Statistics (Dupuy, 1984). These studies yielded thorough

evaluations of the Vitality scale's psychometric properties and documented its item-discriminant validity and scale reliability. Moreover, this health domain scale's sensitivity to the impact of disease and treatment has been demonstrated in clinical trials involving patients with hypertension (Croog et al., 1986), prostate disease (Fowler et al., 1988), and those differing in severity of AIDS (Watchel et al., 1992; Wu et al., 1991). The advent of the SF-36v2 resulted in the elimination of one of the SF-36 VT scale's response choices (*a good bit of the time*), which was not found to consistently scale across countries (Keller et al., 1998).

Social Functioning (SF)

In contrast to physical and mental health concepts that tend to "end at the skin" (Ware, 1986, p. 206), the Social Functioning health domain scale extends measurement beyond the individual respondent to capture both the quantity and quality of social activities. The SF-36v2 retains an improved version of a social functioning item from the SF-20 and includes a second item. These two items, a subset of the long-form social functioning items developed for the MOS, assess health-related effects on social activities (Stewart & Ware, 1992). Most measures of social activities ask respondents to report the number of contacts and activities or frequency of participation in different activities (Donald & Ware, 1984). Furthermore, such measures usually do *not* ask respondents to indicate whether their social activities have been affected by their own health problems. Thus, most of the variation reported in social activities reflects non-health-related factors (Stewart, Hays, & Ware, 1988). However, the SF-36v2 items specifically ask about the impact of either physical health or emotional problems on social activities to measure health outcomes.

Role-Emotional (RE)

The SF-36v2's role-functioning scales define two sets of items that distinguish between role limitations due to physical health and those due to mental health. The former are assessed by the RP scale (see previous section); the latter, sometimes overlooked by surveys that do not explicitly ask about limitations due to emotional problems (McHorney et al., 1992; Stewart & Ware, 1992), are assessed by the Role-Emotional health domain scale. This RE scale discriminates well between groups known to differ in psychiatric conditions (Sherbourne et al., 1992). Similar to the RP scale, all three of the the SF-36v2 RE scale's items utilize five-level response choices, in place of the SF-36's dichotomous (*yes* or *no*) response choices.

Table 13.3

Mean Current Health Scores for Respondents Choosing Each Level of SF-36v2 Item 1

Response to Item 1	Mean Current Health		Recommended Scoring	
	Screening Sample ($N = 18,573$)	Baseline Sample ($N = 3,054$)	1-5 Scale	0-100 Scale
<i>Excellent</i>	87.9 (=5)	86.9 (=5)	5.0	100
<i>Very good</i>	75.5 (4.36)	75.4 (4.40)	4.4	85
<i>Good</i>	57.6 (3.43)	55.9 (3.37)	3.4	60
<i>Fair</i>	30.0 (2.00)	30.6 (2.04)	2.0	25
<i>Poor</i>	10.8 (=1)	10.8 (=1)	1.0	0

Adapted from "Preliminary Tests of a 6-Item General Health Survey: A Patient Application," by J. E. Ware, Jr., E. C. Nelson, C. D. Sherbourne, and A. L. Stewart, 1992. In *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach*, A. L. Stewart and J. E. Ware, Jr. (Eds.), p. 299. Durham, NC: Duke University Press. Copyright 1992 by Duke University Press.

Mental Health (MH)

For the SF-36v2, the five-item version of the Mental Health Inventory (MHI-5; Veit & Ware, 1983) that was used in the SF-20 was retained, with modifications only to format. The MHI-5, in use for two decades (Berwick et al., 1991; Croog et al., 1986; Fowler et al., 1988; Read, Quinn, & Hoefler, 1987; Stewart, Hays, & Ware, 1988; Stewart & Ware, 1992; Wachtel et al., 1992; Wu et al., 1991), was constructed from the five items that best predicted the summary score for the 38-item MHI. The MHI-38, which served as the “gold standard” in constructing the MHI-5, is discussed elsewhere (Davies, Sherbourne, Peterson, & Ware, 1988; Veit & Ware, 1983; Ware et al., 1979). Furthermore, evidence of the MHI-38’s empirical validity includes published studies of (a) groups of patients known to differ in medical and psychiatric conditions (Cassileth et al., 1984; Dupuy, 1984; Smith et al., 1986); (b) predictive validity in terms of subsequent utilization of mental health services (Ware, Manning, Duan, Wells, & Newhouse, 1984), utilization of general medical services (Manning et al., 1982), and mental health measured after 3 years (Williams, Ware, & Donald, 1981); (c) the negative impact of stressful life events and the utility of social supports (Williams et al., 1981); (d) construct validity based on factor analysis (Cassileth et al., 1984; Ware, Davies-Avery, & Brook, 1980); and (e) correlations with other health status measures (Cassileth et al., 1984; Dupuy, 1984; Nelson et al., 1983; Read et al., 1987).

The MH health domain scale includes one or more items from each of the four major mental health dimensions (anxiety, depression, loss of behavioral/emotional control, and psychological well-being), confirmed in factor-analytic studies of the full-length MHI (Veit & Ware, 1983). Further, the simple sum of the five MH items (without weights) correlates well (.95) with the full-length MHI-38. A correlation of .93 was found on cross-validation, using data from the HIE. As with the VT scale, one of the SF-36 response choices (*a good bit of the time*) was eliminated during the SF-36v2’s development.

Self-Evaluated Transition (SET)

In addition to the five items in the GH scale, both versions of the SF-36 include a sixth general health rating item that asks respondents about the amount of change in their health, in general, over a 1-week period (acute form) or a 1-year period (standard form). This item, called the *Self-Evaluated Transition* item (SET; formerly referred to as the Reported Health Transition item), is not used to score any of the eight multi-item

scales; however, it can be analyzed as a categorical variable or as an ordinal- or interval-level scale for research or clinical purposes. The SET item has also proven to be useful in longitudinal studies, in studies of the importance of changes (better or worse) to individual patients, and in the prediction of death and the trajectory of health.

Scale composition, number of score levels, highest and lowest possible *T* scores, and meanings of the highest and lowest scores for each of the eight SF-36v2 health domain scales and the SET item are summarized in Table 7.1.

Differences Between the SF-36 and the SF-36v2

The SF-36v2 was derived from more than 8 years of experience with the SF-36 and from the findings of the thousands of publications that had accumulated by the time of its development. In particular, many lessons were learned from the translation studies conducted as part of the IQOLA Project (see Chapter 1). As a result of these studies, words that did not translate well from English to other languages were identified and replaced with words that did; double negatives were eliminated from item wording; and the response option *a good bit of the time* was eliminated from MH and VT items because it did not consistently pass psychometric tests across translations (Keller et al., 1998). In 1998, a national calibration study, designed to evaluate the effects of all the survey improvements and to ensure that comparable scores could be computed across versions, was completed in the United States. Note that some of the results from this study are presented later in this chapter.

As noted in Chapter 1, relative to its predecessor, improvements found in the SF-36v2 include:

- Revised wording of instructions and survey items, designed to shorten and simplify the text, making it more familiar and less ambiguous.
- Redesigned layout for questions and response choices in the self-administered survey form, making them easier to read and complete and thus reducing the number of missing responses.
- Greater comparability with the translations and cultural adaptations widely used in the United States and in other developed countries.
- Five-level response choices in place of dichotomous (*yes* or *no*) response choices for the items in the two role-functioning scales (RP and RE).

- Five-level response choices in place of six-level response choices, designed to simplify the items in the MH and VT scales.
- Adoption of the *T*-score metric for both the health domain scales and the component summary measures, based on 1998 U.S. general population data (with 2009 U.S. general population norms now available).

Layout

The layout of the SF-36v2 standard and acute forms is based on the cognitive design principles for formatting HRQOL instruments recommended by Mullin, Lohr, Bresnahan, and McNulty (2000). All responses choices are printed in a left-to-right (horizontal) format, rather than the mixture of horizontal and vertical listings found in the SF-36. Mixed formats of response choices can confuse respondents and lead to missing and inconsistent responses, particularly among older populations. Other improvements to the form's layout include more consistent use of indenting, numbering of instructions, deletion of item labels, formatting of the response boxes that are marked by respondents, and more "white space," which makes the text easier to read, particularly for individuals with visual impairment.

Type Size and Bolding

A larger type size has been adopted throughout the survey form. In addition, instructions and questions, but *not* response choices, are printed in bold text to simplify the look and feel of the SF-36v2. These and other refinements have been incorporated on the basis of lessons learned from health care surveys, as well as from surveys in other fields.

Wording Changes

Evidence from qualitative research, including numerous focus group studies and formal cognitive tests, and from empirical studies conducted in more than a dozen countries supported the improvements made in item wording and the changes in the terms used to describe health status in the SF-36v2. These improvements were designed to make the English-language SF-36v2 easier to understand and administer, as well as making it more objective. In addition, the SF-36v2 is more comparable with translations of the survey. Because most of the improvements in item wording were developed during the process of translating and adapting the SF-36 for use in other countries during the IQOLA Project, the SF-36v2 is also referred to as the "international" version.

A comparison of Short Form item wording from the original MOS PAQ (Stewart & Ware, 1992) through to the SF-36v2 is presented in Table 13.4. Note that the SF-

36, as such, was never administered in its final form in the MOS. The column labeled *Original MOS PAQ Items* lists the verbatim content and location of candidate items that were embedded throughout eight sections of the baseline MOS PAQ. Numerous changes to the original MOS PAQ versions of SF-36 items and instructions were adopted for the developmental (pre-publication) form; likewise, substantial changes were necessary when making the transition from the developmental version to the standard version (i.e., the SF-36). Given the widespread adoption of the developmental version at the time, a very high priority was placed on maintaining comparability between these two surveys.

Table 13.4 also shows that the differences between the developmental version and the original SF-36 are fewer than those found between the SF-36 and the SF-36v2, which is a result of the changes that were incorporated into the SF-36v2 to make it an improvement over its predecessor. Table 13.5 presents a brief summary of these changes in item wording from the SF-36 to the SF-36v2.

Five-Choice Response Scales

Although the two role-functioning scales measure what is arguably one of the most important health outcomes—participation in usual role activities—these scales are the most coarse and least precise of the SF-36 scales. The reasons for this include their reliance on dichotomous (binary) response choices and the relatively narrow range of levels of functioning that they cover. The SF-36v2 development team had known for nearly 8 years that categorical rating scales would solve these problems; however, the issue was the choice of *which* categorical rating scale to use.

Studies aimed at solving these problems with the role-functioning scales were initiated in the early 1990s (see Kantz, Harris, Levitsky, Ware, & Davies, 1992). These studies began by investigating the SF-36 role-functioning questions, using a combination of yes/no response choices and categories of impairment (e.g., yes, all of the time; yes, most of the time; yes, some of the time; yes, a little of the time; no, none of the time). The hope was that the responses to the categorical role-functioning scales, when scored dichotomously (yes/no), would be comparable with the original (yes/no) responses for these scales and that the new version could also be scored using the five rating scale categories for each item. Unfortunately, comparability was not achieved because test subjects responded differently to the dichotomous choices when they were combined with other categories of impairment, and the match between the original questions and response categories of impairment was poor. Further, this approach did not eliminate

Table 13.4

Comparison of Items in the SF-36v2 Health Survey, SF-36 Health Survey, Developmental Version of the SF-36 Health Survey, and Original MOS PAQ

Item Number	Scale	SF-36v2	SF-36	Developmental (Pre-Publication) Version	Original MOS PAQ Items
Overall Instructions		<p>This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities. <i>Thank you for completing this survey!</i></p> <p>For each of the following questions, please mark an <input type="checkbox"/> in the one box that best describes your answer.</p>	<p>INSTRUCTIONS: This survey asks for your views about your health. This information will help keep track of how you feel and how well you are able to do your usual activities.</p> <p>Answer every question by marking the answer as indicated.^a If you are unsure about how to answer a question, please give the best answer you can.</p>	<p>INSTRUCTIONS: This survey asks for your views about your health. This information will be summarized in your medical record and will help your doctors keep track of how you feel and how well you are able to do your usual activities.</p> <p>Answer every question by circling the appropriate number, 1, 2, 3, ... If you are unsure about how to answer a question, please give the best answer you can and make a comment in the <u>left margin</u>.</p>	<p>1. Please answer every question (unless you are asked to skip questions because they don't apply to you). Some questions may look alike, but each one is different.</p> <p>2. Answer the questions by circling the appropriate number 1 2 or filling in the answer as requested.</p> <p>3. If you are unsure about how to answer a question, please give the best answer you can and make a comment in the left margin. We will read all your comments, so feel free to make as many as you wish.</p>
Question 1	GH	In general, would you say your health is:	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 373, Section 1: Health and Daily Activities, Q2*</i>
Q1 Responses	GH	Excellent Very good Good Fair Poor	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2
Question 2	SET	<u>Compared to one year ago</u> , how would you rate your health in general <u>now</u> ? ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Compared to one year ago, how would you rate your health in general <u>now</u> ? ^c <i>12-month PAQ, Page 3, Section 1: Your Health Now Compared to One Year Ago, Q6</i>
Q2 Responses	SET	Much better now than one year ago Somewhat better now than one year ago About the same as one year ago Somewhat worse now than one year ago Much worse now than one year ago	Same as SF-36v2	Much better now than one year ago Somewhat better now than one year ago About the same Somewhat worse now than one year ago Much worse now than one year ago	Same as Developmental (Pre-Publication) Version

Table 13.4 (continued)

Comparison of Items in the SF-36v2 Health Survey, SF-36 Health Survey, Developmental Version of the SF-36 Health Survey, and Original MOS PAQ

Item Number	Scale	SF-36v2	SF-36	Developmental (Pre-Publication) Version	Original MOS PAQ Items
Question 3	PF	The following questions 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?	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?	The following items are about activities you might do during a typical day. Does <u>your health</u> now limit you in these activities? If so, how much?	The following items are activities you might do during a typical day. Does <u>your health limit you</u> in these activities? <i>Page 375, Section 2: Physical Health, Q1</i>
3a	PF	<u>Vigorous activities</u> , such as running, lifting heavy objects, participating in strenuous sports	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 375, Section 2, Q1a</i>
3b	PF	<u>Moderate activities</u> , such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 375, Section 2, Q1b</i>
3c	PF	Lifting or carrying groceries	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 375, Section 2, Q1c</i>
3d	PF	Climbing <u>several</u> flights of stairs	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 375, Section 2, Q1d</i>
3e	PF	Climbing <u>one</u> flight of stairs	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 375, Section 2, Q1e</i>
3f	PF	Bending, kneeling, or stooping	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 375, Section 2, Q1f</i>
3g	PF	Walking <u>more than a mile</u>	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 375, Section 2, Q1g</i>
3h	PF	Walking <u>several hundred yards</u>	Walking <u>several blocks</u>	Same as SF-36 ^b	Same as SF-36 <i>Page 375, Section 2, Q1h</i>
3i	PF	Walking <u>one hundred yards</u>	Walking <u>one block</u>	Same as SF-36 ^b	Same as SF-36 ^b <i>Page 375, Section 2, Q1i</i>
3j	PF	Bathing or dressing yourself	Same as SF-36v2	Bathing and dressing yourself	Same as SF-36v2 <i>Page 375, Section 2, Q1j</i>

(continued on next page)

Table 13.4 (continued)

Comparison of Items in the SF-36v2 Health Survey, SF-36 Health Survey, Developmental Version of SF-36 Health Survey, and Original MOS PAQ

Item Number	Scale	SF-36v2	SF-36	Developmental (Pre-Publication) Version	Original MOS PAQ Items
Q3 Responses	PF	Yes, limited a lot Yes, limited a little No, not limited at all	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2
Question 4	RP	During the <u>past 4 weeks</u> , how much of the time 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>	During the <u>past 4 weeks</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>	During the <u>past 4 weeks</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> (Please answer <u>YES</u> or <u>NO</u> for each question by circling 1 or 2 on each line).	During the <u>past 4 weeks</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> (Please answer YES or NO for each question). <i>Page 380, Section 4: Daily Activities, Q1</i>
4a	RP	Cut down on the <u>amount of time</u> you spent on work or other activities	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Cut down the <u>amount of time</u> you spent on work or other activities. <i>Page 380, Section 4, Q1b</i>
4b	RP	<u>Accomplished less</u> than you would like	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 380, Section 4, Q1c</i>
4c	RP	Were limited in the <u>kind</u> of work or other activities	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 380, Section 4, Q1e</i>
4d	RP	Had <u>difficulty</u> performing the work or other activities (for example, it took extra effort)	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 380, Section 4, Q1f</i>
Q4 Responses	RP	All of the time Most of the time Some of the time A little of the time None of the time	Yes No	Same as SF-36	Same as SF-36
Question 5	RE	During the <u>past 4 weeks</u> , how much of the time have you had any of the following problems with your work or other regular daily activities <u>as a result of any emotional problems</u> (such as feeling depressed or anxious)?	During the <u>past 4 weeks</u> , have you had any of the following problems with your work or other regular daily activities <u>as a result of any emotional problems</u> (such as feeling depressed or anxious)?	During the <u>past 4 weeks</u> , have you had any of the following problems with your work or other regular daily activities <u>as a result of any emotional problems</u> (such as feeling depressed or anxious)? (Please answer <u>YES</u> or <u>NO</u> for each question by circling 1 or 2 on each line).	During the <u>past 4 weeks</u> , have you had any of the following problems with your work or other regular daily activities <u>as a result of any emotional problems</u> (such as feeling depressed or anxious)? (Please answer YES or NO for each question). <i>Page 380, Section 4: Daily Activities, Q2</i>

Table 13.4 (continued)

Comparison of Items in the SF-36v2 Health Survey, SF-36 Health Survey, Developmental Version of the SF-36 Health Survey, and Original MOS PAQ

Item Number	Scale	SF-36v2	SF-36	Developmental (Pre-Publication) Version	Original MOS PAQ Items
5a	RE	Cut down on the <u>amount of time</u> you spent on work or other activities	Same as SF-36v2 ^b	Cut down the <u>amount of time</u> you spent on work or other activities	Same as Developmental (Pre-Publication) Version ^b <i>Page 380, Section 4, Q2a</i>
5b	RE	<u>Accomplished less</u> than you would like	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 380, Section 4, Q2b</i>
5c	RE	Did work or other activities <u>less carefully</u> than usual	Didn't do work or other activities as <u>carefully</u> as usual	Same as SF-36 ^b	Same as SF-36 ^b <i>Page 380, Section 4, Q2c</i>
Q5 Responses	RE	All of the time Most of the time Some of the time A little of the time None of the time	Yes No	Same as SF-36	Same as SF-36
Question 6	SF	During the <u>past 4 weeks</u> , to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Same as SF-36v2 ^b <i>Page 375, Section 1: Health and Daily Activities, Q7</i>
Q6 Responses	SF	Not at all Slightly Moderately Quite a bit Extremely	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2
Question 7	BP	How much <u>bodily pain</u> have you had during the <u>past 4 weeks</u> ?	Same as SF-36v2 ^b	Same as SF-36v2 ^b	How much <u>bodily pain</u> have you generally had during the <u>past 4 weeks</u> ? <i>Page 374, Section 1: Health and Daily Activities, Q4</i>
Q7 Responses	BP	None Very mild Mild Moderate Severe Very severe	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2

(continued on next page)

Table 13.4 (continued)

Comparison of Items in the SF-36v2 Health Survey, SF-36 Health Survey, Developmental Version of the SF-36 Health Survey, and Original MOS PAQ

Item Number	Scale	SF-36v2	SF-36	Developmental (Pre-Publication) Version	Original MOS PAQ Items
Question 8	BP	During the <u>past 4 weeks</u> , how much did <u>pain</u> interfere with your normal work (including both work outside the home and housework)?	Same as SF-36v2 ^b	Same as SF-36v2 ^b	Did you experience any bodily pain in the <u>past 4 weeks</u> ? (Yes/No) ^d <i>Page 378, Section 3: Pain, Q1</i> <i>IF YES TO Q1:</i> During the <u>past 4 weeks</u> , how much did pain interfere with the following things? <i>Page 379, Section 3, Q4</i> Your normal work (including both work outside the home and housework) <i>Page 379, Section 3, Q4d</i>
Q8 Responses	BP	Not at all A little bit Moderately Quite a bit Extremely	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2
Question 9	VT and MH	These questions are about how you feel and how things have been with you <u>during the past 4 weeks</u> . 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 <u>past 4 weeks</u> ...	Same as SF-36v2 ^b	These questions are about how you feel and how things have been with you <u>during the past month</u> . For each question, please indicate the one answer that comes closest to the way you have been feeling. How much of the time during the <u>past month</u> ...	
Question 9	MH only				These questions are about how you feel and how things have been with you during the past month. For each question, please circle a number for the one answer that comes closest to the way you have been feeling. <i>Page 381, Section 5: Your Feelings</i>
Question 9	VT only				How often during the <u>past 4 weeks</u> ... ^e <i>Page 377, Section 2: Physical Health, Q11</i>

Table 13.4 (continued)

Comparison of Items in the SF-36v2 Health Survey, SF-36 Health Survey, Developmental Version of the SF-36 Health Survey, and Original MOS PAQ

Item Number	Scale	SF-36v2	SF-36	Developmental (Pre-Publication) Version	Original MOS PAQ Items
9a	VT	Did you feel full of life?	Did you feel full of pep?	Same as SF-36	Same as SF-36 <i>Page 377, Section 2, Q11e</i>
9b	MH	Have you been very nervous?	Have you been a very nervous person?	Same as SF-36	How much of the time, during the <u>past month</u> , have you been a very nervous person? <i>Page 384, Section 5, Q9</i>
9c	MH	Have you felt so down in the dumps that nothing could cheer you up?	Same as SF-36v2	Same as SF-36v2	How much of the time, during the <u>past month</u> , have you felt so down in the dumps that nothing could cheer you up? <i>Page 390, Section 5, Q26</i>
9d	MH	Have you felt calm and peaceful?	Same as SF-36v2	Same as SF-36v2	How much of the time, during the <u>past month</u> , have you felt calm and peaceful? <i>Page 386, Section 5, Q16</i>
9e	VT	Did you have a lot of energy?	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 377, Section 2, Q11c</i>
9f	MH	Have you felt downhearted and depressed?	Have you felt downhearted and blue?	Same as SF-36	How much of the time, during the <u>past month</u> , have you felt downhearted and blue? <i>Page 387, Section 5, Q18</i>
9g	VT	Did you feel worn out?	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 377, Section 2, Q11a</i>
9h	MH	Have you been happy?	Have you been a happy person?	Same as SF-36	During the <u>past month</u> , how much of the time have you been a happy person? <i>Page 392, Section 5, Q33</i>
9i	VT	Did you feel tired?	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 378, Section 2, Q11i</i>
Q9 Responses	VT, MH	All of the time Most of the time Some of the time A little of the time None of the time	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	Same as SF-36	Same as SF-36

(continued on next page)

Table 13.4 (continued)

Comparison of Items in the SF-36v2 Health Survey, SF-36 Health Survey, Developmental Version of the SF-36 Health Survey, and Original MOS PAQ

Item Number	Scale	SF-36v2	SF-36	Developmental (Pre-Publication) Version	Original MOS PAQ Items
Question 10	SF	During the <u>past 4 weeks</u> , how much of the time has your <u>physical health</u> or <u>emotional problems</u> interfered with your social activities (like visiting with friends, relatives, etc.)?	Same as SF-36v2 ^b	Has your <u>health</u> <u>limited your social activities</u> (like visiting with friends or close relatives)? ^f	During the <u>past 4 weeks</u> , how much of the time has your <u>physical health</u> or <u>emotional problems</u> interfered with your social activities (like visiting with friends, relatives, etc.)? <i>Page 394, Section 6: Social Activities, Q1</i>
Q10 Responses	SF	All of the time Most of the time Some of the time A little of the time None of the time	Same as SF-36v2	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	Same as SF-36v2
Question 11	GH	How TRUE or FALSE is <u>each</u> of the following statements for you?	Same as SF-36v2 ^b	Please choose the answer that best describes how <u>true</u> or <u>false</u> each of the following statements is for you.	Same as SF-36v2 ^b <i>Page 397, Section 8: Your Health</i>
11a	GH	I seem to get sick a little easier than other people	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 398, Section 8, Q21</i>
11b	GH	I am as healthy as anybody I know	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 398, Section 8, Q27</i>
11c	GH	I expect my health to get worse	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 398, Section 8, Q24</i>
11d	GH	My health is excellent	Same as SF-36v2	Same as SF-36v2	Same as SF-36v2 <i>Page 399, Section 8, Q33</i>
Q11 Responses	GH	Definitely true Mostly true Don't know Mostly false Definitely false	Same as SF-36v2	Definitely true Mostly true Not sure ^g Mostly false Definitely false	Same as SF-36v2

* Page numbers and item numbers refer to the placement of the original item in the baseline MOS Patient Assessment Questionnaire (PAQ), sections of which are reproduced as shown in the Appendix of Stewart & Ware (1992).

^a Scannable forms have read "Answer every question by marking the appropriate oval."

^b Underscored words may be emphasized differently in earlier versions. The emphasis may be boldfaced, underlined, or both. These are considered equivalent for the purpose of this comparison.

^c This item was first fielded in the PAQ12 (12-month follow-up survey) rather than the PAQ00.

^d A positive response to this question was necessary for administration of the next two questions (Q4 and Q4d on page 379 of the PAQ).

^e In the PAQ VT questions 9a, 9e, 9g, and 9i were administered separately in one grid, with health distress and other vitality items, MH questions 9b, 9c, 9d, 9f, and 9h were administered as separate items without a grid, within a section of 29 general mental health questions, 6 cognitive functioning questions, and 5 emotional ties questions. The other three forms include both the MH and VT items in one grid.

^f This item was included as part of a grid in question 9 of the developmental version, along with MH and VT items. It is a single item in the SF-36.

^g Empirical studies have shown that the numerical scale values for *don't know* and *not sure* do not differ significantly. Therefore, the categories are considered interchangeable.

Table 13.5*Summary of Item Wording Changes From the SF-36 to the SF-36v2*

Item Number	SF-36 Wording	SF-36v2 Wording
3, introduction	Items	Questions
3h	Several blocks	Several hundred yards
3i	One block	100 hundred yards
4, introduction	---	How much of the time
4, response choices	<i>Yes/no</i>	<i>All of the time, most of the time, some of the time, a little of the time, none of the time</i>
5, introduction	---	How much of the time
5, response choices	<i>Yes/no</i>	<i>All of the time, most of the time, some of the time, a little of the time, none of the time</i>
5c	Didn't do work or other activities as carefully as usual	Did work or other activities less carefully than usual
9, response choices	Six choices, including <i>a good bit of the time</i>	Five choices, <i>a good bit of the time</i> dropped
9a	Pep	Life
9b	A very nervous person	Very nervous
9f	Blue	Depressed
9h	A happy person	Happy

the double negative problem inherent in Item 5c of the SF-36 RE scale. After studying a variety of different categorical rating scales (e.g., categories of frequency, severity, endorsement), the development team decided to adopt five well-tested “frequency” categories (ranging from *all of the time* to *none of the time*) that had proven their usefulness and efficiency in other SF-36 scales. These five categories are familiar to SF-36v2 users and accomplish the objectives of ease of administration, increased score reliability, and coverage of a much wider range of participation levels.

There is considerable empirical evidence from independent investigators showing that the SF-36v2 five-level response choices substantially improved the two role-functioning scales (e.g., Jenkinson, Stewart-Brown, Petersen, & Paice, 1999; Linder & Singer, 2003; Perry et al., 2003; Ricci et al., 2004; Taft, Karlsson, & Sullivan, 2000; Wang, Taylor, Pearl, & Chang, 2004). These response choices extend the range measured and greatly increase score precision without increasing respondent burden. Specifically, the SF-36v2: (a) achieves a fourfold increase in the number of levels defined by both RP and RE and more than a fivefold increase in the range measured; (b) substantially reduces the standard deviation; (c) substantially reduces the percentage of respondents who score at the ceiling and floor for both role scales; (d) increases reliability; and (e) improves construct validity through increases in correlations with the physical component.

Meanwhile, the elimination of one of the six response choices (*a good bit of the time*) from the MH and

VT items was based on results from studies using the Thurstone method of equal-appearing intervals (Thurstone & Chave, 1929). Specifically, in studies of SF-36 translations, this response category was not consistently ordered between the *most of the time* and *some of the time* response categories, as was hypothesized (Keller et al., 1998). Eliminating this response category simplified the format of the survey form with little or no loss of information. Subsequent studies using item response theory (IRT) supported this decision.

A comparison of the number of health domains measured, the number of items measuring each domain, and the associated scale levels across the SF-36 and SF-36v2 is presented in Table 13.6.

Scoring

The SF-36v2 has many features that, in addition to representing improvements over the original SF-36, facilitate or improve the user's ability to make meaningful interpretations of results. Most important among these are the updated norms and the development of *T* scores for the health domain scales.

The 1998 norms were selected to introduce the use of the *T*-score metric for the eight health domain scales because these norms were more up-to-date and reflected a greater cross-section of the general population as compared to the 1990 normative data set. Significantly, the 1998 norms and *T*-score algorithms provided the long-awaited link necessary to compare results across studies utilizing the eight health domain scales and/or two summary measures from *any* of the adult Short Form

Table 13.6

Comparison of Number of Health Domain Items and Scale Levels for the SF-36v2 and SF-36

Health Domain	SF-36v2		SF-36	
	# Items	# Levels	# Items	# Levels
Physical Functioning	10	21	10	21
Role-Physical	4	17	4	5
Bodily Pain	2	11	2	11
General Health	5	21	5	21
Vitality	4	17	4	21
Social Functioning	2	9	2	9
Role-Emotional	3	13	3	4
Mental Health	5	21	2	26
Self-Evaluated Transition	1	5	1	5

Adapted from “The MOS 36-Item Short Form Health Survey (SF-36). I. Conceptual Framework and Item Selection” by J. E. Ware, Jr., and C. D. Sherbourne, 1992, *Medical Care*, 30, 473–483. Copyright 1992 by Lippincott-Raven Publishers.

surveys. As described in Chapter 5 of this manual, SF-36v2 scoring algorithms allow for the computation of *T* scores for all eight health domain scales, utilizing the same standardization of scoring (mean = 50, *SD* = 10) that had made the SF-36 component summary measures easier to interpret.

Advantages of *T* Scores

The interpretation of SF-36v2 results has been greatly simplified with the availability of the *T*-score metric for scoring the health domain scales and component summary measures, and it is recommended that users base their interpretations on these *T* scores. The advantage of *T* scores can be illustrated by comparing the SF-36 profile scored using the original 0–100 metric with the profile based on *T*-score algorithms for the same sample. For the purposes of this comparison, surveys were scored both ways for a sample of asthmatic patients who participated in a clinical trial (Okamoto, Noonan, Boisblanc, & Kellerman, 1996).

The original 0–100 scoring metric produced the profile shown in Figure 13.1. The shape of this profile—the peaks and valleys due to higher and lower scores, respectively, across scales—reflect both the impact of asthma on health domains, as well as arbitrary differences in the ceilings and floors of the scales. Three scales—namely GH, VT, and MH—measure relatively wide score ranges and set the ceiling relatively high by measuring very favorable levels of those health domains (Ware, Snow, Kosinski, & Gandek, 1993). Other scales, such as PF and RP, assess a narrower range based on a lower ceiling. For these scales, the most favorable levels (a score of 100 using the original SF-36 algorithms) represent the absence of limitations and do not extend the range

into well-being. Thus, when using the original 0–100 metric, the average score for each scale substantially differs across the profile for reasons that have nothing to do with asthma (see *Norm* in Figure 13.1). Ignoring these norms, a reasonable inference from the profile in Figure 13.1 is that asthma has a greater impact on the Vitality (VT) scale than on the Physical Functioning (PF) scale; however, this inference is incorrect.

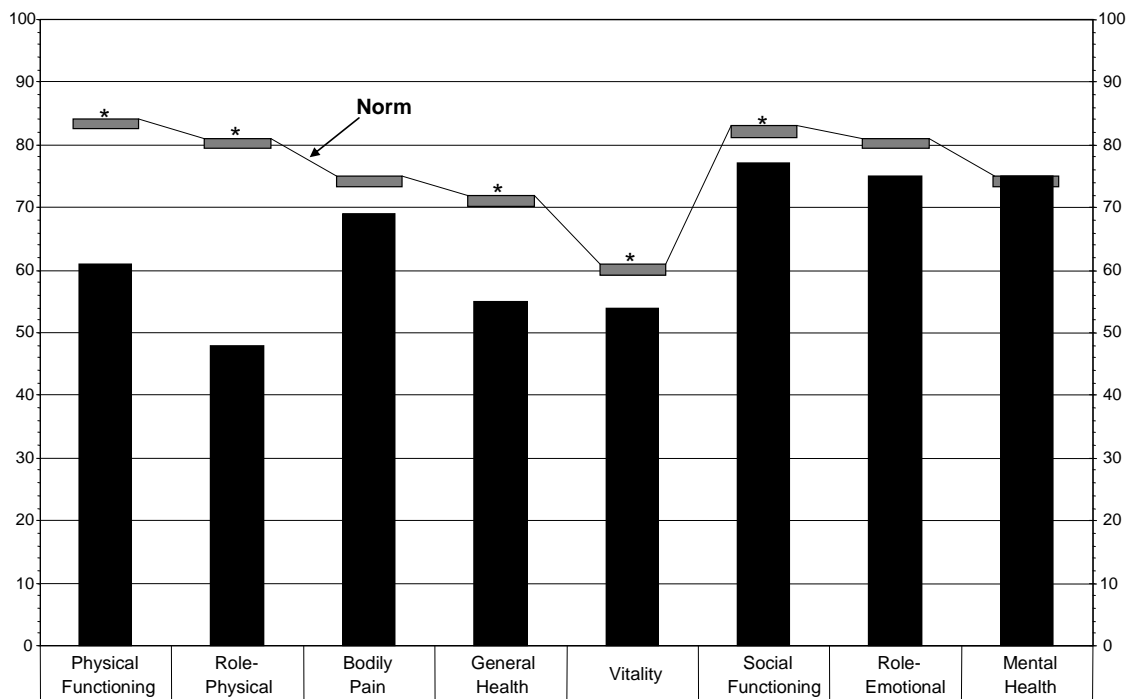
General population norms provide a basis for meaningful comparisons across scales. For example, the PF scale general population norm is between 80 and 90 while the VT norm is around 60 on the 100-point scale (see Figure 13.1). In relation to these norms, the impact of asthma is actually much larger on the PF scale than on the VT scale, although both are statistically significant. Using the original 0–100 scoring, these differences in norms must be kept in mind when interpreting a profile. Differences in standard deviations (which substantially vary across some of the scales) must also be considered when comparing 0–100 scoring results across scales.

With *T* scores, each scale has the same average (50) and the same standard deviation (10), meaning each point equals one-tenth of a standard deviation. As a result, without referring to tables of norms, this method makes it clear that whenever an individual respondent’s scale score is below 45, or a group mean scale score is below 47, the implication is that health status is below the average range (see Chapter 7). As shown in Figure 13.2, *T*-score differences in scale scores much more clearly reflect the impact of the disease—in this example, the impact of asthma. Using *T* scores, clinicians can more quickly and appropriately interpret the effect of asthma, or any other health issue, on an SF-36v2 profile.

Other advantages of *T*-scores are evident when examining Figures 13.2 and 13.3. First, results for the PCS and MCS measures, which have always been transformed to *T* scores, can be compared directly with results for the eight health domain scales when all are standardized on a common metric in relation to population norms. Because the PCS and MCS measures take into account the correlations amongst the eight health domain scales, it is clear from the example in Figure 13.2 that asthma has a very broad impact on the physical component of health.

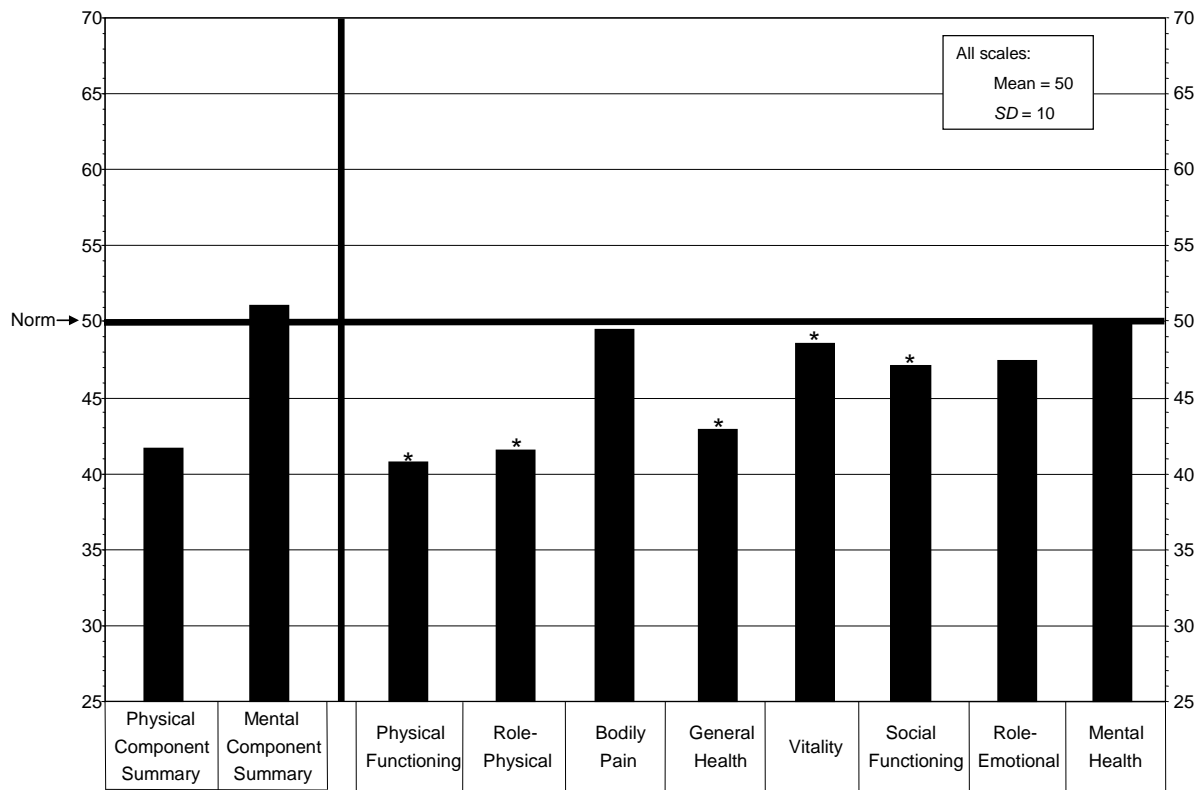
Second, assessing treatment effects in a clinical trial is made easier through the use of *T* scores, as is illustrated in Figure 13.3. Relative to baseline after 16 weeks of treatment, patients treated using an inhaler showed statistically significant improvements (represented by the shaded portions of the bars in Figure 13.3) on the PCS measure and on the PF, RP, and GH scales (i.e., three of the four scales most closely associated with physical functioning).

Figure 13.1 SF-36 Profile of 0–100 Scores: Adults With Asthma Compared With U.S. General Population Norms

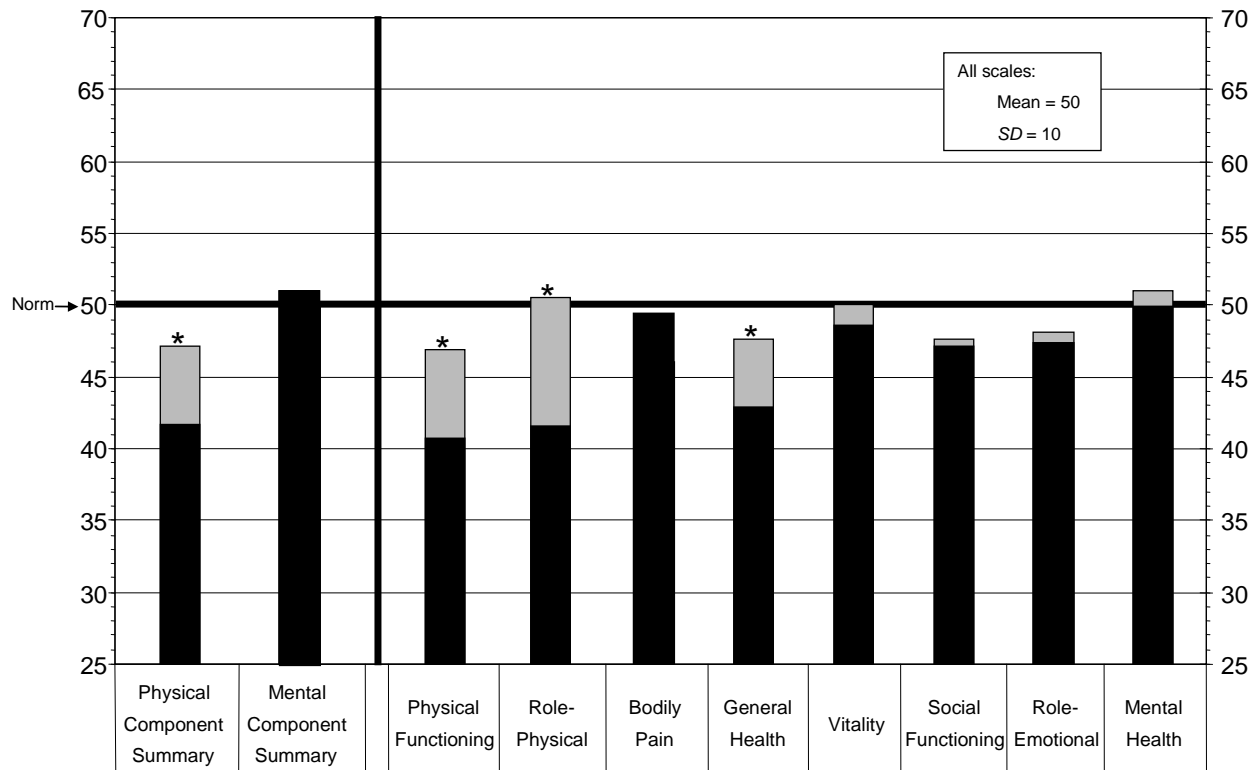


Note. Adapted from “Fluticasone Propionate Improves Quality of Life in Patients With Asthma Requiring Oral Corticosteroids” by L. J. Okamoto, M. Noonan, B. P. Boisblanc, and D. J. Kellerman, 1996, *Annals of Allergy, Asthma and Immunology*, 76, 455–461. Copyright 1996 by Elsevier.
 * Norm significantly higher.

Figure 13.2 SF-36 Profile of *T* Scores: Adults With Asthma Compared With U.S. General Population Norms



Note. Adapted from “Fluticasone Propionate Improves Quality of Life in Patients With Asthma Requiring Oral Corticosteroids” by L. J. Okamoto, M. Noonan, B. P. Boisblanc, and D. J. Kellerman, 1996, *Annals of Allergy, Asthma and Immunology*, 76, 455–461. Copyright 1996 by Elsevier.
 * Norm significantly higher

Figure 13.3 Interpreting SF-36 Treatment Outcomes Among Adults With Asthma

Note. Adapted from “Fluticasone Propionate Improves Quality of Life in Patients With Asthma Requiring Oral Corticosteroids” by L. J. Okamoto, M. Noonan, B. P. Boisblanc, and D. J. Kellerman, 1996, *Annals of Allergy, Asthma and Immunology*, 76, 455–461. Copyright 1996 by Elsevier.

* Significantly improvement with treatment.

To summarize, the main advantage of *T* scores is simplified interpretation. Specifically, when interpreting *T* scores, one no longer has to remember the different norms for the eight health domain scales because the general population norm is built into the scoring algorithm. For example, for all scales and measures, individual respondent scores below 45 and group mean scores below 47 can be interpreted as being below the average range for the general population. And because the standard deviations for all scales and measures are equalized at 10, it is easier to see, in standard deviation units, exactly how far below or above the general population mean a given score is. Moreover, comparisons across SF-36v2 health domain scale and component summary measure scores can be directly made. Note that the use of the *T*-score metric continues with the release of the 2009 SF-36v2 norms (see Chapter 14).

Finally, for those conducting research, it is important to *not* “mix” or combine *T* scores and 0–100 scores from either measure for the purpose of analyzing or reporting data. The mixed scores that have been reported in the published literature have resulted in erroneous conclusions about the hypotheses being tested. It is also important to clearly document the norms and

scoring algorithms used in reports of study methods that accompany results based on the SF-36v2 2009 U.S. general population norms. Further, because tables and figures are sometimes distributed separately, it is also important to include explicit references to *SF-36v2 2009 U.S. general population norms* and to *T* scores in any tables and/or figures presenting results based on the more current normative data (see Chapter 4).

1998 Norms

SF-36v2 norms were first derived by QualityMetric Incorporated from the responses of representative adult samples of the 1998 noninstitutionalized U.S. general population who participated in the 1998 National Survey of Functional Health Status (NSFHS), which included the SF-36v2 survey. Norms for the 4-week standard form ($N = 7,069$) and 1-week acute form ($N = 7,837$) were developed separately. Because health status scores for some constructs significantly differ across age groups, as well as for men and women, norms were developed in different age groups for the total population and separately for males and females. Standard and acute form norms were also developed for 18 disease- or condition-specific populations.

The 1998 SF-36v2 norms have since been replaced by the 2009 norms, collected as part of the QualityMetric 2009 Norming Study. Development of this most current set of norms is discussed in detail in Chapter 14.

Comparability of Results

The use of the *T*-score metric made it possible for the results from the two SF-36 versions to be directly compared, with the mean and standard deviation being 50 and 10, respectively, in the 1998 U.S. general population. This shared metric allowed users to take advantage of the advances achieved with the SF-36v2 while retaining the option of comparing results with SF-36 results. In addition, cross-sectional and longitudinal norms for the general population were estimated for the SF-36 using the *T*-score metric as the basis for scoring the eight health domain scales and two component summary measures, thus making SF-36 scores easier to interpret and directly comparable to SF-36v2 scores. Extensive discussions regarding the psychometric comparability of SF-36v2 health domain scales and component summary measures with those of the SF-36 are presented in the following sections of this chapter.

Psychometric Characteristics and Comparability

Comparison of the SF-36v2 with the SF-36. Prior to the 1998 norming survey, little data were available for the purposes of testing whether changes made to SF-36v2 items would impact the psychometric properties of SF-36 health domain scale scores. To address this topic, a large U.S. general population study was conducted to compare the psychometric properties of the two versions of the survey. Participants were randomized to complete either the SF-36 or SF-36v2, and then further randomized to either the acute (1-week) form or the standard (4-week) form of one version of the survey. It was hypothesized that changes made to the SF-36v2 would not affect whether its scales conform to the assumptions underlying its predecessor's scoring and scaling; that the adoption of five-choice response categories for the two role scales, RP and RE, would substantially reduce ceiling and floor effects, greatly reduce their variances, and increase their item-scale correlations; and that the physical and mental health constructs (principal components) underlying the SF-36 would be replicated in the SF-36v2.

Using the 1998 norms data, the study's results presented in Tables 13.7 through 13.10 show that the revised wording adopted for the SF-36v2 did not change the empirical validity of the items or the assumptions underlying the scoring of the SF-36 scales. For example, a change in the characterization of distance from *blocks*

to *yards* in two PF items had no effect on the relationship between those items and the PF scale, with their correlations (corrected for overlap) being virtually identical across the two versions of the survey. Furthermore, the means (thus the relative difficulty) and standard deviations were nearly the same for these two PF items across the two versions. Specifically, walking "several blocks" and walking "several hundred yards" were both more difficult than walking "one block" and "walking 100 hundred yards," respectively. The remaining wording changes and the reduced number of response choices for VT and MH scale items also did not impact the psychometric properties of either scale. Item-scale correlations remained substantial in magnitude, and item means and standard deviations were approximately equivalent across items within their respective scales.

As hypothesized, an increase in the number of response choices from two to five resulted in higher item-scale correlations for both the RP and RE scales. This is attributed to the increased range covered by these items as a result of the additional response choices. Importantly, the relative difficulty of each item within each health domain scale was unchanged by the increase in response choices. For example, the "accomplish less than you would like" item was the most difficult in the RP scale, as indicated by the lower item mean, across both versions of the survey; meanwhile, the "cut down the amount of time spent on work" item was the easiest RP item (i.e., highest mean score).

A decrease in the number of response choices from six to five for items in both the VT and MH scales did not change item-scale correlations or the order of item difficulty (i.e., the order of item means stayed the same) within each version of these scales. Also, as was previously found by Keller et al. (1997), differences in recall period (4-weeks vs. 1-week) did not affect whether SF-36 scales conformed to the scaling assumptions underlying their construction and scoring; however, the score distributions presented here differed from the Keller et al. study's results. Because these differences in score distributions cannot be confirmed by the 1998 SF-36v2 data, the Keller et al. findings should be viewed with caution.

Tables 13.11 and 13.12 summarize the results of tests of scaling assumptions for the standard and acute forms, respectively, of both versions of the survey, based on 0–100 scoring. Scaling success rates were perfect (100%) for both SF-36 and SF-36v2 standard and acute forms, supporting the grouping of items into the eight scales. No differences were observed in the percentage of completed items or in the percentage of computable scale scores across versions.

Table 13.7

SF-36v2 Standard (4-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (N = 5,038)

Scale	Item	Mean	SD	PF	RP	BP	GH	VT	SF	RE	MH
PF	3a	2.11	0.80	.65 ^a	.61	.52	.52	.37	.39	.35	.19
	3b	2.65	0.62	.81 ^a	.72	.54	.52	.41	.50	.47	.26
	3c	2.73	0.55	.80 ^a	.68	.52	.48	.39	.51	.48	.27
	3d	2.48	0.71	.80 ^a	.64	.51	.52	.40	.45	.42	.24
	3e	2.74	0.55	.83 ^a	.65	.49	.46	.37	.48	.46	.25
	3f	2.52	0.67	.75 ^a	.63	.54	.47	.38	.43	.41	.22
	3g	2.49	0.75	.81 ^a	.65	.51	.51	.40	.44	.43	.22
	3h	2.70	0.61	.84 ^a	.65	.48	.47	.37	.46	.46	.23
	3I	2.78	0.53	.79 ^a	.59	.43	.42	.34	.44	.47	.23
	3j	2.89	0.38	.57 ^a	.45	.32	.30	.26	.37	.40	.23
RP	4a	4.37	1.08	.72	.88 ^a	.61	.56	.47	.61	.61	.31
	4b	4.07	1.20	.68	.87 ^a	.61	.58	.52	.60	.60	.33
	4c	4.24	1.18	.76	.91 ^a	.65	.58	.48	.60	.57	.31
	4d	4.24	1.15	.75	.90 ^a	.67	.61	.51	.62	.60	.34
BP	7	4.37	1.29	.53	.58	.76 ^a	.56	.50	.50	.38	.35
	8	4.22	1.01	.61	.71	.76 ^a	.59	.53	.64	.50	.41
GH	1	3.54	0.93	.57	.57	.55	.70 ^a	.53	.51	.42	.37
	11a	4.25	1.01	.34	.41	.42	.54 ^a	.45	.47	.39	.43
	11b	3.74	1.16	.47	.50	.46	.69 ^a	.51	.49	.39	.40
	11c	3.62	1.12	.35	.35	.37	.48 ^a	.41	.33	.29	.31
	11d	3.57	1.23	.55	.59	.57	.79 ^a	.60	.54	.45	.45
VT	9a	3.50	0.95	.40	.46	.46	.56	.67 ^a	.57	.46	.60
	9e	3.21	1.03	.45	.50	.50	.61	.71 ^a	.56	.45	.56
	9g	3.50	1.01	.33	.40	.43	.48	.69 ^a	.49	.40	.55
	9i	3.21	0.96	.35	.40	.42	.50	.72 ^a	.49	.39	.52
SF	6	4.36	1.00	.53	.65	.58	.57	.59	.78 ^a	.66	.58
	10	4.34	1.01	.50	.58	.55	.56	.60	.78 ^a	.60	.59
RE	5a	4.50	0.94	.51	.61	.43	.47	.47	.64	.88 ^a	.53
	5b	4.32	1.06	.49	.59	.43	.47	.51	.63	.88 ^a	.55
	5c	4.50	0.91	.50	.58	.43	.46	.46	.60	.83 ^a	.52
MH	9b	4.16	0.96	.23	.26	.30	.37	.44	.45	.43	.62 ^a
	9c	4.45	0.88	.25	.30	.33	.40	.50	.55	.51	.72 ^a
	9d	3.48	0.97	.21	.26	.33	.43	.60	.47	.40	.63 ^a
	9f	4.20	0.96	.24	.29	.32	.40	.55	.54	.52	.74 ^a
	9h	3.76	0.86	.20	.25	.30	.42	.55	.48	.40	.64 ^a

^aItem-scale correlation corrected for overlap (relevant item removed from its scale for correlation) and hypothesized to be highest in same row.

With the exception of the RP and RE scales, ceiling and floor effects were similar across the SF-36 and SF-36v2 standard and acute forms. For the SF-36v2's RP and RE scales, the increased number of response choices significantly decreased the observed ceiling and floor effects as compared to the SF-36 role scales. For example, the ceiling and floor effects of the RP scale were reduced from 61.9% to 47.4% and 13.6% to 2.1%, respectively, on the standard form.

Scale means and standard deviations across the SF-36 and SF-36v2 standard and acute forms were similar (e.g., within 1 to 2 points of each other) for all scales except RP and RE, whose mean scale scores were 4 to 5 points higher on the SF-36v2. Consistent with

hypotheses, the standard deviations for SF-36v2 RP and RE scales were substantially smaller than the standard deviations for the SF-36 RP and RE scales.

As shown in Tables 13.11 and 13.12, internal consistency reliability coefficients were above the recommended level for group comparisons (.70) and did not differ between the SF-36 and SF-36v2 standard and acute forms for six of the eight health domain scales. As expected, internal consistency reliability estimates were substantially higher among the SF-36v2 RP and RE scales (.95 and .93, respectively) as compared to the SF-36 (.88 and .82, respectively), reflecting the item improvements that were incorporated into these revised scales.

Table 13.8

SF-36 Standard (4-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (N = 2,031)

Scale	Item	Mean	SD	PF	RP	BP	GH	VT	SF	RE	MH
PF	3a	2.06	0.80	.62 ^a	.53	.50	.49	.38	.37	.27	.20
	3b	2.63	0.63	.81 ^a	.64	.56	.53	.43	.54	.36	.26
	3c	2.73	0.55	.79 ^a	.58	.51	.47	.38	.52	.33	.23
	3d	2.47	0.72	.82 ^a	.58	.51	.53	.45	.47	.32	.24
	3e	2.73	0.58	.83 ^a	.55	.48	.46	.39	.47	.31	.21
	3f	2.50	0.69	.78 ^a	.56	.54	.46	.38	.45	.30	.21
	3g	2.47	0.75	.81 ^a	.57	.50	.51	.42	.45	.32	.23
	3h	2.67	0.65	.85 ^a	.58	.50	.49	.40	.49	.34	.23
	3i	2.81	0.50	.77 ^a	.49	.43	.42	.33	.46	.31	.21
	3j	2.90	0.37	.55 ^a	.34	.31	.30	.22	.38	.24	.17
RP	4a	1.83	0.38	.55	.71 ^a	.55	.45	.43	.59	.42	.28
	4b	1.67	0.47	.55	.72 ^a	.55	.48	.48	.56	.47	.31
	4c	1.75	0.43	.64	.79 ^a	.61	.47	.46	.56	.39	.27
	4d	1.75	0.43	.61	.79 ^a	.61	.50	.49	.58	.41	.30
BP	7	4.31	1.26	.53	.57	.76 ^a	.55	.52	.51	.33	.34
	8	4.19	1.02	.62	.71	.76 ^a	.56	.54	.67	.46	.39
GH	1	3.52	0.95	.59	.52	.54	.72 ^a	.57	.52	.36	.40
	11a	4.22	1.02	.33	.37	.42	.54 ^a	.46	.47	.37	.47
	11b	3.68	1.17	.47	.44	.47	.70 ^a	.53	.47	.33	.38
	11c	3.66	1.16	.36	.34	.36	.52 ^a	.39	.33	.25	.31
	11d	3.49	1.24	.55	.50	.54	.79 ^a	.59	.51	.36	.44
VT	9a	3.60	1.27	.47	.51	.53	.60	.76 ^a	.55	.39	.53
	9e	3.59	1.30	.46	.50	.50	.59	.75 ^a	.52	.38	.55
	9g	4.28	1.23	.35	.41	.44	.52	.71 ^a	.51	.37	.54
	9i	3.96	1.21	.35	.43	.45	.50	.75 ^a	.51	.37	.52
SF	6	4.33	1.01	.56	.65	.60	.54	.56	.76 ^a	.59	.51
	10	4.30	1.02	.49	.58	.57	.55	.59	.76 ^a	.53	.58
RE	5a	1.86	0.35	.34	.44	.36	.38	.38	.54	.70 ^a	.45
	5b	1.76	0.43	.31	.43	.36	.36	.41	.52	.70 ^a	.50
	5c	1.87	0.34	.34	.40	.35	.35	.35	.50	.68 ^a	.42
MH	9b	5.06	1.11	.16	.22	.25	.36	.36	.39	.38	.59 ^a
	9c	5.36	1.04	.26	.28	.31	.40	.50	.55	.48	.73 ^a
	9d	3.97	1.25	.19	.27	.34	.42	.58	.43	.39	.66 ^a
	9f	5.09	1.06	.23	.28	.30	.40	.50	.51	.49	.71 ^a
	9h	4.38	1.14	.24	.29	.34	.43	.54	.44	.38	.66 ^a

^aItem-scale correlation corrected for overlap (relevant item removed from its scale for correlation) and hypothesized to be highest in same row.

Overall, the results of this investigation indicate that the SF-36v2 is a comparable yet improved version of the SF-36. Changes to wording and to the number of response choices for SF-36v2 items resulted in substantial improvements, particularly in the RP and RE scales. Specifically, both ceiling and floor effects for these scales were substantially reduced. Furthermore, the standard deviations were reduced and the internal consistency reliabilities were improved for both of these SF-36v2 scales.

Comparison of the standard (4-week recall) and acute (1-week recall) forms. Tables 13.7 and 13.9 present item means, standard deviations, and correlations between items and health domain scales in the 1998 U.S.

general population for the SF-36v2 standard (4-week recall) and acute (1-week recall) forms, respectively. (Note that similar comparisons based on 2009 U.S. general population data are presented in Chapter 14.) The item-scale correlations that have been corrected for overlap (i.e., each item's score was removed from its parent scale's score before the correlation was calculated) and that are hypothesized to be the highest in the same row are noted. Examination of Tables 13.7 and 13.9 reveals that within each scale, correlations between items and their hypothesized scale were roughly equal and exceeded the .40 standard for internal consistency (Helmstader, 1964) for both the standard (4-week) and acute (1-week) forms. Also, for all scales except

Table 13.9

SF-36v2 Acute (1-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (N = 6,137)

Scale	Item	Mean	SD	PF	RP	BP	GH	VT	SF	RE	MH
PF	3a	2.09	0.81	.64 ^a	.60	.52	.53	.38	.39	.35	.21
	3b	2.63	0.64	.82 ^a	.72	.58	.53	.42	.52	.45	.27
	3c	2.71	0.58	.79 ^a	.69	.54	.49	.38	.52	.45	.28
	3d	2.48	0.73	.81 ^a	.65	.53	.53	.43	.47	.41	.27
	3e	2.72	0.58	.83 ^a	.66	.51	.48	.39	.49	.44	.27
	3f	2.51	0.69	.77 ^a	.63	.56	.48	.38	.44	.40	.25
	3g	2.47	0.76	.82 ^a	.67	.53	.52	.42	.46	.42	.26
	3h	2.68	0.63	.84 ^a	.67	.51	.48	.39	.49	.43	.26
	3i	2.76	0.55	.79 ^a	.64	.48	.44	.35	.47	.43	.25
	3j	2.90	0.37	.55 ^a	.45	.32	.30	.25	.38	.32	.21
	RP	4a	4.36	1.11	.73	.88 ^a	.62	.55	.49	.64	.61
4b		4.06	1.25	.70	.87 ^a	.63	.57	.54	.62	.61	.38
4c		4.20	1.22	.77	.90 ^a	.66	.58	.50	.62	.58	.34
4d		4.23	1.18	.76	.91 ^a	.67	.59	.53	.65	.61	.36
BP	7	4.49	1.30	.56	.59	.78 ^a	.55	.51	.50	.40	.36
	8	4.27	1.01	.65	.72	.78 ^a	.57	.54	.65	.52	.42
GH	1	3.53	0.94	.58	.56	.54	.72 ^a	.54	.49	.41	.39
	11a	4.21	1.04	.36	.41	.41	.56 ^a	.46	.46	.38	.42
	11b	3.75	1.16	.48	.49	.46	.70 ^a	.51	.47	.38	.40
	11c	3.61	1.15	.36	.35	.36	.51 ^a	.40	.32	.27	.30
	11d	3.55	1.23	.56	.57	.55	.78 ^a	.59	.54	.44	.46
VT	9a	3.47	1.00	.43	.49	.49	.59	.69 ^a	.58	.49	.62
	9e	3.19	1.05	.47	.53	.51	.60	.74 ^a	.55	.48	.58
	9g	3.51	1.04	.34	.42	.44	.47	.71 ^a	.50	.40	.56
	9i	3.19	0.98	.35	.42	.44	.50	.75 ^a	.50	.41	.55
SF	6	4.38	1.02	.54	.65	.58	.56	.58	.77 ^a	.65	.59
	10	4.35	1.04	.52	.62	.55	.54	.60	.77 ^a	.63	.61
RE	5a	4.50	0.96	.48	.62	.46	.46	.49	.66	.88 ^a	.56
	5b	4.33	1.08	.47	.61	.45	.45	.52	.65	.88 ^a	.58
	5c	4.52	0.93	.47	.59	.45	.44	.47	.62	.84 ^a	.53
MH	9b	4.22	0.95	.25	.30	.31	.37	.45	.46	.45	.62 ^a
	9c	4.49	0.88	.26	.33	.32	.38	.51	.56	.53	.71 ^a
	9d	3.49	1.00	.22	.29	.34	.42	.61	.48	.42	.65 ^a
	9f	4.26	0.95	.26	.32	.33	.41	.57	.56	.54	.74 ^a
	9h	3.75	0.88	.25	.29	.34	.43	.59	.49	.43	.66 ^a

^aItem-scale correlation corrected for overlap (relevant item removed from its scale for correlation) and hypothesized to be highest in same row.

PF, item means and standard deviations were roughly equal across items within each scale, for both forms. The implication of these results is that items in each hypothesized scale contained approximately the same proportion of information about the health domain being measured. These results support the comparability of the scales from the standard and acute forms, as well as the use of the summated ratings method as the first step in scoring SF-36v2 scales (see Chapter 5). For the PF scale, items measuring easier physical tasks, such as bathing and dressing, had lower standard deviations than items measuring more demanding tasks, such as vigorous activities or climbing several flights of stairs. However, additional analyses by IRT models have shown that these

differences are due to larger floor effects for the easy items in the samples analyzed, not because these items contain more information than the difficult items (e.g., Haley, McHorney, & Ware, 1994). Thus, the summated ratings method is appropriate for the PF scale as well.

Physical and Mental Component Summary Measures

The Physical Component Summary (PCS) and Mental Component Summary (MCS) measures are referred to as *component* summary measures (Ware, Kosinski, Bayliss, et al., 1995) because they were derived and

Table 13.10

SF-36 Acute (1-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 1998 U.S. General Population Sample (N = 1,700)

Scale	Item	Mean	SD	PF	RP	BP	GH	VT	SF	RE	MH
PF	3a	2.08	0.80	.65 ^a	.57	.55	.53	.43	.43	.32	.22
	3b	2.61	0.66	.80 ^a	.63	.58	.50	.42	.51	.35	.25
	3c	2.71	0.57	.79 ^a	.61	.56	.47	.37	.51	.38	.27
	3d	2.46	0.73	.82 ^a	.59	.54	.51	.43	.47	.36	.24
	3e	2.70	0.59	.84 ^a	.58	.53	.48	.38	.47	.35	.21
	3f	2.48	0.70	.77 ^a	.59	.58	.47	.39	.43	.35	.24
	3g	2.45	0.77	.80 ^a	.59	.54	.51	.42	.47	.39	.25
	3h	2.61	0.69	.84 ^a	.59	.54	.49	.40	.49	.39	.26
	3i	2.78	0.53	.78 ^a	.53	.48	.42	.34	.44	.36	.20
	3j	2.88	0.39	.59 ^a	.40	.38	.32	.26	.41	.29	.20
RP	4a	1.83	0.38	.57	.72 ^a	.54	.44	.44	.58	.47	.28
	4b	1.68	0.47	.57	.72 ^a	.56	.51	.52	.56	.49	.32
	4c	1.76	0.43	.67	.77 ^a	.61	.49	.46	.54	.41	.26
	4d	1.75	0.43	.63	.80 ^a	.62	.53	.50	.57	.45	.30
BP	7	4.45	1.30	.57	.58	.77 ^a	.56	.50	.54	.36	.35
	8	4.26	1.01	.67	.71	.77 ^a	.58	.53	.65	.45	.38
GH	1	3.50	0.97	.60	.54	.56	.73 ^a	.56	.51	.39	.38
	11a	4.23	1.04	.33	.40	.41	.54 ^a	.45	.47	.32	.39
	11b	3.72	1.16	.47	.45	.46	.72 ^a	.51	.47	.34	.37
	11c	3.62	1.16	.37	.34	.40	.51 ^a	.40	.34	.27	.34
	11d	3.51	1.22	.55	.54	.57	.79 ^a	.59	.54	.40	.43
VT	9a	3.60	1.31	.46	.52	.52	.60	.75 ^a	.55	.41	.53
	9e	3.57	1.36	.46	.51	.49	.60	.78 ^a	.56	.41	.53
	9g	4.33	1.25	.35	.43	.43	.48	.72 ^a	.53	.42	.51
	9i	3.97	1.22	.37	.45	.45	.52	.76 ^a	.52	.42	.51
SF	6	4.41	0.99	.56	.63	.60	.56	.59	.76 ^a	.60	.53
	10	4.37	1.03	.51	.59	.58	.54	.58	.76 ^a	.56	.56
RE	5a	1.85	0.35	.40	.48	.38	.39	.42	.56	.71 ^a	.49
	5b	1.75	0.43	.36	.45	.36	.38	.44	.55	.70 ^a	.50
	5c	1.85	0.36	.38	.44	.37	.37	.39	.51	.67 ^a	.43
MH	9b	5.13	1.11	.19	.20	.21	.31	.32	.36	.38	.55 ^a
	9c	5.43	0.98	.28	.32	.35	.39	.49	.56	.54	.71 ^a
	9d	3.99	1.28	.23	.28	.34	.43	.55	.45	.38	.64 ^a
	9f	5.16	1.05	.25	.27	.30	.37	.47	.47	.48	.66 ^a
	9h	4.38	1.19	.18	.24	.30	.38	.50	.43	.37	.63 ^a

^aItem-scale correlation corrected for overlap (relevant item removed from its scale for correlation) and hypothesized to be highest in same row.

scored using the factor analytic method of principal *components* analysis (Harman, 1976). Principal components analyses of correlations among the eight health domain scales have consistently identified two components. On the strength of the pattern of their correlations with the eight scales, they have been interpreted as *physical* and *mental* components of health status. These physical and mental components account for 81.5% of the reliable variance in the SF-36 scales in the U.S. general 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 HIE Medical History Questionnaire (Ware, Brook, et al., 1980),

the MOS Functioning and Well-Being Profile (Hays & Stewart, 1990), and the Sickness Impact Profile (SIP; Bergner, Bobbitt, Carter, & Gibson, 1981). With their factor analytic studies yielding recognizably similar physical and mental components, these comprehensive survey results offer further support of the generalizability of this two-dimensional model of health.

The measurement model underlying the construction of the multi-item health domain scales and component summary measures of both versions of the SF-36 is illustrated in Chapter 2 of this manual (see Figure 2.1). Recall that this model has three levels: (a) items, (b) health domain scales that aggregate items, and (c) component summary measures that aggregate the health

domain scales. All but one of the survey's 36 items (the Self-Evaluated Transition [SET] item) are used to score the eight health domain scales. As shown in Figure 2.1, each of these 35 items is used in scoring only one health domain scale. Tests of assumptions underlying the algorithms used in scoring the eight health domain scales have been strongly supported in the United States (McHorney, Ware, Lu, & Sherbourne, 1994; Ware et al., 1993) and in other countries (Ware et al., 1998).

The health domain scales also form two distinct, higher-ordered clusters, which are reflected in the ordering of the eight health domain scales on the Short Form profile. As previously noted, the eight scales are ordered from left to right according to the extent to which they measure physical and mental health. Three scales (PF, RP, and BP) correlate most highly with the physical component and contribute most to the scoring of the PCS measure. The mental component correlates most highly with the MH, RE, and SF scales, which contribute the most to the scoring of the MCS measure. Three of the scales have noteworthy correlations with both components: the VT scale correlates substantially with both; the GH scale correlates with both but higher with the physical component; and the SF scale correlates much higher with the mental component. Reasons for these patterns of correlations are discussed in McHorney et al. (1993).

The psychometric approach to summarizing health measures illustrated here is in contrast to a utility index, such as the SF-6D (Brazier, Roberts, & Deverill, 2002; Brazier, Usherwood, Harper, & Thomas, 1998; see also Chapter 2), in which measures are aggregated without regard to their interrelationships. 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 measures described here is their value in distinguishing a physical health outcome from a mental health outcome.

The development of the SF-36v2 PCS and MCS measures mirror and are built upon the development of the SF-36 PCS and MCS measures; therefore, they will be discussed here in detail, as they served as the bases for the SF-36v2 PCS and MCS measures.

Methodological Issues

Principal component analysis and factor analysis have proven to be very useful in testing hypotheses about the structure of health and in evaluating the construct validity of the Short Form and other health surveys (Derogatis, 1986; Goldberg & Hillier, 1979; Hall, Epstein, & McNeil, 1989; Hays & Stewart, 1990; Mason, Anderson, & Meenan, 1988; McHorney et al., 1993; Schag, Heinrich, Aadland, & Ganz, 1990; Veit & Ware, 1983; Ware, 1976a;

Ware, Davies-Avery, & Brook, 1980; Wiklund, Lindvall, Swedberg, & Zupkis, 1987). Considerable attention was given to the implications of different methods of extraction and rotation. In many cases, conclusions did not vary across methods. When such conclusions do vary, the choice among methods depends on the purpose(s) of the analysis (Nunnally & Bernstein, 1994).

Such choices for the SF-36 studies stemmed from the developers' earlier work and considerations of work published by others (Snyder & Ware, 1974; Ware, Davies-Avery, & Brook, 1980; Ware, Miller, & Snyder, 1973; Ware & Snyder, 1975). Consistent with guidelines suggested by Harris and Harris (1971), choice of method was not of great consequence in arriving at conclusions about the structure of the SF-36 because it 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, 1976a; Ware & Karmos, 1976a, 1976b; Ware et al., 1973), 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. Such advantages are also well documented in empirical studies of personality variables (Comrey, 1973).

The two-component structure of both versions of the SF-36 has also been shown to satisfy criteria for "simple structure" (Nunnally & Bernstein, 1994) across patient and general population samples in the United States and other countries. To facilitate reanalyses by others, the matrices of correlations for the SF-36v2 standard and acute forms are reproduced in Tables 13.13 and 13.14, respectively.

Principal Components

As previously noted, the interpretation of the two components as physical health and mental health has been straightforward and robust across methods. Thus, the choice of analytic method was not governed by considerations for interpreting the components. Rather, the choice of the principal components method was based on other considerations, including the ease of estimation of component scores for the two summary measures, estimation of the content of the eight SF-36 health domain scales in relation to physical and mental health status, the explanatory power of the components, and the components' validity in discriminating between physical and mental health status. Each of these considerations is briefly discussed later in this chapter.

Table 13.11

Comparison of Descriptive Statistics and Reliability Estimates for the Standard (4-Week Recall) Forms of the SF-36v2 and SF-36, 1998 U.S. General Population Sample Using 0-100 Scoring

	PF		RP		BP		GH		VT		SF		RE		MH	
	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1
Mean	80.6	79.8	80.8	75.1	70.1	68.7	70.1	69.2	58.8	57.2	83.7	82.9	86.3	82.7	75.2	75.5
25th %ile	70	70	69	50	51	51	57	57	50	45	75	75	75	67	65	65
50th %ile	90	90	94	100	74	72	72	87	62	60	100	100	100	100	80	80
75th %ile	100	100	100	100	84	84	87	100	75	75	100	100	100	100	90	88
SD	25.1	25.5	26.9	36.9	24.3	24.0	21.3	21.9	20.7	21.4	23.6	23.8	22.5	32.2	18.4	17.8
Skewness	-1.5	-1.5	-1.4	-1.1	-0.6	-0.5	-0.8	-0.8	-0.5	-0.5	-1.5	-1.4	-1.8	-1.7	-1.1	-1.1
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	32.6	29.7	47.4	61.9	22.0	19.1	6.0	6.2	2.1	1.0	55.7	53.4	60.2	73.6	5.1	3.2
% Floor	1.0	1.0	2.1	13.6	1.0	0.9	0.0	0.3	1.1	1.0	1.0	1.1	1.1	8.8	0.1	0.1
% Complete	87.6	89.8	93.8	95.9	97.5	97.2	94.0	94.9	93.5	94.5	95.7	96.9	95.6	97.2	92.3	93.6
% Computable	98.9	99.5	98.1	98.3	99.3	99.5	98.8	99.3	98.0	98.5	99.4	99.6	98.2	98.8	98.7	99.4
Scaling Success	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Median Item Internal Consistency	.80	.79	.89	.75	.76	.76	.69	.70	.68	.75	.78	.76	.53	.69	.50	.66
Reliability	.94	.94	.95	.88	.85	.85	.83	.84	.85	.88	.88	.86	.93	.82	.85	.85

Note. v2 = SF-36v2, v1 = SF-36

Table 13.12

Comparison of Descriptive Statistics and Reliability Estimates for the Acute (1-Week Recall) Forms of the SF-36v2 and SF-36, 1998 U. S. General Population Using 0-100 Scoring

	PF		RP		BP		GH		VT		SF		RE		MH	
	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1	v2	v1
Mean	79.8	78.8	80.5	75.2	72.2	71.6	69.8	69.2	58.6	57.3	84.2	84.6	86.4	81.4	76.0	76.3
25th %ile	70	67	69	50	52	51	57	57	44	40	75	75	83	67	65	68
50th %ile	90	90	94	100	74	74	72	72	63	60	100	100	100	100	80	80
75th %ile	100	100	100	100	100	100	87	87	75	75	100	100	100	100	90	88
SD	25.8	26.4	27.8	36.9	24.6	24.7	21.8	22.1	21.5	22.1	24.1	23.6	23.2	33.3	18.7	17.5
Skewness	-1.4	-1.3	-1.4	-1.2	-0.7	-0.6	-0.8	-0.8	-0.6	-0.5	-1.6	-1.6	-1.9	-1.6	-1.2	-1.2
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	32.1	31.1	48.5	62.3	27.0	26.8	6.2	5.8	2.5	1.1	58.1	59.7	62.2	72.2	3.4	3.4
% Floor	0.9	0.7	2.8	13.4	0.9	1.0	0.2	0.3	1.4	0.9	1.2	0.8	1.7	9.7	0.4	0.1
% Complete	88.1	87.9	94.1	96.2	96.9	97.2	93.5	94.8	93.6	94.2	95.8	96.5	95.6	93.4	92.7	93.4
% Computable	99.0	99.2	97.9	98.3	99.3	99.3	98.8	98.9	98.9	98.2	99.3	98.6	98.4	98.9	98.6	98.9
Scaling Success	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Median Item Internal Consistency	.79	.79	.89	.75	.78	.77	.70	.72	.72	.75	.77	.76	.86	.70	.66	.64
Reliability	.94	.94	.95	.88	.86	.86	.84	.84	.87	.88	.87	.86	.93	.83	.86	.83

Note. v2 = SF-36v2, v1 = SF-36

Table 13.13

Product-Moment Correlations and Reliability Coefficients for the SF-36v2 Standard (4-Week Recall) Form Scales in the 1998 U.S. General Population (N = 7,069)

	PF	RP	BP	GH	VT	SF	RE	MH
PF	.94							
RP	.76	.95						
BP	.59	.67	.85					
GH	.59	.63	.60	.83				
VT	.46	.54	.54	.64	.85			
SF	.55	.68	.59	.60	.62	.88		
RE	.50	.63	.44	.50	.51	.67	.93	
MH	.28	.36	.38	.50	.65	.61	.58	.85

Note. Cronbach's alpha coefficients for health domain scales are indicated in **bold**.

Table 13.14

Product-Moment Correlations and Reliability Coefficients for the SF-36v2 Acute (1-Week Recall) Form Scales in the 1998 U.S. General Population (N = 7,837)

	PF	RP	BP	GH	VT	SF	RE	MH
PF	.94							
RP	.78	.95						
BP	.62	.67	.86					
GH	.60	.62	.59	.84				
VT	.47	.57	.55	.64	.87			
SF	.56	.69	.59	.59	.62	.87		
RE	.50	.64	.46	.48	.52	.68	.93	
MH	.30	.39	.39	.50	.66	.62	.60	.86

Note. Cronbach's alpha coefficients for health domain scales are indicated in **bold**.

The advantages of components analysis over principal factor analysis are noteworthy for the purposes of achieving (a) a simple additive model of content facilitating the interpretation of each scale, (b) summary measures that explain as much of the variance in the eight health domain scales as possible, (c) summary scales that are easy to estimate statistically, and (d) summary scores that are interpretable as physical and mental components of health. Seeing that the goal when constructing the PCS and MCS measures was to explain as much of the variance in the eight health domain variables as possible with only two summary measures (Ware, Kosinski, Bayliss, et al., 1995), components analysis was the logical choice as it attempts to do just that. In contrast, principal factors analysis attempts to reproduce the original correlation matrix (Harman, 1976). Further, the computation of scores for each principal component is a straightforward estimation using scores for the observed variables (i.e., the eight health domain scale scores), in contrast to the approximations involved in estimating scores for principal factors. These differences, along with the advantages of components analysis, are discussed in numerous texts

on factor analysis and psychometric methods (e.g., Harman, 1976; Nunnally & Bernstein, 1994).

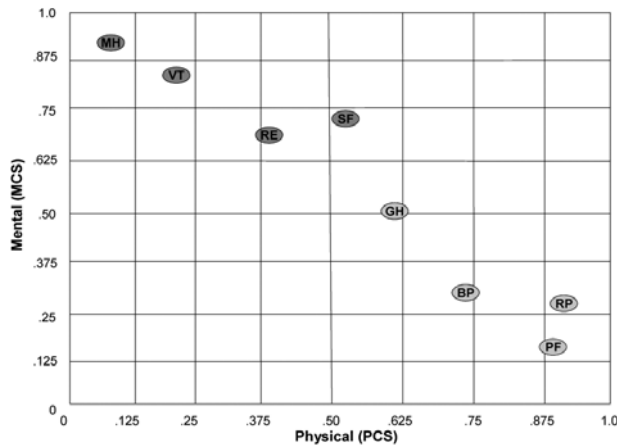
Orthogonal Components

There are good theoretical arguments for the use of both orthogonal and oblique factor rotations (Harman, 1976; Nunnally & Bernstein, 1994). As argued in numerous texts, orthogonal components proved to be ideal for the purposes of SF-36v2 development. The initial objective in analyzing the correlations amongst the health domain scales was to test the construct validity of the SF-36v2 and to establish guidelines for interpreting each scale on the basis of its physical and mental health components (McHorney et al., 1993; Ware, Kosinski, Bayliss, et al., 1995; Ware et al., 1993). For this purpose, orthogonal components, which are not correlated, have clear advantages. For example, *factor loadings*, which are product-moment correlations between scales and components, can be squared and summed across components to estimate the amount of variance in each scale accounted for by each component and the amount of variance in each scale that is explained by all components (i.e., the communality). As a result, the implications for the interpretation of each scale are more straightforward (Ware et al., 1993).

To provide a visual representation of the contribution of the eight health domain scales to the physical and mental components, their correlations with each component (i.e., factor loadings) are plotted in Figure 13.4 for the U.S. general population. This plot reveals a progression from the upper-left corner, with the MH scale correlating most highly with the MCS measure and least with the PCS measure, to the lower-right corner, with the PF scale correlating most highly with PCS and least with MCS. In between is a progression of scale correlations from MH (upper left) to PF (lower right) that corresponds to their order in the SF-36v2 profile.

Examining Figure 13.4, it is also apparent that the eight health domain scales form two distinct clusters, one with four scales (MH, RE, SF, and VT) correlating highest with the MCS measure and lowest with the PCS measure, and one with four scales (PF, RP, BP, and GH) correlating highest with the PCS measure and lowest with the MCS measure. The component score coefficients used to score the PCS and MCS measures correspond to these two clusters. Specifically, the highest positive physical health scale coefficients (.42 to .25) are used to weight the four best physical scales in scoring the PCS, and the highest positive mental health scale coefficients (.49 to .24) are used to weight the four best mental scales in scoring the MCS. Because the component score coefficients take into account the correlations

Figure 13.4 Plot of SF-36v2 Scale Factor Loadings on Orthogonal Physical and Mental Components for the 1998 U.S. General Population ($N = 7,069$)



among the eight scales, they differ from the factor loadings in that some are negative. Negative component score coefficients are also observed in oblique principal factor solutions and are not unique to the principal components method (see Ware & Kosinski, 2001a). Oblique solutions, which can allow substantial correlations between health components (factors), can facilitate the identification of factors but they also 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, thus 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 shown to be substantially correlated on cross-validation, there would have been good reason to favor an oblique solution. However, it is clear that physical and mental health are only weakly positively correlated. Correlations between the best physical (PF) and best mental (MH) health measures among the eight health domain scales are low, with medians ranging from only .22 to .30 based on 39 patient and general population studies in the United States, Germany, Sweden, and the United Kingdom. Cross-validation of the orthogonal two-component model (using United States factor score coefficients) in these samples has demonstrated very low empirical correlations between PCS and MCS scores, with medians ranging from $-.01$ to $.07$ across the 39 estimates available from these studies. These correlations would have been much larger upon cross-validation if the orthogonal solution and scoring were grossly distorted for any reason.

Further convincing evidence favoring the use of orthogonal principal components in summarizing SF-36 information about physical and mental health comes in

the form of their superiority in discriminating between physical and mental health outcomes in empirical tests. Comparisons of alternative scoring strategies revealed that much of the interpretive gains made when using orthogonal solution with the SF-36 PCS and MCS was lost when the physical and mental components were scored with an oblique solution (see Ware & Kosinski, 2001a).

Scoring the Component Summary Measures: Use of Positive and Negative Component Weights

What is the content of the SF-36 and SF-36v2 health domain scales and how do positive and negative component weights improve their validity in discriminating between physical and mental health outcomes? To answer these questions central to the scoring and interpretation of the surveys, the total variance in each health domain scale was divided into four parts: (a) the physical (PCS) component of health; (b) the mental (MCS) component of health; (c) the unique reliable variance, meaning the variance that is reproducible but not accounted for by either component; and (d) the error, which equals 1 minus the reliability coefficient. Each health domain scale manifests a different pattern, all of which are measured with some error, usually 10% to 20%, or less. PF and MH are the purest measures of the physical and mental components, respectively. The most complicated scales, in terms of factor content, are the middle four on the continuum: BP, GH, VT, and SF. (This finding should not be surprising; the SF items, for example, address both physical and mental health status.) When these middle scales are aggregated to compute a summary score, each scale adds information about more than one component of health, thereby substantially confounding the results. Note that negative coefficients remove the health scores that otherwise would be counted twice.

Principal component scores provide a proven solution to the confounding problem. The positive and negative coefficients involved are fundamental to the scoring of health domain scales that have complicated factor content (i.e., significantly correlate with two or more components). The use of positive and negative weights is not unique to orthogonal components. For example, when scoring the PCS measure, scores are aggregated using positive weights for the first five scales of the profile (PF, RP, BP, GH, and VT), which are the scales that contribute the most information about the physical component of health. However, these scales also have substantial correlations with the component of mental health. For a given individual, if the scores for the mental health scales (i.e., a different health outcome) are above

the mean, then they must be subtracted out to avoid inflating the score estimate for the physical component.

Conversely, if the mental health scores are *below* the mean, they must be added back in to avoid introducing downward bias into the estimate of the physical component score. This same logic holds true for the four scales on the right side of the profile (VT, SF, RE, and MH), scales which are positively weighted in estimating the mental component. When these scores are added to the MCS score, substantial information about a different health outcome (i.e., physical health) is also added to the estimate of the mental health component score. Therefore, to correct for the confounding of physical and mental health, negative coefficients for some scales subtract out the unwanted variance.

Proponents of orthogonal and of oblique scoring algorithms for summary health measures differ in the amount of correction they make; that is, in the amount of confounding, or overlap, they are comfortable with when scoring and interpreting summary scores. The PCS and MCS scores presented in this manual are orthogonal. PCS is an aggregation of the physical component of health as measured by all eight health domain scales, whereas MCS is an aggregation of the mental component of health as measured by the same eight scales. By minimizing their overlap (i.e., confounding), their validity in measuring a single component of health outcomes is maximized. In sharp contrast, oblique higher order factors derived from both versions of the SF-36 have substantial overlap, as evidenced by their substantial (i.e., high) interfactor correlations (approximately ranging from .50 to .70 in published studies; see Ware & Kosinski, 2001a). Also, oblique solutions have the disadvantage of resulting in negative scoring weights. At this time, the practical implications of this approach are largely unknown.

In conclusion, while some researchers have objected to the scoring method used for the PCS and MCS measures (in particular the negative weights; Nortvedt, Riise, Myhr, & Nyland, 2000; Taft, Karlsson, & Sullivan, 2001) and still other researchers have suggested alternative summary scores (Hays, Sherbourne, & Mazel, 1993), it is the authors' opinion that a careful examination of all empirical evidence supports the orthogonal principal components scoring of the PCS and MCS measures so described in this manual (see also Ware & Kosinski, 2001a).

Development of the PCS and MCS Measures

When the SF-36v2 was being developed, a decision was made to retain the same coefficients that had been developed for the SF-36 component summary measures. This decision was based on the results of studies showing that component score coefficients derived from the

1990 SF-36 normative data did not significantly differ from those derived from the 1998 SF-36v2 normative data. In addition, PCS and MCS scores obtained from the two sets of coefficients were found to correlate .99 or better, and no systematic bias was noted at any score level of either component summary measure. Moreover, maintaining the same component score coefficients provided a means of maintaining continuity between the revised instrument and the original survey. The only change made to the scoring of the SF-36v2 PCS and MCS measures was centering of each measure on a mean of 50 with a standard deviation (*SD*) of 10 in the 1998 U.S. general population. The meanings of the highest and lowest scores for the PCS and MCS measures are summarized in Chapter 7 (see Table 7.1).

Comparability of the SF-36 and SF-36v2 PCS and MCS Measures

Results of principal component analyses of SF-36v2 scales confirmed the two-component structure that has been well-documented for its predecessor. Results were consistent for both the standard and acute forms (see Table 13.15), as found in previous studies of the SF-36. For example, the PF scale had highest correlation with the *physical* component and the MH scale had highest correlation with the *mental* component. Furthermore, the magnitude and pattern of scale-to-component correlations across the SF-36 and SF-36v2 scales replicated the findings from previous studies of the SF-36 (McHorney et al. 1993; Ware, Kosinski, Bayliss, et al., 1995; Ware, Kosinski, & Keller, 1994), with the exception of the VT and RE scales. The differences for these two health domain scales require further empirical evaluation. Finally, the two components accounted for more than 70% of the total variance and more than 80% of reliable variance in the eight health domain scale scores across standard and acute forms of both versions of the survey.

In summary, the physical and mental health constructs underlying the SF-36 were replicated in the SF-36v2 for both standard and acute recall periods. The implication is that the health domain scales have the same content and interpretation, regardless of the version of the survey or the recall period. These findings support the construction and scoring of the two SF-36v2 component summary measures. Furthermore, to maintain the comparability of interpretations, SF-36 component score coefficients are used in scoring the SF-36v2. Moreover, as with the SF-36 component summary measures, it is clear that the SF-36v2 measures are able to reduce the eight health domain scales to two summary measures without substantial loss of information (Ware, Kosinski, & Keller, 1994, 1995).

Table 13.15

Correlations Between Health Domain Scales and Rotated Physical and Mental Health Components Across SF-36v2 and SF-36 Standard and Acute Forms, 1998 U.S. General Population

Scales	SF-36v2 Standard		SF-36 Standard		SF-36v2 Acute		SF-36 Acute	
	Physical	Mental	Physical	Mental	Physical	Mental	Physical	Mental
PF	.88	.14	.87	.13	.89	.16	.87	.16
RP	.89	.29	.83	.28	.85	.33	.82	.30
BP	.74	.32	.77	.31	.77	.30	.79	.29
GH	.61	.51	.61	.50	.61	.49	.64	.47
VT	.35	.77	.43	.70	.36	.76	.43	.69
SF	.53	.67	.55	.65	.52	.68	.55	.66
RE	.45	.64	.25	.70	.42	.68	.30	.73
MH	.08	.93	.10	.91	.09	.93	.11	.90
Variance Explained								
Total	74%		70%		75%		72%	
Reliable	84%		82%		85%		83%	

Note. Standard and acute correlations are slightly different from those presented in the first edition of the SF-36v2 manual (see Ware, Kosinski, & Dewey, 2000, Table 4.8). The differences reflect the ability of the SF-36v2 developers to use Missing Score Estimation (MSE) via the SF Health Outcomes Scoring Software (Saris-Baglama et al., 2004; see also Chapter 5) to maximize the useable data from the standard and acute norm groups (see Chapter 14). MSE was not available when the first edition of this manual was published.

The SF-6D

Although the SF-36 was not originally designed for use in economic evaluations or for determining quality adjusted life years (QALYs), research has shown that a meaningful health state classification measure—the SF-6D—can be created by applying a scoring method developed by Brazier and colleagues (Brazier et al., 2002; Brazier et al., 1998). The SF-6D consists of 11 items and focuses on seven of the eight health domains covered by the SF-36v2: physical functioning, role participation (combined role-physical and role-emotional), social functioning, bodily pain, mental health, and vitality. Only the general health domain is not included. The specific SF-36v2 areas or activities contributing to the scoring of this index include: ability to engage in both moderate and vigorous activities; ability to bathe and dress oneself; limitations in the kind of work or other activities as the result of physical health; accomplishing less due to emotional problems; bodily pain and its interference with normal work; nervousness, depression, and energy level; and interference with social activities due to physical or emotional problems.

Based on SF-36v2 data (or SF-12v2 data, when applicable), individual respondents can be classified on any of four to six levels of functioning or limitations for each of six domains (with RP and RE considered a single dimension and GH not included), thus allowing a respondent to be classified into any of 18,000 possible unique health states (O'Brien et al., 2003). Brazier et al. (2002) used the standard gamble valuation technique to

obtain utility values on 249 of the possible health states for 836 respondents in the United Kingdom. The resulting SF-6D index, scored from 0.0 (worst health state) to 1.0 (best health state), can be used in the assessment of the QALYs and the cost-effectiveness of various health care interventions. Note that utility weights for the SF-6D have been developed specifically for Great Britain. For a discussion of the advantages and disadvantages of developing country-specific weights, see Brazier et al. (2002), Brazier and Roberts (2004), and Walters and Brazier (2003).

Currently, the SF-36v2 and the SF-12v2 are the only health status measures available that can provide both a description of health (through their eight health domain scales and two component summary measures) and the means to conduct an economic evaluation (via the SF-6D utility index).

Advances Accompanying SF-36v2 Development

Several significant improvements in the Short Form instruments came about during or after the development of the SF-36v2. Notable among these improvements are changes to the standard profile in which Short Form scores are presented, calibration of scores from the Short Form instruments on a common metric, advances in estimating missing scores, and the development of translated versions of the SF-36v2 and additional translations of other Short Form instruments

The SF-36v2 Profile

The SF-36v2 was constructed to achieve at least the minimum standards of precision necessary for group comparisons across the eight health domains. Moreover, it was designed to yield a profile of scores that would facilitate understanding of individual respondent or population differences in physical and mental health status, the health burden of chronic diseases and other medical conditions, and the effects of treatment on general health status.

Figure 2.2 (see Chapter 2) illustrates the important features of the SF-36v2 profile of scores. The first scores presented, at the left side of the profile, are the PCS and MCS measure scores. Placement of the component summary measures at the beginning (left side) of the profile emphasizes the importance of first considering individual respondent or group results with regard to overall functioning in the physical and mental health dimensions when interpreting the data (see Chapters 7, 11, and 12). Doing so provides a context in which relative deficits and strengths, as indicated by the eight health domain scale scores in the profile, can be more accurately interpreted. Following the component summary measures, note that the score profile orders the eight health domain scales from left to right, from the best physical health measure (PF) on the left side to the best mental health measure (MH) on the right side. This ordering facilitates interpretation of the profile, with differences on the left side of the health domain section of the profile reflecting physical health status and differences on the right side reflecting mental health status. The empirical evidence for this ordering of the scales is presented earlier in this chapter and in Chapter 16 as well.

Calibration of Scales on a Common Metric

As illustrated in Figure 13.5, each Short Form scale measuring the same health domain, including those dynamically administered and across all Short Form instruments, is scored on the same “ruler,” or metric. Each assessment tool differs only in terms of precision, with the single-item scales distinguishing fewer levels (only 5 or 6) in comparison with the SF-36v2 scales, which distinguish from 9 to 21 levels. Moreover, for each health domain, an item bank has been created that contains SF-8, SF-12v2, and SF-36v2 items, along with many others; is calibrated on a common metric (mean = 50, $SD = 10$); and can be dynamically administered using QualityMetric Incorporated’s DYNHA software (see Chapter 1), which utilizes computerized adaptive testing (CAT) logic. When using this mode of administration, items from the item bank are selected and administered

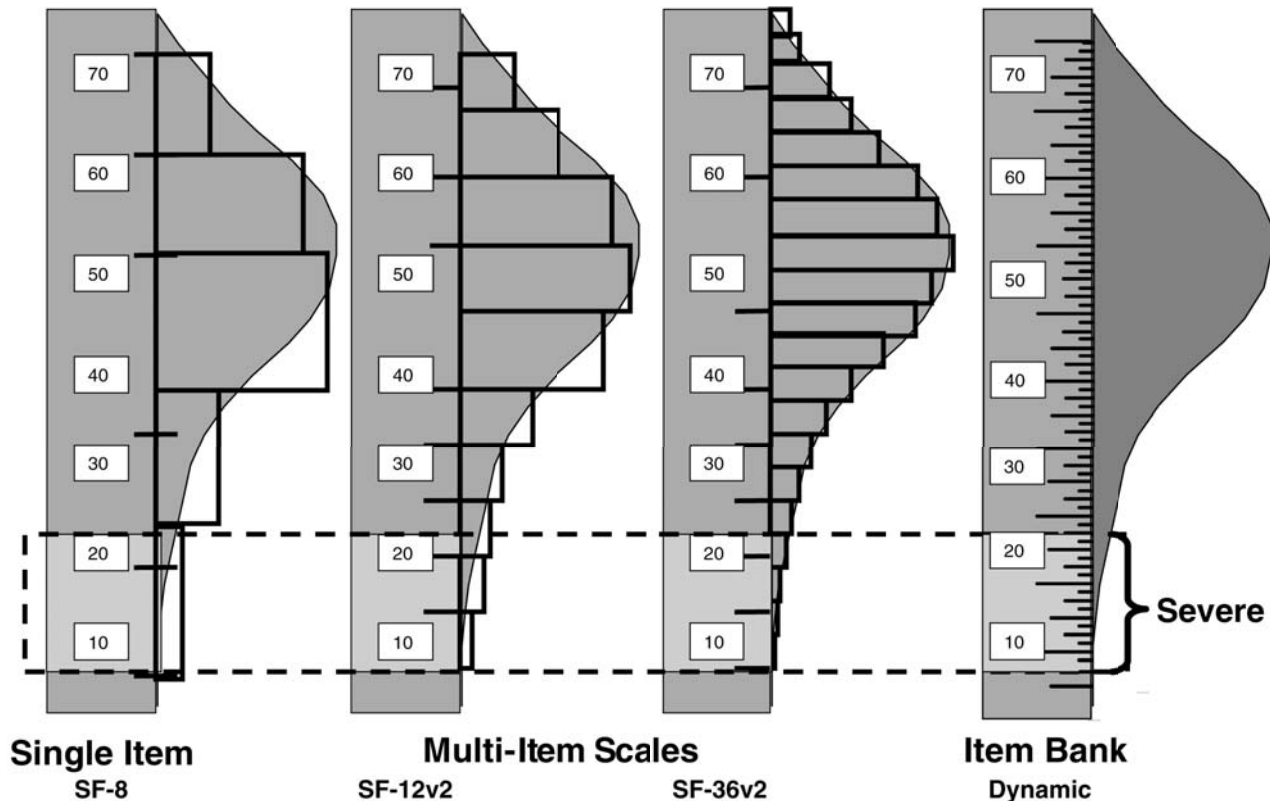
only if they match a given respondent’s level of health. Computerized dynamic methods used for this purpose have been shown to very quickly provide reliable estimates of health status scores throughout the range of health levels (Bjorner & Ware, 1998; Ware, Bjorner, & Kosinski, 2000; Ware, Kosinski, et al., 2003).

In comparison with the other methods illustrated in Figure 13.5, the SF-8 ruler is the most coarse, defining only five or six levels of each health concept. In comparison, the SF-36v2 scales, particularly the two role-functioning scales, define more levels over a wider range and yield more reliable score estimates. However, in the 1998 and 2009 U.S. general population samples, the SF-8, SF-12v2, and SF-36v2 scales all have means of 50 and standard deviations of 10 when using the *T*-score algorithms described in Chapters 5 and 14 of this manual. Thus, *T*-score algorithms enable the direct comparison of SF-8, SF-12v2, SF-36v2, and DYNHA PCS and MCS scores. As such, in studies based on large representative samples, the SF-8, SF-12v2, and SF-36v2 measures yield directly comparable estimates of average population scores. In principle, these estimates differ only, on average, in terms of their precision; however, shorter scales are also more prone to ceiling and floor effects.

Missing Score Estimation

As described in Chapter 5, the standard approach for handling missing response data has been to replace the missing score with a person-specific estimate derived by calculating the mean response value to the answered items in the same scale when a respondent has answered at least half of the items in that scale. For example, if a respondent has one missing item response on the five-item MH scale, the average final item value across the four completed items would replace the missing value. This standard algorithm for estimating missing scores is referred to as the *Half-Scale Rule* in the QualityMetric Health Outcomes Scoring Software 5.0 (see Chapter 5; see also Saris-Baglana et al., 2011). If users of the Scoring Software 5.0 prefer not to apply a method of missing score estimation, the option to choose the *No Missing Score Estimation* method for scoring, which requires all items on a given scale be completed, is also available.

Although the Half-Scale Rule for missing score estimation makes it possible to estimate scores for health domain scales and component summary measures when some data are missing, the resulting missing score estimates are biased in scales that are constructed from items presented in a hierarchical order (e.g., the PF scale) and require that the respondent answer at least half of

Figure 13.5 Scoring of the SF-8, SF-12v2, SF-36v2, and Dynamic Health Assessments on a Common Metric

the items in each of the eight scales. Additionally, the PCS and MCS scores cannot be estimated when one or more scale scores are missing. This may be problematic at times, particularly when analyzing data obtained from older respondents or other populations that have higher rates of missing data (McHorney, Ware, et al., 1994).

Because of the rules regarding application of standard missing data estimation algorithms, improved algorithms have been developed and evaluated (Kosinski, Bayliss, Bjorner, & Ware, 2000). These improved methods of missing score estimation were obtained through (a) dropping the Half-Scale Rule and adopting the *Full Missing Score Estimation (Full MSE)* method, so that a given health domain scale score (except for PF) can be estimated when the respondent provides a response to at least one item in said scale; (b) using item response theory (IRT) to develop a model for estimating scores on the PF scale; and (c) using regression methods to estimate component summary measure scores on the basis of the available scales.

Modern psychometric methods provide for improved accuracy in the scoring of surveys with missing data (Bjorner & Ware, 1998). For the PF scale, an estimated score generated with an IRT model is preferred to the person-specific mean final item response value because items vary greatly in difficulty across the scale.

For example, Item 3a defines physical functioning in terms of one's limitations in engaging in "vigorous activities," whereas Item 3j defines physical functioning in terms of one's limitations in "bathing and dressing." Depending on which items are answered, the mean response value may not yield a very precise estimate of a respondent's level of physical functioning. Through IRT models, one can generate item parameters that indicate the probability of a respondent selecting a particular response to a particular item, based on their responses to previously answered items. This method results in a more precise estimation of a missing value. A more detailed discussion of how IRT is used to estimate a score on the PF scale when only a few of the PF items have been answered can be found in Appendix C of this manual.

While IRT models are able to improve the accuracy of missing data estimation for the PF items, these models do not appear to substantially improve the accuracy of estimates obtained through person-specific mean substitution for the remaining seven health domain scales. This may be due to the internal consistency of the items that are comprised by the other scales, causing the estimates based on person-specific data to be psychometrically sound. Based on this evaluation, the IRT method for missing data estimation for the other seven scales has not been

adopted for the scoring of either version of the SF-36.

Furthermore, the traditional scoring algorithms of the Half-Scale Rule allowed calculation of the PCS and MCS scores only when a respondent had scores for all eight scales. To investigate scoring algorithms that would allow for the calculation of the PCS and MCS scores when a scale score is missing, multiple regression models were developed using data from the 1990 National Survey of Functional Health Status (NSFHS) and the MOS (Kosinski, Bayliss, et al., 2000). One set of models used the PCS score as the dependent variable and different combinations of SF-36v2 health domain scales as independent variables, whereas another set of models used the MCS score as the dependent variable and different combinations of SF-36 scales as independent variables. The stability of regression coefficients in predicting PCS and MCS scores across general and clinical populations was evaluated and confirmed. These analyses made it possible to generate algorithms for calculating the PCS score when a respondent has at least seven health domain scale scores and the PF scale score is not missing, and likewise for calculating the MCS score when a respondent has at least seven health domain scale scores and the MH scale score is not missing.

Using the standard missing score estimation algorithms, Kosinski, Bayliss, et al. (2000) found that the component summary measure scores could not be estimated for approximately 7% of elderly respondents in the NSFHS and approximately 17% of elderly respondents in the MOS. However, these percentages were significantly reduced (to 3.97% and 12.42%, respectively) when using the new missing score estimation algorithms. In addition, Kosinski et al. evaluated the accuracy of these estimated component summary measure scores by introducing missing data amongst respondents with complete data in the NSFHS and MOS studies. This allowed for a direct comparison between a respondent's estimated PCS and MCS scores and their actual PCS and MCS scores based on complete data. Although there may have been a small loss of precision for component summary measure scores estimated with seven scales, the degree of agreement between estimated and actual PCS and MCS scores was very high. Correlations between the estimated and actual component summary measure scores ranged from .95 to .99 for PCS and from .94 to .99 for MCS. Further, the mean estimated PCS and MCS scores never differed by more than 1.1 *T*-score points in comparison with the observed scores in either study population. These results suggest that one can use established norms for interpreting estimated PCS and MCS scores.

Translations

As of August 2011, more than 140 translations and English-language adaptations of the SF-36v2 had been completed pursuant to the International Quality of Life Assessment (IQOLA) Project in 1991 (see Chapter 1). The IQOLA Project began with the goal of translating the SF-36 for international use in 14 countries. To meet this goal, the IQOLA Project team adopted a multistage translation procedure designed to assure that translations of the SF-36 were not only conceptually equivalent to the U.S. source form but also linguistically and culturally relevant (Aaronson et al., 1992; Bullinger et al., 1998). In brief, the IQOLA translation process included the development of one initial forward translation from multiple independent translations, backward translation of the forward translation into English, a review of the backward translation for conceptual equivalence with the source form, and pilot testing of the translation among native speakers. In several countries, independent judges rated the translation on clarity, use of common language, conceptual equivalence, and overall acceptability. A Thurstone-like scaling exercise was also used to inform the selection of response choices in many countries (Keller et al., 1998). In addition, a harmonization meeting was held amongst investigators from the first dozen countries to join the project (Wagner et al., 1998). Overall, the psychometric properties of many of the IQOLA translations have been thoroughly evaluated, as documented elsewhere (see Gandek & Ware, 1998).

As previously discussed, many of the changes found in the SF-36v2 had already been incorporated into translations of the SF-36 during the translation process. For example, the SF-36 item "full of pep" (Item 9a) was not translated literally, as *pep* is not a common word outside of the United States; rather, synonyms for *pep* (energy, life) were used in the SF-36 translations (Wagner et al., 1998). Thus, the SF-36 translation of this item often could be used in the corresponding SF-36v2 translation, as the translated item was already equivalent to the SF-36v2 version of this item (i.e., "full of life"). Moreover, translation of the "block" items in the PF scale (Items 3h and 3i) led to a discussion amongst the IQOLA investigators as to what a *block* means in the United States, since a block in New York City could be quite different than a block in Texas or California for example. Also, because people in European countries (which made up the majority of the original 14 IQOLA Project countries) generally estimate distances in terms of meters, translations used a standard of "100 hundred meters" for "one block." This is approximately equal to the "100 hundred yards" terminology used in the SF-36v2. Similarly, many

of the IQOLA teams struggled with the negative wording of the RE item “didn’t do work or other activities as carefully as usual” (Item 5c), leading to the adoption of “did work or other activities less carefully than usual” for the translations, which is also the SF-36v2 wording for this item.

Thus, the initial development of an SF-36v2 translation began with an evaluation of which changes represented in the SF-36v2 source form were already incorporated in the SF-36 translations. For example, if Item 3i (“walking 100 hundred yards”) was already translated as “walking 100 hundred meters” in the SF-36 translation, then the translation of that item was simply retained in the SF-36v2 translation. SF-36v2 items that were not equivalent were translated using the process of forward and backward translation previously described, unless the changes were very minor. In many cases, a

full, independent review of a given SF-36 translation was undertaken prior to finalization of the equivalent SF-36v2 translation. Development of new SF-36v2 translations follows the standard process adopted by the IQOLA team, which is comparable to the guidelines recommended by such organizations as the Scientific Advisory Committee of the Medical Outcomes Trust (2002), the ISPOR Task Force for Cultural Adaptation and Translation (Wild et al., 2005), and many European organizations (Acquadro et al., 2003; Apolone, De Carli, Brunetti, & Garattini, 2001; Chassany, Sagnier, Marquis, Fullerton, & Aaronson, 2002).

Information related to translation methodology can be found at <http://www.iqola.org>. For additional information regarding translations of the Short Form instruments, visit <http://www.sf-36.org> or <http://www.qualitymetric.com>.

14

2009 Normative Data

Normative data make it possible to interpret SF-36v2 health domain scale and component summary measure scores by comparing them with the distribution of scores for other respondents. As such, scores can then be understood as departures from expected or typical scores, which are referred to as *norms* and can be computed at the individual respondent or group level. At the individual respondent level, norms are the scores that are typical of a respondent under stable conditions. At the group level, norms are the average values for a given group and can be calculated based on a sample from the population of interest. Importantly, norm-based comparisons require valid norms for a comparison group of interest. When such norms are available, norm-based interpretation can help to determine whether an observed score is typical. In other words, whether the observed score is one that would be expected for a given respondent or group of respondents.

With the passage of more than a decade since the development of the 1998 norms, the developers of the SF-36v2 determined that updated norms were necessary to ensure that all the Short Form surveys remained current and relevant to their users' needs. The normative data that were collected during the QualityMetric 2009 Norming Study allowed for this important updating of the SF-36v2's norms, as well as to the norms for the SF-12v2 health domain scales and component summary measures. Note that SF-8 normative data were also gathered during the 2009 study.

A primary goal of the QualityMetric 2009 Norming Study was the development of updated norms for the SF-36v2, SF-12v2, and SF-8 based on a large, representative sample of the U.S. general population. Normative data for other HRQOL surveys published by QualityMetric Incorporated were also collected as part of this project. Simultaneously collecting normative data for these other instruments allowed not only for the updating and/or further validation of these surveys but also for the fur-

ther validation of the SF-36v2 and the development of additional ways to interpret the meanings of SF-36v2 scores (see Chapter 9).

The purpose of this chapter is to detail the sampling and data collection methods used in the QualityMetric 2009 Norming Study, as well as to discuss the aspects of the study that directly pertain to the development of the updated SF-36v2 norms. Also included in this chapter is a presentation of the 2009 SF-36v2 total-sample U.S. general population normative data for the standard and acute forms, descriptions of the characteristics of the samples completing each of the forms, and a discussion of the development of the 2009 SF-36v2 supplemental (age, gender, disease-specific) norms and benchmarks, which are available through the scoring services offered by QualityMetric Incorporated and its authorized resellers. In addition, general discussions of the norming study's data collection instruments, formats for online item presentation, development of scoring algorithms, and finalization of the 2009 normative samples are presented here. Finally, this chapter presents a comparison of the 2009 and 1998 SF-36v2 normative data.

How the SF-36v2 Was Renormed

The QualityMetric 2009 Norming Study was conducted via the Internet using a sample of adults, aged 18 years and older, drawn from the U.S. general population panel maintained by Knowledge Networks (KN). The primary purpose of this study was to develop updated normative data for three of QualityMetric's Short Form health status instruments. In addition to the SF-36v2, SF-12v2, and SF-8 surveys, data also were collected to develop or update normative and validation data for other new and established QualityMetric proprietary measures, including: the 12- and 6-item Medical Outcomes Study Sleep Scale-Revised measures (MOS Sleep-R),

Medical Outcomes Study Cognitive Functioning Scale–Revised (MOS COG–R), Premenstrual Symptoms Impact Survey (PMSIS™), and Pain Impact Questionnaire™ (PIQ-6™). In all, normative data were collected for the following: the SF-36v2, standard (4-week) and acute (1-week) recall forms; SF-12v2, standard and acute recall forms; SF-8, standard, acute, and 24-hour recall forms; 12- and 6-item MOS Sleep–R, standard and acute recall forms; MOS COG–R, standard and acute recall forms; PIQ-6, standard and acute recall forms; and PMSIS standard recall form. (Note that the PMSIS was not included in any of the three study forms that were used to collect SF-36v2 data.) Other data were also collected for purposes of the validation and interpretation of the instruments' findings. A full description of each of the four study forms can be found later in this chapter.

Sampling

Source. The 2009 SF-36v2 normative data came from a national probability sample of U.S. noninstitutionalized adults aged 18 years and older, drawn from the KnowledgePanel maintained by Knowledge Networks (KN). An oversample of respondents aged 65 years or older was also included. Ninety percent of the total sample was selected at random from the entire KN adult panel, and the remaining 10% of the total sample was selected at random from those panel participants aged 65 years or older.

Sample size determination. Sample sizes were determined based on how large a sample was needed to analyze specific age and gender subgroups for the SF-36v2 standard (4-week) recall form. First, the smallest subgroup of interest was identified, based on the 1998 normative data, and the largest standard deviation (*SD*) across SF-36v2 scales and summary measures for that group was determined. Note that all summary measures and domain scales were scored on the *T*-score metric, with a mean of 50 and an *SD* of 10. Examination of the 1998 SF-36v2 norms data revealed that the smallest subgroup of interest was males aged 75 years or older and that the largest *SD* for that subgroup (12.79) was found in the Role–Emotional scale.

Second, study investigators determined that a 95% confidence interval of ± 2 *T*-score points provided an acceptable degree of imprecision, because a 3-point difference is considered significant for interpretation purposes. Thus, using an *SD* of 13 for the SF-36v2 group of males aged 75 years or older, it was determined that a sample size of 169 males aged 75 years or older was required for the renorming study. According to the 2009 U.S. Census data, males aged 75 years or older composed approximately 3% of the U.S. adult population aged 20

years or older. As previously mentioned, participants aged 65 years or older were to be oversampled, and the *SD*s for most scales in the other age/gender groups were generally in the 9-to-10 *T*-score point range. Therefore, investigators determined that the total sample size required was approximately 4,000 (169/.04). As a result, the target sample size was 8,000 respondents, of whom 4,000 were to complete the SF-36v2 (standard form) and other instruments; 2,000 were to complete the SF-36v2 (acute form) and other instruments; and the final 2,000 were to complete the SF-12v2 Health Survey–Mental Health Enhanced (SF-12v2–MH Enhanced; Ware et al., 2010) with the Self-Evaluated Transition (SET) item (standard form), and other instruments.

Data Collection Forms

As indicated in Table 14.1, SF-36v2 normative data were gathered using three of the four study forms (Forms A, B, and C), which also collected data on other HRQOL instruments and related variables. Forms A and B included SF-36v2 standard (4-week recall) form items, whereas Form C included SF-36v2 acute (1-week recall) form items. A variety of published HRQOL surveys and other sets of items were administered along with the SF-36v2 items in each of the three study forms. Each collateral survey and item set, summarized in Table 14.1, is described in the following sections.

SF-8 Health Survey. The SF-8 (Ware, Kosinski, Dewey, & Gandek, 2001) contains 8 items, only one of which is identical to any of the items found in the SF-36v2. Although the SF-8 items are not a direct subset of SF-36v2 items, both the SF-8 and the SF-36v2 measure the same eight health domains. Whereas the SF-36v2 uses between 2 and 10 items to measure each health domain, the SF-8 uses just one item for each health domain, making it less burdensome to complete and a good alternative to the SF-36v2 for large-scale population survey efforts. Similar to the SF-36v2, the PCS and MCS measures can be calculated from SF-8 results. Among the disadvantages of the SF-8 is that its scores generally cover a narrower range of the measured constructs, are more coarse (i.e., define fewer levels) for some scales, and are less precise. The SF-8 is available in a standard (4-week recall) form, an acute (1-week recall) form, and a 24-hour recall form.

MOS Sleep Scale–Revised (MOS Sleep–R). The MOS Sleep Scale–Revised (MOS Sleep–R) is a brief, generic, self-administered assessment designed to measure key aspects of sleep, such as disturbance, adequacy, somnolence, and quantity, in either general or clinical populations. Two versions of the scale are available: a 12-item version and a 6-item version, with each having

Table 14.1

Composition of QualityMetric 2009 Norming Study Forms Used to Collect SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Normative and Validation Data

Survey Template	# Items	Form A ^a (N = 2,039)	Form B ^a (N = 2,005)	Form C ^b (N = 2,062)
SF-36v2, standard recall, single-item format	36	•		
SF-36v2, standard recall, grid format	36		•	
SF-36v2, acute recall, single-item format	36			•
SF-8, standard recall	8	•	•	
SF-8, acute recall	8			•
MOS Sleep–Revised, 12-item, standard recall	12	•		
MOS COG–Revised, 6-item, standard recall	6		•	
PIQ-6 (the 4 non–SF-36v2 items), standard recall	4		•	
MOS Sleep–Revised, 12-item, acute recall	12			•
Validation items	12	•	•	•
Healthcare utilization items	7	•	•	•
Condition checklist	40 ^c	•	•	•
Background items	8	•	•	•

^aUsed to collect SF-36v2 standard form data.

^bUsed to collect SF-36v2 acute form data.

^cTwenty-six *yes/no* items asking “Have you ever been told by a doctor or other health care professional that you have...” (grouped into 4 grids of 6 items each and 1 grid of 2 items) and 14 *yes/no* items asking “Do you *now* have...” (grouped into 2 grids of 6 items each and 1 grid of 2 items). For each endorsed condition, the respondent was also asked how much the condition limited “your usual activities or enjoyment of everyday life.”

both standard (4-week recall) and acute (1-week recall) forms. Each version yields a sleep problem summary score, with the 12-item version also yielding scores on several subscales.

The original MOS Sleep Scale was developed for use in the Medical Outcomes Study (MOS; Stewart & Ware, 1992; Tarlov et al., 1989; Ware, Bayliss, Rogers, Kosinski, & Tarlov, 1996) to assess an HRQOL concept—sleep—that is relevant to everyone’s health status and well-being (Ware, 1987, 1990a) and known to be directly affected by disease and treatment (Stewart & Ware, 1992). The revised version of the MOS Sleep Scale, the MOS Sleep–R, was constructed for and administered to adults during the 2009 norming study. Otherwise identical to the original, the revised version uses five response options, instead of six, for the items with similar response styles. The MOS Sleep–R also demonstrates improved psychometric properties with regard to response distributions and utilizes a *T*-score metric.

MOS Cognitive Functioning Scale–Revised (MOS COG–R). Also developed for the MOS, the original MOS Cognitive Functioning Scale was available in both four- and six-item versions (Stewart, Ware, Sherbourne, & Wells, 1992). The revised version of the original scale, the MOS Cognitive Functioning Scale–Revised (MOS COG–R), is a six-item, self-administered assessment that measures a range of day-to-day problems in cognitive functioning, such as memory, attention, and reasoning. Two forms of the scale are available: a standard (4-week recall) form and an acute (1-week recall) form.

Each form yields a summary score indicating the general level of cognitive functioning. The MOS COG–R was constructed for and administered to adults during the 2009 norming study. Otherwise identical to the original six-item version, the revised scale uses five response options, instead of six, for all the items. Moreover, like the MOS Sleep–R, the MOS COG–R demonstrates improved psychometric properties with regard to response distributions and utilizes a *T*-score metric.

Pain Impact Questionnaire (PIQ-6). The Pain Impact Questionnaire (PIQ-6; Becker, Saris-Baglama, Kosinski, & Bjorner, 2005) is a brief, six-item, patient-based assessment designed to measure pain severity and the impact of pain on an individual’s emotional well-being and work and leisure activities. Two of the PIQ-6’s items compose the SF-36v2 BP scale and were administered in the norming study as part of the SF-36v2. The PIQ-6 is available in a standard (4-week recall) form and an acute (1-week recall) form.

Validation items. This set of 12 items asked the respondent questions regarding his or her employment status and performance, health and quality of life, and the presence of depressive symptoms.

Health care utilization items. These seven items addressed the respondent’s utilization of and satisfaction with health care services.

Condition checklist with global disease impact items. This group of *yes/no* items asked, “Have you *ever* been told by a doctor or other health professional that you have...” regarding a list of 26 conditions and asked,

“Do you *now* have...” regarding a list of 14 conditions. For any acknowledged condition, the respondent was then asked, “In the *past 4 weeks*, how much did your [condition] limit your usual activities or enjoyment of everyday life?”

Background items. Eight items were used to elicit a variety of information about the respondent, including his or her physical attributes (e.g., height) and habits, use of tobacco and alcohol, and level of stress.

Data Collection

Because both the 1990 SF-36 and 1998 SF-36v2 norming surveys were conducted in the autumn, the study investigators wanted to collect the 2009 normative data in the autumn as well; however, they did not want to wait until autumn of 2009 to start collecting data. Therefore, data were collected during two time periods: June and July 2009 (Wave 1) and September and October 2009 (Wave 2). Approximately three fourths of the data were collected in Wave 1, with the remaining one fourth collected in Wave 2. In both data collection waves, all KN panel members who were selected to participate received an invitation from KN. Wave 1 participants were randomly assigned to complete one of the four survey forms and to either the QualityMetric (QM) server or the KN server. All Wave 2 participants were assigned to the KN server.

It is important to note is that all 2009 SF-36v2 normative data were collected via online technology, using

one of two item-presentation formats. SF-36v2 items administered as part of Form A were presented as single items (see Figure 14.1). On the other hand, SF-36v2 items administered as part of Form B were presented in an item-grid format (see Figure 14.2). Only the single-item presentation format was used to collect SF-36v2 acute form data (Form C).

In Wave 1, data were collected via two servers, the QM server and the KN server. The original plan called for all data to be collected on the QM server, using a portal that was developed at QualityMetric. However, about 20% of the KN panel did not own a computer. For the sample to be as representative of the U.S. general population as possible, it was important to include individuals who did not own computers. Therefore, this subgroup completed surveys using a WebTV connection supplied by Knowledge Networks. Due to limitations in WebTV technology, the survey forms could not be viewed on the QM server via the WebTV connection. As such, it was necessary for all WebTV users to complete the survey on the KN server. In addition to those participants who completed surveys via WebTV, a smaller group of participants who also did not own computers completed surveys using a laptop PC that was supplied by Knowledge Networks. As a result, respondents who completed surveys in Wave 1 fell into three groups: those who used their own PC, those connected via WebTV, and those who used a KN-supplied laptop.

Figure 14.1 Sample SF-36v2 Single-Item Screen Presentation Used for the QualityMetric 2009 Norming Study

Section 1: Your Health and Well-Being Exit

In general, would you say your health is:

Excellent
Very good
Good
Fair
Poor

← PREVIOUS PROGRESS NEXT →

Figure 14.2 Sample SF-36v2 Item-Grid Screen Presentation Used for the QualityMetric 2009 Norming Study

Section 1: Your Health and Well-Being Exit

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	Some of the time	A little of the time	None of the time
a. Did you feel full of life?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
b. Have you been very nervous?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
c. Have you felt so down in the dumps that nothing could cheer you up?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
d. Have you felt calm and peaceful?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
e. Did you have a lot of energy?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

← PREVIOUS
PROGRESS
NEXT →

This split sample ultimately resulted in approximately 25% of Wave 1 respondents completing surveys in the standard KN format on the KN server. In other words, if the participant was assigned to the KN server, then he or she completed the entire survey on that server. After agreeing to complete the survey, participants were given a brief, online introduction to the survey, which was followed by the first survey question. Each respondent then completed the entire survey on the KN server, with the final items asking about 2008 household income and household size.

The remaining Wave 1 respondents began their surveys on the KN server but were then passed to a QualityMetric-maintained site to complete the majority of the survey. The QM server was located offsite at Rackspace, which had the capacity to allow a large number of subjects to complete the survey at the same time and offered 24-hour monitoring of the server. As a precaution, a feasibility test of the QM server was first conducted to ensure that surveys could be simultaneously completed by multiple users.

Wave 1 participants assigned to the QM server began the survey on the KN server. Participants were first given a description of the survey, including information about how to complete the survey on the QM server and assurances that data were being collected by a trusted partner (KN). Each respondent was then seamlessly transferred to the QM server to complete the assigned survey form's items. Upon completion of the survey, respondents were transferred back to the KN server and asked the same questions about 2008 household income and household

size as those participants assigned to the KN server. In other words, all participants started and ended their surveys on the KN server. Note that Knowledge Networks had previously conducted similar surveys in which participants were transferred from the KN server to a non-KN server for data collection; thus, most of the KN panel likely had prior experience in completing surveys outside of the KN server.

At the conclusion of the Wave 1 data collection, survey data were downloaded from the QM server. After limited data cleaning, a data file was then sent to KN. The initial and resulting sample sizes and completion rates for Wave 1, by assigned server, are provided in Table 14.2.

In Wave 2, all three categories of participants (own PC, WebTV, KN-supplied laptop) were sampled and all data were collected via the KN server. At the conclusion of both data collection waves, KN merged the survey data from the KN and QM servers with other data regarding the participants that KN had already collected (e.g., sociodemographic data) and created sampling weights. A preliminary file, containing Wave 1 data from both the QM and KN servers, was made available by KN, which was analyzed until the final (Waves 1 and 2 combined)

Table 14.2

Survey Wave 1 Sample Sizes and Completion Rates, by Server

Server	Sample Size	Surveys Completed	Completion Rate
Knowledge Networks (KN)	2,247	1,502	66.8%
QualityMetric (QM)	6,643	4,309	64.9%

Table 14.3*Sample Group Summary, by Study Wave*

	Sample Type		Survey Form				Method of Internet Access		
	General Population	65+ Oversample	A	B	C	D	PC	Web TV	Laptop
Wave 1									
QM server	3,730	579	1,073	1,067	1,124	1,045	4,231	0	78
KN server	1,433	69	387	385	360	370	333	1,094	75
Total	5,163	648	1,460	1,452	1,484	1,415	4,564	1,094	153
Wave 2									
Total	2,115	186	579	553	578	591	1,655	433	213
Combined									
Total	7,278	834	2,039	2,005	2,062	2,006	6,219	1,527	366

data file was available. Table 14.3 summarizes the respondent totals by sample type, survey form, method of Internet access, survey wave, and overall.

As previously mentioned, the study design called for an oversampling of the participants aged 65 or older. In addition, respondents were randomized to one of four survey forms, each differing in content. Finally, efforts were made to ensure that respondents were culled from across the groups using PCs, WebTV units, and KN-provided laptops to complete surveys. The initial and resulting sample sizes and completion rates for each survey wave and for the two waves combined are provided in Table 14.4.

Survey Readministration

As just discussed, data were collected in two waves during the renorming study. Within the sample, a subsample of 607 study participants completed the same survey form in both waves. Approximately the same number of respondents twice completed each of the four study forms. Note that this subsample's second wave data were not included in the main analysis; however, these data were used to study the stability of the study forms' instruments and the predictive validity of the SF-36v2 (see Chapter 15) and SF-12v2.

Sample Characteristics

Table 14.5 summarizes important demographic characteristics of the participants who completed

study forms that included the SF-36v2 standard form items (Forms A and B). Similarly, Table 14.6 presents the demographic characteristics of participants who completed the study form that included the SF-36v2 acute form items (Form C).

2009 U.S. General Population Norms

Development of the Health Domain Scale Scoring Algorithms

The algorithms used to score the SF-36v2 were designed to be as simple as possible, to satisfy the assumptions of the methods used to construct the SF-36v2 health domain scales and component summary measures, and to maximize comparability with the 1998 SF-36v2 scores throughout their in-common range, so as to preserve the original interpretations of the scales and measures. The only scoring change concerns the centering of each scale. Specifically, a linear *T*-score transformation method was used so that the norm-based scores for each of the health domain scales and component summary measures have a mean of 50 and an *SD* of 10, based on the 2009 U.S. general population sample. Thus, scores above and below 50 are above and below the average, respectively, found in the 2009 U.S. general population. Also, because the *SD* is 10, each 1-point difference or change in scores has a direct interpretation; that is, 1 point is one-tenth of an *SD*, or an effect size of 0.10.

Development of the PCS and MCS Scoring Algorithms

When the 2009 SF-36v2 scoring algorithms were being developed, a decision was made to retain the same coefficients that had been developed for the 1990 SF-36 and 1998 SF-36v2 component summary measure norms. This decision was based on the results of studies showing that component score coefficients derived

Table 14.4*Overall Sample Sizes and Completion Rates, by Study Wave*

Survey Wave	Sample Size	Surveys Completed	Completion Rate
1	8,890	5,811	65.4%
2	3,399	2,301	67.7%
Total	12,289	8,112	66.0%

Table 14.5*Demographic Characteristics of the SF-36v2 Standard (4-Week Recall) Form, 2009 U.S. General Population (N = 4,040)*

Characteristic	n	Total %	Study Form		Method of Internet Access		
			A ^a %	B ^b %	PC %	Web TV %	Laptop %
Gender							
Male	1,995	49.4	25.2	24.2	38.2	9.4	1.7
Female	2,045	50.6	25.2	25.4	38.6	9.5	2.5
Age							
18–24	324	8.0	4.1	3.9	7.6	0.2	0.3
25–34	530	13.1	6.2	6.9	12.4	0.3	0.5
35–44	654	16.2	8.3	7.9	13.8	1.7	0.7
45–54	694	17.2	8.6	8.6	11.7	4.3	1.1
55–64	771	19.1	9.8	9.3	12.3	5.8	1.0
65–74	749	18.5	9.8	8.7	14.0	4.2	0.4
75+	318	7.9	3.6	4.3	5.2	2.5	0.3
Marital status							
Married	2,116	52.4	26.2	26.2	44.5	6.3	1.6
Widowed	264	6.5	3.3	3.2	3.6	2.7	0.3
Divorced	522	12.9	6.6	6.4	7.6	4.4	0.9
Separated	67	1.7	0.8	0.9	0.9	0.5	0.2
Never married	834	20.6	10.8	9.8	15.6	4.3	0.7
Living with partner	237	5.9	2.8	3.0	4.7	0.7	0.5
Employment status							
Working	2,153	53.3	26.6	26.7	44.7	6.8	1.9
Retired/disabled	1,304	32.3	16.9	15.4	20.9	9.9	1.5
Temporarily laid off/looking for work	300	7.4	3.6	3.9	5.5	1.3	0.6
Not working/other	283	7.0	3.3	3.7	5.7	0.9	0.4
Ethnicity/race							
White, non-Hispanic	3,109	77.0	39.0	38.0	60.4	13.9	2.6
Black/African American, Non-Hispanic	370	9.2	4.6	4.6	5.2	2.9	1.1
Other, non-Hispanic	162	4.0	2.0	2.0	3.0	0.8	0.3
Hispanic	399	9.9	4.9	5.0	8.3	1.3	0.4
Household income							
Below \$35,000	1,247	31.3	15.5	15.9	10.3	18.6	2.5
\$35,000 or more	2,409	60.5	31.2	29.4	6.7	52.8	1.1
Refused/don't know	323	8.1	3.9	4.3	2.1	5.4	0.7
Education							
Less than high school	342	8.5	4.2	4.3	5.6	1.8	1.0
High school	1,221	30.2	15.2	15.1	22.0	6.7	1.6
Some college/other training	1,248	30.9	15.9	15.0	22.8	6.9	1.2
Bachelor's degree or higher	1,229	30.4	15.2	15.2	26.5	3.5	0.5

^aStudy Form A included the SF-36v2 standard (4-week recall) form, single-item format; SF-8 standard (4-week recall) form; MOS Sleep-R standard (4-week recall) form; condition checklist; and validation, healthcare, and background items.

^bStudy Form B included the SF-36v2 standard (4-week recall) form, item-grid format; SF-8 standard (4-week recall) form; MOS COG-R standard (4-week recall) form; PIQ-6 standard (4-week recall) form; condition checklist; and validation, health care, and background items.

from the 2009 SF-36v2 normative data did not significantly differ from those derived from the 1990 SF-36 or 1998 SF-36v2 normative data. Moreover, using the same component score coefficients provides continuity between the revised instrument and the earlier versions of the survey. As with the health domain scales, the only PCS and MCS scoring change concerns the centering of each measure on a mean of 50 with an *SD* of 10, based on the 2009 U.S. general population sample.

Comparison of Item Format Presentations

One of the first steps in the development of 2009 U.S. general population norms was to determine whether SF-36v2 standard form data gathered using two different item-presentation formats could be combined. To this end, SF-36v2 standard form data collected using the single-item presentation format (Form A) was compared to data collected using the item-grid presentation format (Form B). This comparison uncovered statistically

Table 14.6*Demographic Characteristics of the SF-36v2 Acute (1-Week Recall) Form, 2009 U.S. General Population (N = 2,061)*

Characteristic	n	Total %	Method of Internet access		
			PC %	Web TV %	LaptoP %
Gender					
Male	1,011	49.05	38.04	9.51	1.50
Female	1,050	50.95	38.77	9.02	3.15
Age					
18–24	163	7.91	7.33	0.19	0.39
25–34	286	13.88	12.95	0.49	0.44
35–44	354	17.18	14.56	1.84	0.78
45–54	337	16.35	11.11	4.17	1.07
55–64	378	18.34	12.32	5.05	0.97
65–74	378	18.34	13.88	3.78	0.68
75+	165	8.01	4.66	3.01	0.34
Marital status					
Married	1,097	53.23	46.09	5.63	1.50
Widowed	135	6.55	3.64	2.38	0.53
Divorced	256	12.42	6.89	4.85	0.68
Separated	36	1.75	1.02	0.34	0.39
Never married	423	20.52	14.94	4.51	1.07
Living with partner	114	5.53	4.22	0.82	0.49
Employment status					
Working	1,096	53.18	44.59	6.79	1.80
Retired/disabled	666	32.31	21.01	9.41	1.89
Temporarily laid off/looking for work	155	7.52	5.39	1.55	0.58
Not working/other	144	6.99	5.82	0.78	0.39
Ethnicity/race					
White, non-Hispanic	1,582	76.76	60.55	13.39	2.81
Black/African American, Non-Hispanic	197	9.56	5.68	2.72	1.16
Other, non-Hispanic	90	4.37	2.96	1.12	0.29
Hispanic	192	9.32	7.62	1.31	0.39
Household income					
Below \$35,000	675	33.30	19.88	10.85	2.57
\$35,000 or more	1,203	59.35	51.85	6.17	1.33
Refused/don't know	149	7.35	4.83	1.73	0.79
Education					
Less than high school	174	8.44	5.34	2.28	0.82
High school	634	30.76	22.61	6.16	1.99
Some college/other training	611	29.65	21.69	6.55	1.41
Bachelor's degree or higher	642	31.15	27.17	3.54	0.44

Note. This sample was administered Study Form C, which included the SF-36v2 acute (1-week recall) form, single-item format; SF-8 acute (1-week recall) form; MOS Sleep-R acute (1-week recall) form; condition checklist; and validation, health care, and background items.

significant differences in the observed data for the SF, RE, and MH health domain scales (see Table 14.7). However, because the difference in these three scales' scores for the differing item formats did not exceed 1.35 *T*-score points, the data from Form A and Form B were combined ($N = 4,040$) to become the basis for the 2009 SF-36v2 standard form norms, while the data collected using Form C served as the basis for the 2009 SF-36v2 acute form norms.

Finalization of the 2009 Normative Samples

To maximize the amount of usable data, the study investigators applied the *Full Missing Score Estimation (Full MSE)* method to data sets with missing item response values for the health domain scales. This method provides an effective and simple solution for dealing with missing data, even when only one item in a given scale is answered. When this occurs, Full MSE assumes that a scale's missing item response(s) would be the same

as the response to said scale's one answered item, or the average of all responses if more than one scale item was answered, and final item response value(s) would be assigned accordingly. More specifically, when a scale has one or more items missing a response and two or more items with responses, the final item response value assigned to the missing response(s) would be the average of the final item response values for the items with responses. However, because of the hierarchical nature of its items, this rule was not used to estimate Physical Functioning (PF) scale item responses in the 2009 norming study; instead, in cases where the PF scale was missing one or more item responses, an estimate was attained through a solution based on item response theory (IRT).

Furthermore, using regression methods, SF-36v2 component summary measure scores can be estimated on the basis of available health domain scale scores. Specifically, the Full MSE method allows the PCS score to be calculated when seven scale scores are available *and* the PF scale score is not missing. Similarly, this method also allows the MCS score to be calculated when at least seven scale scores are available *and* the MH scale score is not missing. Note that when calculating summary measure scores, QualityMetric's MSE software capability selects a unique scoring algorithm based on which particular health domain scale score is missing.

Component Summary Measure, Health Domain Scale, and SF-6D 2009 U.S. General Population Norms

The 2009 U.S. general population normative data for SF-36v2 component summary measures, health domain scales, and SF-6D are presented in Tables 14.8 and 14.9 for the standard and acute forms, respectively. The data in each of these tables include the *T*-score mean, median (50th percentile), 25th and 75th percen-

tiles, *SD*, observed range of scores, and the percentages of the norm group scoring the highest possible score (i.e., the ceiling) and the lowest possible score (i.e., the floor) for each scale. With the exception of SF-6D data, the normative data presented in Tables 14.8 and 14.9 are *T* scores. Note that with the exception of the VT scale, the median (50th percentile score) for each standard and acute scale, measure, and SF-6D was generally higher than its mean score. This reflects some skewness of the score distributions in the U.S. general population, with more respondents scoring above the mean than not.

Tables 14.10 and 14.11 present item mean raw scores, *SD*s, and correlations between items and health domain scales in the 2009 U.S. general population for the SF-36v2 standard and acute forms, respectively. The item-scale correlations that were corrected for overlap (i.e., an item's score was removed from its parent scale's score before the correlation was calculated) and that were hypothesized to be the highest in the same row are noted. Examination of Tables 14.10 and 14.11 reveals that within each scale, correlations between items and their hypothesized scales exceeded the .40 standard for internal consistency (Helmstader, 1964) for both the standard and acute forms. Also, with the exception of one standard form item (Item 3a), each item correlated higher with its parent scale than with any of the other scales. Moreover, for all scales except BP, item means and standard deviations were roughly equal across items within each scale, for both forms. The implication of these results is that items in each hypothesized scale contain approximately the same proportion of information about the health domain being measured. These results support the comparability of the standard and acute form scales, as well as the use of the summed ratings method as the first step in scoring the SF-36v2's scales.

Table 14.7

Comparison of SF-36v2 Standard (4-Week Recall) Form Mean Health Domain Scale T Scores Based on Single-Item and Item-Grid Presentation Formats, 2009 U.S. General Population

	Single-Item Format (<i>N</i> = 2,037)			Item-Grid Format (<i>N</i> = 2,003)			<i>t</i>	<i>p</i>
	<i>N</i>	Mean	<i>SD</i>	<i>N</i>	Mean	<i>SD</i>		
PF	2,036	50.09	10.06	1,998	49.91	9.94	0.56	.57
RP	2,036	50.06	10.00	1,991	49.94	10.01	0.35	.72
BP	2,032	50.18	10.22	1,995	49.82	9.77	1.13	.26
GH	2,036	49.91	10.09	2,000	50.10	9.91	-0.60	.55
VT	2,032	49.88	10.10	1,996	50.12	9.89	-0.76	.44
SF	2,032	50.36	10.06	1,997	49.64	9.93	2.29	.02
RE	2,033	50.42	9.60	1,993	49.58	10.38	2.66	.01
MH	2,031	50.67	9.90	1,997	49.32	10.06	4.29	< .0001

For the PF scale, items measuring easier physical tasks, such as bathing and dressing, were found to have lower standard deviations than items measuring more demanding tasks, such as vigorous activities or climbing several flights of stairs. However, additional analyses using IRT models have shown that these differences were caused by larger floor effects for the easy items in the samples analyzed and do not suggest that these simpler items contain more information than the difficult items (e.g., Haley, McHorney, & Ware, 1994). Thus, the summated ratings method is appropriate for the PF scale as well.

2009 Supplemental Norms and Benchmarks

As with previous Short Form norming studies conducted by QualityMetric Incorporated, supplemental sets of norms and benchmarks were developed using the data collected during the 2009 norming project. These age, gender, and disease-specific norms and benchmarks can provide important comparison information when inter-

preting results from individual respondents or groups of respondents (see Chapter 7).

Supplemental Norms for Age, Gender, and Gender-by-Age Groups

QualityMetric and its authorized resellers make available age, gender, and gender-by-age norms for the SF-36v2 standard and acute form health domain scales, component summary measures, and SF-6D. Seven different age groupings were used when developing these supplemental norms: 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, and 75 years or older. These age groups were selected (a) to be large enough to satisfy minimum standards for precision, (b) to correspond with standard practices for defining age-specific groups, and (c) to correspond with the age groupings used when reporting norms for the SF-36 and its translations (Apolone, Mosconi, & Ware, 1997; Bjorner et al., 1997; Fukuhara, Suzukamo, Bito, & Kurokawa, 2001; Jenkinson, Layte, Wright, & Coulter, 1996; Leplège, Ecosse, Pouchot, Coste, & Perneger, 2001; Sullivan, Karlsson, & Ware, 1994; Ware & Kosinski, 2001b; Ware, Kosinski, & Keller, 1994; Ware, Snow, Kosinski, & Gandek, 1993).

Table 14.8

SF-36v2 Standard (4-Week Recall) Form Normative Data, 2009 U.S. General Population

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH	SF-6D
Mean	50.01	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	0.74
25th %ile	45.04	45.24	46.06	45.93	42.64	43.68	43.69	47.31	45.72	43.02	0.64
50th %ile	53.07	52.94	53.71	54.91	51.51	50.81	49.63	57.34	56.17	53.48	0.75
75th %ile	57.24	57.13	57.54	57.16	55.55	57.94	58.54	57.34	56.17	58.72	0.85
SD	9.95	10.16	10.01	10.01	10.01	10.01	10.01	10.01	10.01	10.01	0.14
Minimum	7.32	5.79	19.26	21.23	21.68	18.95	22.89	17.23	14.39	11.63	0.30
Maximum	70.14	69.91	57.54	57.16	62.00	66.50	70.42	57.34	56.17	63.95	1.00
% Floor	N/A	N/A	0.8	2.4	1.1	0.2	1.1	1.2	1.3	0.2	0.1
% Ceiling	N/A	N/A	31.4	46.2	20.	3.8	2.0	54.2	61.2	4.9	2.2
N	4,024	4,024	4,034	4,027	4,027	4,036	4,028	4,029	4,026	4,028	3,856

Note. Except for the SF-6D, the norms presented are *T* scores.

Table 14.9

SF-36v2 Acute (1-Week Recall) Form Normative Data, 2009 U.S. General Population

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH	SF-6D
Mean	50.01	49.98	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	0.76
25th %ile	44.56	44.81	46.02	46.11	45.47	44.23	44.65	46.85	48.00	45.33	0.64
50th %ile	53.40	53.44	53.74	57.12	52.97	52.17	50.10	56.74	55.64	52.76	0.78
75th %ile	57.66	57.23	57.60	57.12	60.87	57.46	58.26	56.74	55.64	57.72	0.89
SD	10.35	10.45	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	0.14
Minimum	10.80	5.62	19.03	21.89	21.39	21.29	25.60	17.20	9.84	13.12	0.30
Maximum	75.51	69.65	57.60	57.12	60.87	65.40	69.15	56.74	55.64	62.67	1.00
% Floor	N/A	N/A	0.7	1.8	1.0	0.3	1.6	1.0	0.8	0.1	0.1
% Ceiling	N/A	N/A	32.1	49.8	26.4	4.1	2.6	59.9	67.5	7.1	3.4
N	2,056	2,056	2,059	2,057	2,056	2,061	2,057	2,057	2,057	2,060	1,997

Note. Except for the SF-6D, the norms presented are *T* scores.

Table 14.10

SF-36v2 Standard (4-Week Recall) Form Item Means, Standard Deviations, and Correlations With Health Domain Scales, 2009 U.S. General Population (N = 4,040)

Scale	Item	Mean	SD	PF	RP	BP	GH	VT	SF	RE	MH
PF	3a	2.12	0.79	.65 ^a	.66	.58	.52	.42	.43	.36	.24
	3b	2.61	0.65	.82 ^a	.74	.61	.52	.43	.52	.44	.29
	3c	2.72	0.56	.80 ^a	.71	.57	.47	.40	.53	.47	.31
	3d	2.45	0.73	.82 ^a	.70	.58	.54	.45	.48	.42	.27
	3e	2.72	0.58	.81 ^a	.66	.53	.46	.39	.49	.44	.28
	3f	2.46	0.69	.75 ^a	.66	.60	.46	.40	.45	.39	.25
	3g	2.46	0.77	.84 ^a	.73	.60	.53	.45	.50	.43	.28
	3h	2.67	0.64	.85 ^a	.71	.56	.49	.41	.51	.45	.28
	3i	2.74	0.56	.79 ^a	.65	.51	.43	.37	.47	.43	.27
	3j	2.88	0.39	.59 ^a	.50	.36	.32	.26	.41	.39	.24
	RP	4a	4.28	1.12	.77	.88 ^a	.65	.55	.50	.64	.59
4b		4.08	1.21	.76	.89 ^a	.67	.58	.56	.64	.59	.40
4c		4.15	1.22	.80	.92 ^a	.70	.59	.52	.64	.56	.37
4d		4.15	1.19	.80	.91 ^a	.70	.59	.54	.65	.58	.39
BP	7	4.37	1.27	.60	.63	.79 ^a	.55	.53	.52	.41	.37
	8	4.19	1.06	.69	.75	.79 ^a	.58	.56	.66	.53	.44
GH	1	3.37	0.92	.55	.54	.53	.71 ^a	.53	.49	.40	.38
	11a	4.22	1.04	.36	.39	.36	.47 ^a	.43	.43	.40	.43
	11b	3.56	1.17	.47	.50	.47	.70 ^a	.53	.47	.38	.42
	11c	3.47	1.12	.35	.36	.36	.47 ^a	.41	.29	.28	.31
	11d	3.31	1.24	.54	.56	.55	.76 ^a	.60	.51	.42	.45
VT	9a	3.45	1.04	.42	.48	.49	.57	.67 ^a	.56	.48	.64
	9e	3.12	1.03	.46	.50	.50	.59	.73 ^a	.53	.43	.57
	9g	3.50	1.01	.40	.46	.48	.53	.72 ^a	.53	.48	.58
	9i	3.16	0.98	.39	.45	.47	.51	.74 ^a	.50	.45	.54
SF	6	4.34	1.02	.55	.64	.58	.53	.57	.73 ^a	.66	.60
	10	4.28	1.08	.54	.62	.57	.53	.60	.73 ^a	.64	.62
RE	5a	4.45	0.98	.51	.61	.48	.48	.52	.68	.88 ^a	.62
	5b	4.36	1.03	.47	.57	.46	.47	.54	.67	.89 ^a	.65
	5c	4.56	0.87	.46	.54	.43	.42	.47	.62	.83 ^a	.58
MH	9b	4.19	0.95	.23	.30	.30	.37	.46	.48	.51	.63 ^a
	9c	4.46	0.86	.31	.38	.36	.41	.51	.59	.62	.73 ^a
	9d	3.46	0.96	.24	.30	.36	.44	.62	.49	.44	.65 ^a
	9f	4.24	0.93	.29	.36	.35	.44	.59	.60	.64	.77 ^a
	9h	3.65	0.89	.26	.33	.35	.45	.60	.51	.46	.67 ^a

^aItem-scale correlation corrected for overlap (relevant item removed from its scale for correlation) and hypothesized to be highest in same row.

These supplemental norms are useful in determining whether a score for a male or a female is above or below the average score for males or females, respectively, in a particular age group of interest in the U.S. general population.

Comparing results across the age groups clearly shows that health status, particularly physical health, is related to age. For example, whereas the mean standard form PF score for the total sample was 50.00, the mean for the 18- to 24-year-old group was higher (54.20) and the mean for the 75 years or older group was lower (41.97). The opposite is generally true for the mental health scales and measure, with the older age groups generally earning higher mean MH and MCS scores

than some of the younger age groups, particularly after the age of 54 years. For example, the mean MCS and MH scores for the 18- to 24-year-old group were 48.00 and 50.09, respectively; for the 75 years or older group, the mean scores for MCS and MH were 53.49 and 52.58, respectively. Finally, normative data indicated that SF-6D mean scores generally decline with increasing age.

Supplemental Benchmarks for Disease-Specific Populations

As previously described, the 2009 U.S. general population normative data collection effort included a checklist of 40 chronic conditions. This checklist asked respondents to indicate whether a doctor or other health

Table 14.11

SF-36v2 Acute (1-Week Recall) Form Items, Standard Deviations, and Correlations With Health Domain Scales, 2009 U.S. General Population (N = 2,061)

Scale	Item	Mean	SD	PF	RP	BP	GH	VT	SF	RE	MH
PF	3a	2.15	0.77	.69 ^a	.66	.59	.59	.45	.40	.31	.25
	3b	2.61	0.65	.82 ^a	.77	.62	.57	.48	.51	.40	.30
	3c	2.74	0.55	.78 ^a	.73	.59	.51	.43	.50	.41	.31
	3d	2.47	0.72	.81 ^a	.69	.56	.56	.47	.44	.32	.27
	3e	2.74	0.56	.82 ^a	.69	.54	.48	.41	.48	.36	.26
	3f	2.50	0.68	.73 ^a	.67	.59	.49	.41	.40	.32	.26
	3g	2.48	0.75	.82 ^a	.70	.58	.55	.45	.45	.33	.26
	3h	2.71	0.60	.83 ^a	.69	.52	.51	.41	.47	.37	.25
	3i	2.76	0.55	.78 ^a	.66	.50	.45	.38	.46	.35	.24
	3j	2.91	0.33	.57 ^a	.52	.40	.34	.31	.42	.33	.24
	RP	4a	4.32	1.06	.77	.87 ^a	.64	.58	.50	.57	.48
4b		4.18	1.15	.79	.92 ^a	.68	.63	.56	.59	.47	.37
4c		4.22	1.17	.81	.92 ^a	.69	.63	.53	.59	.48	.36
4d		4.22	1.16	.81	.92 ^a	.71	.63	.54	.60	.48	.36
BP	7	4.53	1.28	.62	.63	.79 ^a	.58	.56	.51	.38	.39
	8	5.30	1.01	.68	.74	.79 ^a	.57	.55	.61	.45	.42
GH	1	3.36	0.94	.59	.58	.53	.74 ^a	.57	.46	.36	.39
	11a	4.25	1.03	.38	.44	.41	.51 ^a	.47	.45	.37	.39
	11b	3.59	1.17	.56	.58	.52	.75 ^a	.61	.50	.39	.45
	11c	3.44	1.14	.36	.37	.37	.51 ^a	.41	.30	.24	.32
	11d	3.32	1.25	.60	.61	.56	.81 ^a	.63	.53	.40	.45
VT	9a	3.41	1.11	.45	.48	.50	.59	.70 ^a	.57	.46	.67
	9e	3.10	1.07	.51	.53	.52	.65	.76 ^a	.55	.44	.63
	9g	3.50	1.02	.42	.46	.51	.53	.70 ^a	.54	.44	.56
	9i	3.21	0.98	.39	.44	.47	.55	.75 ^a	.52	.43	.56
SF	6	4.41	1.00	.49	.55	.52	.50	.55	.71 ^a	.63	.61
	10	4.39	1.02	.53	.60	.56	.55	.63	.71 ^a	.64	.66
RE	5a	4.54	0.89	.43	.52	.43	.44	.50	.65	.87 ^a	.64
	5b	4.53	0.88	.40	.46	.41	.43	.52	.66	.91 ^a	.69
	5c	4.67	0.77	.36	.44	.38	.38	.45	.61	.82 ^a	.61
MH	9b	4.34	0.91	.23	.28	.31	.34	.46	.49	.53	.61 ^a
	9c	4.56	0.82	.30	.36	.38	.40	.54	.65	.65	.74 ^a
	9d	3.48	0.99	.26	.30	.37	.46	.67	.52	.49	.71 ^a
	9f	4.34	0.91	.29	.33	.37	.42	.59	.63	.68	.79 ^a
	9h	3.61	0.96	.25	.29	.34	.45	.66	.54	.49	.72 ^a

^aItem-scale correlation corrected for overlap (relevant item removed from its scale for correlation) and hypothesized to be highest in same row.

professional had *ever* told them that they had any of 26 conditions or if they *currently* had any of 14 conditions. This information enabled the development of specific sets of benchmarks for each of the conditions and disease states found on the condition checklist. SF-36v2 standard and acute form disease- and condition-specific benchmarks, unadjusted for differences in sociodemographic characteristics and comorbid conditions, are available from QualityMetric and its authorized resellers. As previously indicated, these supplemental benchmarks can provide important comparison information when interpreting results from individual respondents or groups of respondents (see Chapter 7).

Additional information regarding the utility of the SF-36v2 disease-specific benchmarks can be found in the percentage of respondents in each disease group who earned the highest possible score (i.e., the ceiling) and the percentage who earned the lowest possible score (i.e., the floor) for each scale. Table 14.12 presents the percentages of respondents in each of the 40 disease groups scoring at the floor and at the ceiling of each standard form health domain scale. This table also includes SF-36v2 data for a “healthy” subgroup of respondents; that is, those from the 2009 U.S. general population sample who reported never having been told they had any of 18 physical conditions or an alcohol or

drug use disorder and were not currently experiencing anxiety or depression. Likewise, Table 14.13 provides the disease groups' ceiling and floor percentages for each of the acute form health domain scales. With very few exceptions (most notably in the RP scale), there appeared to be relatively little floor effect across the eight standard form health domain scales. However, significant ceiling effects were noted for RP, SF, RE, and, to a somewhat lesser degree, PF and BP (see Table 14.12). Note that findings were quite similar for the SF-36v2 the acute form (see Table 14.13). With some exceptions, these findings are generally comparable to those reported for the SF-36v2 general population normative sample (see Tables 14.8 and 14.9) and thus reflect limitations in the measurement of functioning at the upper end of certain scales.

Further information regarding study participants who completed the SF-36v2 standard or acute form and indicated the presence of one or more chronic conditions is presented in Tables 14.14 and 14.15, respectively. Included in both of these tables are the three most prevalent comorbid conditions for each disease group's members and the percentage of the disease group reporting each of its three most-indicated comorbidities. For a given chronic condition, any of the other 39 conditions on the checklist that were indicated by the respondent to be present was considered a comorbidity, regardless of whether it was one that the respondent had *ever* had or *currently* had. Examination of Tables 14.14 and 14.15 reveals both expected and unexpected findings with regard to the conditions' most common comorbidities. For example, in Table 14.14, the most common comorbidity for seasonal allergies was nasal allergies or rhinitis (69.87%); however, for chronic lung disease other than asthma, the most common comorbidity was arthritis of any kind or rheumatism (64.80%), followed by COPD (60.71%). Interestingly, across the standard and acute forms, 32 (80%) of the 40 conditions share the same most common comorbidity.

SF-36v2 score reports incorporating 2009 age, gender, and disease-specific normative information are available through the scoring services offered by QualityMetric and its authorized resellers.

Comparability of 2009 and 1998 Normative Data: Preliminary Findings

Updating the norms of an established survey instrument commonly elicits legitimate questions regarding the comparability of findings derived from the previous

norm set to the those obtained using the updated norms. Such concerns generally center around the appropriateness and/or practicality of (a) making comparisons using data collected before and after the publication of the updated norms, (b) combining these two sets of data into one data set for research or reporting purposes, and (c) applying interpretive guidelines based on the previous norm set to data obtained using the updated norm set.

To address the comparability of the 1998 and 2009 SF-36v2 norms, data from the 1998 and 2009 normative studies underwent three very basic comparisons. Note that the results of these comparisons (briefly discussed in the following sections) should be considered preliminary and that the issue of comparability will be more fully addressed in studies that will be undertaken in the future.

Comparison of 2009 and 1998 Mean Item Raw Scores

Table 14.16 presents SF-36v2 standard and acute form mean item raw scores and *SDs* for both the 2009 and 1998 U.S. general population normative samples. Standard form mean item raw score differences (2009 score minus 1998 score) ranged from -0.26 (Item 11d) to 0.06 (Item 5c). For the acute form, differences ranged from -0.23 (Item 11d) to 1.03 (Item 8). In general, the difference in mean item raw scores is no more than one quarter of a raw-score point. The exception is the difference in mean raw scores for acute form Item 8 (extent to which pain interfered with normal work). This is approximately equivalent to a one response-category change and may be considered clinically significant.

Comparison of Mean Health Domain Scale *T* Scores

As part of the investigation into comparability, regression analyses were conducted to investigate the differences in SF-36v2 normative scale and summary measure *T* scores between the 1998 and 2009 normative samples. A multivariate regression analysis was conducted for each SF-36v2 scale and summary measure that included age, gender, and 18 self-reported chronic conditions as independent variables in the model. The purpose of these analyses was to adjust for as many differences in sample characteristics as possible, but most notably for the differences in prevalence of chronic conditions, which presumably have the greatest impact on scores. For these analyses the 0-100 scores for each of the 8 SF-36v2 health domain scales were used, since norm-based scoring methods differ for the two samples.

Table 14.12

Percentage Scoring at the Floor and Ceiling of Each SF-36v2 Standard (4-Week Recall) Form Health Domain Scale by Self-Reported Disease Group, 2009 U.S. General Population

	PF		RP		BP		GH		VT		SF		RE		MH	
	%F ^a	%C ^b	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C
Healthy	<1.0	51.2	<1.0	68.7	<1.0	32.2	<1.0	7.5	<1.0	<1.0	3.9	<1.0	71.7	<1.0	76.7	<1.0
Hypertension	1.4	13.9	4.0	30.1	1.8	12.3	0.3	1.2	1.6	1.3	1.3	1.7	47.7	2.2	55.6	0.3
Heart attack	8.2	6.1	12.2	12.2	4.1	14.3	0.0	0.0	2.0	0.0	0.0	4.1	20.4	6.1	38.8	0.0
Congestive heart failure	2.9	2.9	7.4	9.6	2.2	6.6	0.7	0.0	2.9	0.7	0.7	3.6	24.8	2.2	37.5	0.0
Angina	4.1	4.7	7.6	12.3	3.5	9.4	1.8	0.0	1.8	0.0	0.0	1.8	34.5	2.9	49.1	0.0
Other heart condition	1.5	10.3	4.5	22.2	1.5	9.5	0.6	0.9	2.8	0.4	0.4	2.2	42.0	1.3	51.1	0.4
Diabetes	2.1	10.7	5.2	23.7	3.6	7.3	1.0	0.2	3.1	0.7	0.7	2.6	38.5	2.4	47.8	0.5
Cancer	2.3	8.0	6.4	22.5	2.3	13.2	1.3	1.9	2.3	0.3	0.3	2.6	43.7	3.5	51.8	0.6
COPD	3.4	2.8	11.3	11.3	4.5	5.1	2.2	0.0	3.4	0.0	0.0	3.4	23.7	5.1	41.2	1.1
Allergies	0.9	23.7	3.1	37.9	1.6	13.8	0.3	2.4	1.8	0.9	0.9	1.8	48.2	1.6	55.3	0.2
Rheumatoid arthritis	1.6	4.8	6.7	14.7	4.1	4.1	0.3	1.0	2.2	0.3	0.3	4.1	29.5	4.1	36.2	0.3
Osteoarthritis	2.2	5.5	7.2	17.5	3.8	1.7	0.7	0.5	2.7	0.2	0.2	2.4	34.5	2.9	48.9	0.2
Osteoporosis	1.6	8.7	5.1	18.5	2.4	5.9	0.4	1.2	1.6	0.0	0.0	2.4	35.3	2.4	45.7	0.0
Kidney disease	6.5	6.5	9.7	12.9	2.2	9.7	1.1	0.0	1.1	0.0	0.0	3.2	30.9	3.2	41.9	1.1
Liver disease	5.6	9.0	10.1	18.0	7.9	11.2	2.2	0.0	3.4	0.0	0.0	6.7	33.7	6.7	33.7	2.2
GERD	1.5	11.6	3.5	24.5	1.7	8.3	0.3	0.9	2.4	0.6	0.6	1.7	41.3	2.6	48.8	0.2
Stomach disease	1.9	12.0	5.7	18.4	6.3	10.1	1.9	0.6	2.5	0.0	0.0	4.4	25.3	5.7	37.3	0.6
IBS	0.6	15.3	5.6	26.9	2.5	9.3	0.3	0.6	2.8	0.6	0.6	3.1	39.3	3.4	45.8	0.3
Obesity	1.6	11.5	4.3	27.1	2.5	6.8	0.4	0.4	2.7	0.1	0.1	2.5	35.4	2.6	41.8	0.3
Stroke	5.7	4.1	13.0	8.9	3.3	9.8	0.0	0.8	2.4	0.0	0.0	0.8	22.8	4.1	35.8	0.0
HIV/AIDS	5.3	0.0	10.5	15.8	5.3	5.3	5.3	0.0	0.0	0.0	0.0	0.0	15.8	5.3	26.3	0.0
Anemia	1.2	17.8	4.2	30.1	1.2	10.5	0.2	0.9	1.9	0.7	0.7	2.1	38.3	1.6	47.9	0.2
Clinical depression	1.4	15.0	5.9	23.4	2.8	8.3	1.0	0.2	3.9	0.0	0.0	5.3	20.3	5.8	22.4	1.0
Alcohol/drug use	1.9	15.5	3.9	21.4	2.9	10.7	1.0	1.0	1.9	0.0	0.0	1.9	21.4	2.9	23.3	1.0
Chronic fatigue	3.6	3.6	11.3	10.1	7.7	1.2	2.4	1.2	8.3	0.0	0.0	7.1	16.6	6.5	30.4	1.2
Migraine	1.6	22.1	5.4	33.3	2.3	13.8	0.5	2.3	3.0	1.4	1.4	3.1	39.0	4.4	46.7	0.5
Sleep apnea	3.2	10.2	8.0	20.0	4.4	7.8	1.0	0.0	3.2	0.2	0.2	3.6	33.3	4.9	42.3	0.5
Chronic allergies	1.2	16.6	4.3	29.0	2.8	12.0	0.3	1.4	1.7	0.9	0.9	2.3	40.4	2.2	48.3	0.3
Seasonal allergies	0.9	27.8	2.4	43.0	1.2	15.2	0.2	3.2	1.3	1.1	1.1	1.4	52.1	1.3	58.2	0.1
Back problems	1.9	7.0	6.2	17.3	3.7	1.7	0.4	1.6	2.9	0.0	0.0	2.9	30.3	2.9	44.2	0.3
Trouble seeing	1.9	10.8	5.1	19.7	3.4	9.1	0.2	0.4	2.5	0.6	0.6	3.6	29.0	3.2	39.1	0.4
Trouble hearing	1.8	11.5	3.3	23.9	2.7	9.8	0.4	0.8	2.3	0.8	0.8	2.5	43.6	1.6	50.0	0.4
Any arthritis	1.8	6.8	5.1	19.7	2.3	4.0	0.4	1.3	2.0	0.5	0.5	1.8	38.8	2.3	50.4	0.2
Skin conditions	1.5	16.8	4.0	30.2	1.7	10.9	0.0	2.0	0.7	0.5	0.5	1.2	45.5	2.2	51.2	0.2
Asthma	2.3	19.2	7.0	28.1	2.6	13.7	0.6	1.5	1.7	0.6	0.6	2.6	36.7	3.5	47.4	0.0
Lung condition other than asthma	4.5	3.0	10.7	12.7	4.6	6.1	2.0	0.0	1.5	0.0	0.0	3.6	24.4	4.6	37.6	1.0
Ulcer	3.1	13.1	9.2	20.8	6.9	7.7	0.8	0.8	2.3	0.8	0.8	3.1	27.7	6.9	33.8	0.0
Depression	2.3	13.1	6.3	21.9	3.2	8.7	0.8	0.0	4.3	0.0	0.0	5.5	13.8	6.2	16.6	1.0
Anxiety	1.9	16.2	6.1	24.6	2.9	8.3	1.0	0.5	3.8	0.0	0.0	5.1	17.0	5.3	21.4	1.0
Severe headaches	2.0	15.7	7.4	25.4	4.7	9.3	0.7	1.2	4.7	1.2	1.2	5.7	23.6	5.7	35.2	0.7
Limited use of arm/leg	4.8	1.9	12.5	7.7	5.8	2.4	0.8	0.0	3.7	0.3	0.3	4.0	23.1	5.6	37.2	0.8

Note. Estimates for PCS and MCS are not provided because the ceiling and floor effects for the summary measures are not significant.
^aPercentage of respondents scoring at the floor (F), the lowest possible score on each scale, rounded to one decimal point.
^bPercentage of respondents scoring at the ceiling (C), the highest possible score on each scale, rounded to one decimal point.

Score differences were later converted to the *T*-score metric. Given that the component summary measures only use norm-based scoring methods, the comparison of results of the analyses are somewhat compromised and will not be reported here. Separate sets of analyses were conducted for the standard form and the acute form.

Subtracting each standard form scale mean 2009 score from its mean 1998 score revealed differences in scores between normative samples that ranged from -0.28 to 0.66 *T*-score points (see Table 14.17). With the exception of the GH scale (-0.28 points), all scores were higher for the 2009 normative sample. Although the results for each scale (except the GH scale) were statistically significant, the magnitude of the mean score differences are not clinically meaningful as all differences between the 2009 and 1998 SF-36v2 samples are less than 1 *T*-score point. Also, all differences of these magnitudes are less than minimally important difference (MID) scores established for the SF-36v2 health domain scales (see Chapter 10).

Results of analyses of the acute form found differences in mean scores between samples that ranged from 0.66 to 2.15 *T*-score points, with all scores being higher for the 2009 normative sample (see Table 14.18). All differences in mean scale scores between 2009 and 1998 samples are statistically significant; however, these differences are less than 2 *T*-score points for the majority of the scales, and roughly 2 points for the RE scale. All differences of these magnitudes are less than the MID scores established for the SF-36v2 scales.

Comparison of Mean Health Domain *T* Scores in the 2009 U.S. General Population, Scored Using 2009 and 1998 Scoring Algorithms

Another approach to examining the comparability of the 1998 and 2009 SF-36v2 norms is to compare the 2009 health domain scale normative data scored using the 2009 scoring algorithms with the same data scored using the 1998 scoring algorithms. The results of these comparisons for the standard and acute forms are presented in Tables 14.19 and 14.20, respectively. For both SF-36v2 forms, when the 2009 normative data are scored using the 2009 scoring algorithms, the mean *T* score and *SD* for each scale was 50 and 10, respectively.

As demonstrated by the standard form results (Table 14.19), when the 2009 normative data were scored using the 1998 scoring algorithms, all 1998 norms-based *T* scores were found to be significantly lower than the 2009 norms-based *T* scores. Again, this is likely due

to the large sample size. The greatest standard form difference (2.64 *T*-score points, or 0.26 *SD* units) was found for the GH scale. Similarly, all 1998 norms-based acute form *T* scores (Table 14.20) were lower than the 2009 norms-based *T* scores, with the exception of the RE scale. Nonsignificant differences were observed for only the acute form RE and BP scales, with the greatest *T*-score difference (2.68 *T*-score points, or 0.27 *SD* units) again being found for the GH scale. Note that the GH *T*-score differences for both the standard and acute forms exceed the MID recommended in Chapter 10.

Summary, Conclusions, and Recommendations

Several sets of analyses were conducted to investigate the comparability of SF-36v2 results derived from the application of its 1998 and 2009 norms-based scoring algorithms. These included comparisons of (a) mean item raw scores for the 1998 and 2009 normative samples, (b) scoring of 2009 normative raw data using 2009 and 1998 algorithms, and (c) mean *T* scores for the 1998 and 2009 normative samples, adjusted for as many differences that may have an impact on scores in sample characteristics. Partially due to large sample sizes, several *statistically significant* differences on SF-36v2 variables were found between the 2009 and 1998 norms-based data. With very few exceptions, however, findings from these analyses conducted to investigate differences in norms between the 1998 and 2009 SF-36v2 normative samples found no *clinically meaningful* differences in normative scores for the SF-36v2. Based on these findings, the same normal or average *T*-score interpretive ranges recommended for 1998 norms-based individual respondent data (*T* score = 45–55) and group-level data (*T* score = 47–53) continues to be recommended for interpreting 2009 norms-based data (see Chapter 7).

Despite these findings, the differences in the centering of scores using 1998 and 2009 norm-based methods will result in differences in scores that could range from 1 to 3 points. Thus, users who have relied on SF-36v2 1998 norm-based *T* scores will have to “re-center” (i.e., re-score) their scale and summary measure scores using 2009 scoring algorithms to validly compare their findings to norms and disease-specific benchmarks from the 2009 normative study. Similarly, those wanting to combine SF-36v2 data based on 1998 scoring algorithms with those scored using the 2009 algorithms for research or other purposes must first re-score the 1998 norms-based data using the 2009 algorithms before merging data sets.

Table 14.13

Percentage Scoring at the Floor and Ceiling of Each SF-36v2 Acute (1-Week Recall) Form Health Domain Scale by Self-Reported Disease Group, 2009 U.S. General Population

	PF		RP		BP		GH		VT		SF		RE		MH	
	%F ^a	%C ^b	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C
Healthy	0.1	54.9	0.1	72.5	0.1	42.7	0.0	8.4	0.1	4.8	0.0	76.3	0.0	81.6	0.0	10.1
Hypertension	1.3	16.3	2.6	33.6	1.4	16.7	0.4	0.9	2.6	1.5	1.3	54.5	0.6	64.6	0.0	6.6
Heart attack	0.0	10.0	7.5	22.5	2.5	20.0	0.0	2.5	0.0	2.5	0.0	42.5	5.0	52.5	0.0	7.5
Congestive heart failure	3.8	7.6	10.1	16.5	2.5	17.7	2.5	0.0	7.6	0.0	3.8	39.2	5.1	60.8	0.0	3.8
Angina	4.2	6.3	7.4	15.8	5.3	11.6	1.1	0.0	4.2	0.0	5.3	48.4	0.0	62.1	0.0	7.4
Other heart condition	2.1	14.6	6.7	25.8	2.9	13.8	0.4	1.7	4.2	0.4	3.3	39.6	1.7	53.3	0.0	3.3
Diabetes	2.6	11.2	6.3	29.6	2.3	11.5	0.7	0.3	3.3	1.3	2.0	47.0	2.0	59.2	0.0	6.3
Cancer	3.2	8.2	5.1	22.2	1.9	8.9	0.0	0.6	1.3	1.3	1.3	50.6	1.9	67.7	0.0	7.0
COPD	4.1	2.0	8.2	10.2	2.0	13.3	2.0	0.0	4.1	1.0	3.1	35.7	2.0	49.0	0.0	4.1
Allergies	0.4	25.5	1.6	42.3	1.3	19.3	0.3	2.3	2.4	1.0	1.0	51.6	0.6	61.1	0.1	3.5
Rheumatoid arthritis	2.2	5.5	6.6	20.3	3.3	6.0	1.1	1.1	4.4	1.1	4.4	36.8	3.8	48.9	0.5	7.7
Osteoarthritis	2.1	5.7	5.4	20.5	2.7	3.9	0.6	0.3	3.6	0.3	3.6	44.1	2.1	59.2	0.6	3.9
Osteoporosis	2.7	6.8	6.8	21.9	3.4	10.3	1.4	0.7	4.1	0.7	2.7	39.7	2.1	52.1	0.7	4.8
Kidney disease	2.9	7.4	7.4	16.2	0.0	10.3	0.0	0.0	4.4	0.0	2.9	36.8	1.5	47.1	0.0	1.5
Liver disease	2.1	10.6	10.6	17.0	0.0	8.5	0.0	0.0	2.1	2.1	0.0	27.7	2.1	36.2	0.0	4.3
GERD	1.2	14.8	4.3	25.8	2.5	12.0	0.3	0.9	3.4	0.3	2.8	42.8	2.2	56.3	0.3	3.7
Stomach disease	1.2	10.7	6.0	20.2	3.6	6.0	1.2	0.0	6.0	0.0	4.8	32.1	2.4	42.9	0.0	2.4
IBS	1.1	14.7	3.8	27.2	2.2	9.2	0.5	2.2	4.3	0.0	2.7	38.6	1.6	46.2	0.5	2.2
Obesity	1.7	10.8	4.2	26.9	2.5	11.1	0.6	0.3	5.0	0.8	3.1	37.8	1.7	51.1	0.3	3.9
Stroke	1.7	1.7	10.3	15.5	3.4	10.3	0.0	0.0	0.0	0.0	1.7	39.7	1.7	69.0	0.0	5.2
HIV/AIDS	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Anemia	2.1	18.5	2.1	32.2	1.7	15.0	0.0	0.9	2.6	0.4	3.9	42.5	1.3	54.5	0.0	2.6
Clinical depression	2.4	12.7	5.2	24.6	3.2	10.7	0.4	0.8	5.2	0.0	5.2	21.8	3.6	23.0	0.4	1.2
Alcohol/drug use	3.6	16.1	3.6	25.0	1.8	14.3	1.8	3.6	0.0	0.0	3.6	32.1	3.6	42.9	0.0	1.8
Chronic fatigue	7.4	3.7	12.3	9.9	4.9	2.5	1.2	0.0	8.6	0.0	8.6	13.6	3.7	40.7	1.2	2.5
Migraine	0.6	18.9	2.8	34.0	1.6	14.5	0.3	1.6	3.8	0.9	2.5	41.2	1.3	54.1	0.3	1.9
Sleep apnea	2.3	9.5	4.5	25.2	1.4	9.0	0.9	0.0	3.6	0.5	3.6	39.2	1.8	51.8	0.0	3.2
Chronic allergies	1.1	20.9	2.1	37.0	1.3	15.0	0.6	2.6	3.0	0.9	1.9	44.7	1.5	56.6	0.2	3.8
Seasonal allergies	0.6	29.1	1.5	47.1	1.1	21.6	0.4	2.1	1.7	1.2	1.2	55.0	0.7	64.4	0.1	4.5
Back problems	1.7	6.4	4.9	20.1	2.1	4.9	0.6	0.6	3.6	0.6	2.6	37.8	1.3	52.5	0.2	2.6
Trouble seeing	3.2	10.8	8.4	22.5	4.4	12.5	1.2	1.6	4.4	0.4	3.6	35.7	3.2	47.6	0.4	4.4
Trouble hearing	1.1	13.8	1.1	31.3	0.4	13.8	1.1	0.7	2.2	1.8	0.7	54.2	0.7	63.9	0.0	6.9
Any arthritis	1.8	8.4	4.3	23.3	2.1	6.3	0.6	1.0	3.4	0.6	2.7	46.1	1.5	60.2	0.3	4.7
Skin conditions	2.0	19.0	3.9	36.6	1.0	14.1	0.5	2.0	3.4	1.0	2.0	44.4	1.0	55.9	0.5	2.4
Asthma	1.0	20.7	3.6	34.2	1.0	16.6	0.5	1.0	5.7	0.5	2.1	39.4	2.1	53.4	0.5	3.1
Lung condition other than asthma	4.2	4.2	7.6	8.4	3.4	10.9	1.7	0.8	4.2	0.0	3.4	26.1	1.7	49.6	0.0	0.8
Ulcer	1.7	10.2	10.2	13.6	1.7	6.8	0.0	0.0	3.4	0.0	3.4	28.8	6.8	39.0	1.7	5.1
Depression	2.8	14.1	4.3	25.4	3.1	8.9	0.6	0.9	5.8	0.0	4.9	18.3	3.4	16.8	0.3	0.3

Table 14.13 (continued)

Percentage Scoring at the Floor and Ceiling of Each SF-36v2 Acute (1-Week Recall) Form Health Domain Scale by Self-Reported Disease Group, 2009 U.S. General Population

	PF		RP		BP		GH		VT		SF		RE		MH	
	%F ^a	%C ^b	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C	%F	%C
Anxiety	1.7	14.9	3.7	29.9	2.5	10.5	0.6	0.0	5.4	0.3	3.9	21.4	2.8	22.3	0.3	0.3
Severe headaches	1.2	18.8	4.5	30.2	2.8	13.8	1.2	0.8	5.7	0.4	4.5	31.4	2.9	44.5	0.8	0.8
Limited use of arm/leg	4.6	2.5	10.7	9.6	4.5	5.6	1.5	0.5	4.1	0.5	4.6	28.9	3.6	49.5	0.0	5.6

Note. Estimates for PCS and MCS are not provided because the ceiling and floor effects for the summary measures are not significant.

^aPercentage of respondents scoring at the floor (F), the lowest possible score on each scale, rounded to one decimal point.

^bPercentage of respondents scoring at the ceiling (C), the highest possible score on each scale, rounded to one decimal point.

Table 14.14
Characteristics of 2009 SF-36v2 Standard (4-Week Recall) Form Disease-Specific Benchmark Samples

	n	Mean Age	% Male	% Female	Most Common Comorbid Conditions		
					1st	2nd	3rd
Hypertension	1,518	60.40	53.36	46.64	Arthritis of any kind or rheumatism (45.65%)	Seasonal allergies (40.91%)	Nasal allergies or rhinitis (38.47%)
Myocardial infarction (within last year)	49	57.45	65.31	34.69	Hypertension (75.51%)	Arthritis of any kind or rheumatism (57.14%)	Angina (53.06%)
Congestive heart failure	136	62.75	56.62	43.38	Hypertension (76.47%)	Other heart conditions (56.62%)	Arthritis of any kind or rheumatism (55.88%)
Angina	171	65.57	67.25	32.75	Hypertension (80.12%)	Arthritis of any kind or rheumatism (61.40%)	Other heart conditions (45.61%)
Other heart conditions	464	60.30	54.09	45.91	Hypertension (60.34%)	Arthritis of any kind or rheumatism (51.51%)	Seasonal allergies (46.77%)
Diabetes	578	59.52	55.71	44.29	Hypertension (74.74%)	Arthritis of any kind or rheumatism (50.17%)	Nasal allergies or rhinitis (41.18%)
Cancer (not skin cancer)	311	65.04	51.45	48.55	Hypertension (57.23%)	Arthritis of any kind or rheumatism (51.77%)	Seasonal allergies (34.73%)
COPD	176	62.59	46.02	53.98	Arthritis of any kind or rheumatism (67.61%)	Chronic lung disease other than asthma (67.61%)	Hypertension (63.64%)
Nasal allergies or rhinitis	1,352	50.97	42.23	57.77	Seasonal allergies (83.36%)	Chronic allergies or sinus trouble (46.52%)	Hypertension (43.20%)
Rheumatoid arthritis	314	59.90	47.45	52.55	Arthritis of any kind or rheumatism (90.76%)	Hypertension (67.52%)	Nasal allergies or rhinitis (46.82%)
Osteoarthritis	582	62.11	35.74	64.26	Arthritis of any kind or rheumatism (93.13%)	Hypertension (62.20%)	Chronic back problems or sciatica (56.87%)
Osteoporosis	254	65.23	17.32	82.68	Arthritis of any kind or rheumatism (65.35%)	Hypertension (56.30%)	Chronic back problems or sciatica (44.49%)
Kidney disease	93	59.52	55.91	44.09	Hypertension (77.42%)	Arthritis of any kind or rheumatism (58.06%)	Diabetes (46.24%)
Liver disease	89	56.54	65.17	34.83	Hypertension (49.44%)	Arthritis of any kind or rheumatism (48.31%)	Chronic back problems or sciatica (42.70%)
GERD	664	58.81	43.52	56.48	Hypertension (58.43%)	Arthritis of any kind or rheumatism (55.72%)	Seasonal allergies (53.46%)
Stomach disease	158	55.82	40.51	59.49	GERD (56.33%)	Nasal allergies or rhinitis (55.06%)	Arthritis of any kind or rheumatism (54.43%)
Chronic bowel disease	321	53.54	33.96	66.04	Nasal allergies or rhinitis (58.26%)	Seasonal allergies (52.65%)	Hypertension (48.91%)
Obesity	704	51.65	37.93	62.07	Hypertension (57.24%)	Seasonal allergies (48.01%)	Nasal allergies or rhinitis (46.02%)
Stroke	123	64.39	52.85	47.15	Hypertension (80.49%)	Arthritis of any kind or rheumatism (59.35%)	Chronic back problems or sciatica (40.65%)
HIV or AIDS	19	54.00	78.95	21.05	Seasonal allergies (68.42%)	Hypertension (52.63%)	Nasal allergies or rhinitis (52.63%)
Anemia	428	52.18	19.63	80.37	Seasonal allergies (53.50%)	Nasal allergies or rhinitis (50.23%)	Hypertension (43.46%)
Clinical depression	505	48.35	34.26	65.74	Depression (76.83%)	Anxiety (59.80%)	Seasonal allergies (54.65%)

Table 14.14 (continued)
Characteristics of 2009 SF-36v2 Standard (4-Week Recall) Form Disease-Specific Benchmark Samples

	<i>n</i>	Mean Age	% Male	% Female	Most Common Comorbid Conditions		
					1st	2nd	3rd
Alcohol- or drug-use disorder	103	49.36	74.76	25.24	Depression (60.19%)	Anxiety (58.25%)	Clinical depression (53.40%)
Chronic fatigue syndrome or fibromyalgia	168	54.11	19.64	80.36	Arthritis of any kind or rheumatism (69.64%)	Nasal allergies or rhinitis (64.29%)	Seasonal allergies (57.74%)
Migraine headaches	574	46.72	25.09	74.91	Seasonal allergies (56.79%)	Nasal allergies or rhinitis (53.66%)	Severe headaches (46.69%)
Sleep apnea	411	56.35	61.07	38.93	Hypertension (61.31%)	Arthritis of any kind or rheumatism (52.80%)	Nasal allergies or rhinitis (48.18%)
Chronic allergies or sinus trouble	858	52.22	41.61	58.39	Nasal allergies or rhinitis (73.31%)	Seasonal allergies (71.45%)	Arthritis of any kind or rheumatism (46.04%)
Seasonal allergies	1,613	49.47	45.26	54.74	Nasal allergies or rhinitis (69.87%)	Hypertension (38.50%)	Chronic allergies or sinus trouble (38.00%)
Chronic back problems or sciatica	893	56.10	46.47	53.53	Arthritis of any kind or rheumatism (60.25%)	Hypertension (53.64%)	Seasonal allergies (50.50%)
Blindness or trouble seeing (even with glasses)	472	56.50	46.61	53.39	Hypertension (54.24%)	Arthritis of any kind or rheumatism (51.06%)	Seasonal allergies (48.31%)
Deafness or trouble hearing	512	62.09	67.19	32.81	Hypertension (58.59%)	Arthritis of any kind or rheumatism (54.49%)	Seasonal allergies (42.58%)
Arthritis of any kind or rheumatism	1,197	61.52	43.27	56.73	Hypertension (57.89%)	Seasonal allergies (46.12%)	Osteoarthritis (45.28%)
Dermatitis or other chronic skin conditions	404	53.83	43.81	56.19	Seasonal allergies (56.93%)	Nasal allergies or rhinitis (52.48%)	Arthritis of any kind or rheumatism (48.02%)
Asthma	342	47.86	34.50	65.50	Seasonal allergies (71.93%)	Nasal allergies or rhinitis (67.84%)	Chronic allergies or sinus trouble (50.88%)
Chronic lung disease other than asthma	196	59.91	43.37	56.63	Arthritis of any kind or rheumatism (64.80%)	COPD (60.71%)	Hypertension (56.12%)
Ulcer	130	54.34	46.15	53.85	Seasonal allergies (54.62%)	Hypertension (53.85%)	Arthritis of any kind or rheumatism (53.85%)
Depression	599	48.17	37.90	62.10	Anxiety (66.78%)	Clinical depression (64.77%)	Seasonal allergies (49.58%)
Anxiety	627	47.57	37.16	62.84	Depression (63.80%)	Seasonal allergies (52.15%)	Nasal allergies or rhinitis (48.96%)
Severe headaches	406	44.15	31.28	68.72	Migraine headaches (66.01%)	Seasonal allergies (55.67%)	Nasal allergies or rhinitis (53.94%)
Limitation in the use of an arm or leg	376	57.94	50.53	49.47	Arthritis of any kind or rheumatism (63.56%)	Hypertension (59.04%)	Chronic back problems or sciatica (52.66%)

Note. The three most prevalent comorbid conditions are shown for each disease group sample, along with the percentage of the disease group's respondents who reported each comorbidity (in parentheses).

Table 14.15*Characteristics of 2009 SF-36v2 Acute (1-Week Recall) Form Disease-Specific Benchmark Samples*

	<i>n</i>	Mean Age	% Male	% Female	Most Common Comorbid Conditions		
					1st	2nd	3rd
Hypertension	797	59.60	51.19	48.81	Arthritis of any kind or rheumatism (45.67%)	Seasonal allergies (42.66%)	Nasal allergies or rhinitis (38.14%)
Myocardial infarction (within last year)	40	58.90	62.50	37.50	Hypertension (87.50%)	Diabetes (60.00%)	Nasal allergies or rhinitis (45.00%)
Congestive heart failure	79	62.25	63.29	36.71	Hypertension (83.54%)	Arthritis of any kind or rheumatism (55.70%)	Diabetes (51.90%)
Angina	95	65.71	63.16	36.84	Hypertension (85.26%)	Arthritis of any kind or rheumatism (61.05%)	Diabetes (49.47%)
Other heart conditions	240	58.11	46.25	53.75	Hypertension (60.00%)	Arthritis of any kind or rheumatism (50.42%)	Nasal allergies or rhinitis (45.83%)
Diabetes	304	59.19	48.36	51.64	Hypertension (73.68%)	Arthritis of any kind or rheumatism (47.70%)	Nasal allergies or rhinitis (39.80%)
Cancer (not skin cancer)	158	64.06	42.41	57.59	Hypertension (60.13%)	Arthritis of any kind or rheumatism (53.16%)	Osteoarthritis (34.81%)
COPD	98	63.21	47.96	52.04	Chronic lung disease other than asthma (69.39%)	Hypertension (64.29%)	Arthritis of any kind or rheumatism (57.14%)
Nasal allergies or rhinitis	707	50.83	39.89	60.11	Seasonal allergies (81.90%)	Chronic allergies or sinus trouble (49.50%)	Hypertension (43.00%)
Rheumatoid arthritis	182	61.69	42.86	57.14	Arthritis of any kind or rheumatism (89.01%)	Hypertension (60.44%)	Seasonal allergies (52.20%)
Osteoarthritis	331	61.79	33.84	66.16	Arthritis of any kind or rheumatism (93.96%)	Hypertension (59.21%)	Chronic back problems or sciatica (52.57%)
Osteoporosis	146	64.71	13.70	86.30	Arthritis of any kind or rheumatism (71.23%)	Hypertension (58.22%)	Osteoarthritis (53.42%)
Kidney disease	68	60.44	57.35	42.65	Hypertension (66.18%)	Arthritis of any kind or rheumatism (51.47%)	Diabetes (47.06%)
Liver disease	47	55.23	53.19	46.81	Hypertension (61.70%)	Chronic back problems or sciatica (55.32%)	Depression (55.32%)
GERD	325	58.59	41.54	58.46	Hypertension (58.46%)	Arthritis of any kind or rheumatism (55.38%)	Nasal allergies or rhinitis (51.38%)
Stomach disease	84	55.10	38.10	61.90	Seasonal allergies (66.67%)	Nasal allergies or rhinitis (63.10%)	Arthritis of any kind or rheumatism (59.52%)
Chronic bowel disease	184	54.32	22.83	77.17	Nasal allergies or rhinitis (57.61%)	Seasonal allergies (57.07%)	Arthritis of any kind or rheumatism (54.35%)
Obesity	360	51.41	33.89	66.11	Hypertension (61.39%)	Seasonal allergies (47.50%)	Arthritis of any kind or rheumatism (46.39%)
Stroke	58	67.00	46.55	53.45	Hypertension (81.03%)	Arthritis of any kind or rheumatism (58.62%)	Diabetes (41.38%)
HIV or AIDS	1	20.00	0.00	100.00	—	—	—
Anemia	233	53.59	14.16	85.84	Arthritis of any kind or rheumatism (50.64%)	Nasal allergies or rhinitis (50.21%)	Seasonal allergies (49.79%)
Clinical depression	252	48.26	29.76	70.24	Depression (81.35%)	Anxiety (62.70%)	Seasonal allergies (51.98%)
Alcohol- or drug-use disorder	56	52.71	73.21	26.79	Hypertension (53.57%)	Clinical depression (50.00%)	Depression (50.00%)
Chronic fatigue syndrome	81	54.44	12.35	87.65	Arthritis of any kind or rheumatism (76.54%)	Chronic back problems or sciatica (64.20%)	Nasal allergies or fibromyalgia or rhinitis (58.02%)

Table 14.15 (Continued)

Characteristics of 2009 SF-36v2 Acute (1-Week Recall) Form Disease-Specific Benchmark Samples

	<i>n</i>	Mean Age	% Male	% Female	Most Common Comorbid Conditions		
					1st	2nd	3rd
Migraine headaches	318	47.58	24.84	75.16	Severe headaches (55.97%)	Seasonal allergies (55.03%)	Nasal allergies or rhinitis (54.40%)
Sleep apnea	222	55.33	59.01	40.99	Hypertension (62.61%)	Arthritis of any kind or rheumatism (50.45%)	Seasonal allergies (45.50%)
Chronic allergies or sinus trouble	468	52.18	40.60	59.40	Nasal allergies or rhinitis (74.79%)	Seasonal allergies (67.95%)	Hypertension (47.22%)
Seasonal allergies	815	49.88	43.07	56.93	Nasal allergies or rhinitis (71.04%)	Hypertension (41.72%)	Chronic allergies or sinus trouble (39.02%)
Chronic back problems or sciatica	468	56.21	43.16	56.84	Arthritis of any kind or rheumatism (58.76%)	Hypertension (51.07%)	Nasal allergies or rhinitis (48.08%)
Blindness or trouble seeing (even with glasses)	249	57.59	48.19	51.81	Hypertension (53.82%)	Arthritis of any kind or rheumatism (51.81%)	Seasonal allergies (43.37%)
Deafness or trouble hearing	275	62.60	63.64	36.36	Hypertension (53.09%)	Arthritis of any kind or rheumatism (53.09%)	Seasonal allergies (43.64%)
Arthritis of any kind or rheumatism	621	61.16	40.74	59.26	Hypertension (58.62%)	Osteoarthritis (50.08%)	Seasonal allergies (48.31%)
Dermatitis or other chronic skin conditions	205	53.25	43.41	56.59	Seasonal allergies (52.68%)	Arthritis of any kind or rheumatism (51.22%)	Nasal allergies or rhinitis (50.73%)
Asthma	193	48.28	35.23	64.77	Nasal allergies or rhinitis (75.13%)	Seasonal allergies (74.09%)	Chronic allergies or sinus trouble (59.59%)
Chronic lung disease other than asthma	119	60.25	40.34	59.66	Arthritis of any kind or rheumatism (58.82%)	COPD (57.14%)	Hypertension (54.62%)
Ulcer	59	56.88	49.15	50.85	Hypertension (54.24%)	Chronic back problems or sciatica (52.54%)	Arthritis of any kind or rheumatism (50.85%)
Depression	327	48.18	36.39	63.61	Anxiety (70.64%)	Clinical depression (62.69%)	Seasonal allergies (48.01%)
Anxiety	355	48.68	33.52	66.48	Depression (65.07%)	Seasonal allergies (52.96%)	Nasal allergies or rhinitis (50.14%)
Severe headaches	245	43.80	31.84	68.16	Migraine headaches (72.65%)	Seasonal allergies (51.84%)	Nasal allergies or rhinitis (49.80%)
Limitation in the use of an arm or leg	197	58.40	47.21	52.79	Arthritis of any kind or rheumatism (64.97%)	Hypertension (60.41%)	Chronic back problems or sciatica (48.73%)

Note. The three most prevalent comorbid conditions are shown for each disease group sample, along with the percentage of the disease group's respondents who reported each comorbidity (in parentheses). No comorbid conditions were reported by the HIV/AIDS patient.

Table 14.16

Mean Item Raw Scores for SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 and 1998 U.S. General Populations

Scale	Item	Standard Form				Acute Form			
		2009		1998		2009		1998	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
		Norm Sample (N = 4,040)		Norm Sample (N = 5,038)		Norm Sample (N = 2,061)		Norm Sample (N = 6,137)	
PF	3a	2.12	0.79	2.11	0.80	2.15	0.77	2.09	0.81
	3b	2.61	0.65	2.65	0.62	2.61	0.65	2.63	0.64
	3c	2.72	0.56	2.73	0.55	2.74	0.55	2.71	0.58
	3d	2.45	0.73	2.48	0.71	2.47	0.72	2.48	0.73
	3e	2.72	0.58	2.74	0.55	2.74	0.56	2.72	0.58
	3f	2.46	0.69	2.52	0.67	2.50	0.68	2.51	0.69
	3g	2.46	0.77	2.49	0.75	2.48	0.75	2.47	0.76
	3h	2.67	0.64	2.70	0.61	2.71	0.60	2.68	0.63
	3i	2.74	0.56	2.78	0.53	2.76	0.55	2.76	0.55
	3j	2.88	0.39	2.89	0.38	2.91	0.33	2.90	0.37
RP	4a	4.28	1.12	4.37	1.08	4.32	1.06	4.36	1.11
	4b	4.08	1.21	4.07	1.20	4.18	1.15	4.06	1.25
	4c	4.15	1.22	4.24	1.18	4.22	1.17	4.20	1.22
	4d	4.15	1.19	4.24	1.15	4.22	1.16	4.23	1.18
BP	7	4.37	1.27	4.37	1.29	4.53	1.28	4.49	1.30
	8	4.19	1.06	4.22	1.01	5.30	1.01	4.27	1.01
GH	1	3.37	0.92	3.54	0.93	3.36	0.94	3.53	0.94
	11a	4.22	1.04	4.25	1.01	4.25	1.03	4.21	1.04
	11b	3.56	1.17	3.74	1.16	3.59	1.17	3.75	1.16
	11c	3.47	1.12	3.62	1.12	3.44	1.14	3.61	1.15
	11d	3.31	1.24	3.57	1.23	3.32	1.25	3.55	1.23
VT	9a	3.45	1.04	3.50	0.95	3.41	1.11	3.47	1.00
	9e	3.12	1.03	3.21	1.03	3.10	1.07	3.19	1.05
	9g	3.50	1.01	3.50	1.01	3.50	1.02	3.51	1.04
	9i	3.16	0.98	3.21	0.96	3.21	0.98	3.19	0.98
SF	6	4.34	1.02	4.36	1.00	4.41	1.00	4.38	1.02
	10	4.28	1.08	4.34	1.01	4.39	1.02	4.35	1.04
RE	5a	4.45	0.98	4.50	0.94	4.54	0.89	4.50	0.96
	5b	4.36	1.03	4.32	1.06	4.53	0.88	4.33	1.08
	5c	4.56	0.87	4.50	0.91	4.67	0.77	4.52	0.93
MH	9b	4.19	0.95	4.16	0.96	4.34	0.91	4.22	0.95
	9c	4.46	0.86	4.45	0.88	4.56	0.82	4.49	0.88
	9d	3.46	0.96	3.48	0.97	3.48	0.99	3.49	1.00
	9f	4.24	0.93	4.20	0.96	4.34	0.91	4.26	0.95
	9h	3.65	0.89	3.76	0.86	3.61	0.96	3.75	0.88

Note. 1998 normative data taken from Ware et al. (2007).

Users are encouraged to score and interpret their SF-36v2 data using the 2009 algorithms and norms for several reasons. First, these algorithms represent the most up-to-date methods and yield results based on the most current norms. Second, SF-36v2 data scored by the 2009 algorithms enable the user to draw upon the extensive content- and criterion-based interpretation data provided in Chapters 8 and 9 of this manual. This allows for much more interpretive information than is available when 1998 scoring is employed. Third,

2009 norm-based scores can be compared directly to the 2009 benchmark data for each of 40 diseases and conditions, thus expanding the SF-36v2's applicability and utility to a much larger patient population than had previously been the case. Finally, the standard scoring services and products offered by QualityMetric and its authorized resellers are now based on the 2009 norms. Scoring of SF-36v2 data using the 1998 norms is available only through special requests made to QualityMetric.

Ultimately, the decision to use SF-36v2 1998 or 2009 norms is up to the user and what best suits his or her needs. It is important to be mindful, however, that the 2009 norm-based scoring centers each scale and

summary measure *T* score to the “average” in the 2009 general U.S. population, and therefore scores will differ slightly than those computed using norm-based scoring methods based on the 1998 general U.S. population.

Table 14.17

Differences in SF-36v2 Standard (4-Week Recall) Form Mean Health Domain Scale T Scores, 2009 and 1998 U.S. General Population

	Mean T-Score Difference*	<i>t</i>	<i>p</i>
PF	0.66	3.78	<.001
RP	0.53	2.91	.004
BP	0.31	1.68	.094
GH	-0.28	-1.39	.163
VT	0.50	2.50	.012
SF	0.62	3.21	.001
RE	0.62	3.25	.001
MH	0.48	2.54	.011

*Mean 2009 *T* score minus mean 1998 *T* score.

Table 14.18

Differences in SF-36v2 Acute (1-Week Recall) Form Mean Health Domain Scale T Scores, 2009 and 1998 U.S. General Population

	Mean T-Score Difference*	<i>t</i>	<i>p</i>
PF	1.78	8.42	<.001
RP	1.67	7.68	<.001
BP	1.22	5.54	<.001
GH	0.66	2.92	.004
VT	0.80	3.48	<.001
SF	1.47	6.51	<.001
RE	2.15	8.79	<.001
MH	1.22	5.64	<.001

*Mean 2009 *T* score minus mean 1998 *T* score.

Table 14.19

Comparison of SF-36v2 Standard (4-Week Recall) Form Mean Health Domain Scale T Scores Using 2009 and 1998 Scoring Algorithms, 2009 U.S. General Population (N = 4,040)

	<i>N</i>	2009 Algorithms		1998 Algorithms		Mean T-Score Difference*	<i>t</i>	<i>p</i>
		Mean	<i>SD</i>	Mean	<i>SD</i>			
PF	4,034	50.00	10.00	48.74	11.00	1.26	5.39	<.0001
RP	4,027	50.00	10.00	49.05	10.91	0.95	4.08	<.0001
BP	4,027	50.00	10.00	49.54	10.49	0.46	2.00	.05
GH	4,036	50.00	10.00	47.36	10.03	2.64	11.86	<.0001
VT	4,028	50.00	10.00	49.36	10.52	0.64	2.80	<.01
SF	4,029	50.00	10.00	48.87	10.89	1.13	4.87	<.0001
RE	4,026	50.00	10.00	48.99	11.17	1.01	4.26	<.0001
MH	4,028	50.00	10.00	49.07	10.77	0.93	4.00	.0001

*Mean 2009 norms-based *T* score minus mean 1998 norms-based *T* score.

Table 14.20

Comparison of SF-36v2 Acute (1-Week Recall) Form Mean Health Domain Scale T Scores Using 2009 and 1998 Scoring Algorithms, 2009 U.S. General Population (N = 2,061)

	<i>N</i>	2009 Algorithms		1998 Algorithms		Mean T-Score Difference*	<i>t</i>	<i>p</i>
		Mean	<i>SD</i>	Mean	<i>SD</i>			
PF	2,059	50.00	10.00	49.05	10.61	0.95	2.96	<.01
RP	2,057	50.00	10.00	48.91	10.84	1.09	3.37	<.001
BP	2,056	50.00	10.00	49.41	10.55	0.59	1.83	.07
GH	2,061	50.00	10.00	47.32	10.65	2.68	8.32	<.0001
VT	2,057	50.00	10.00	48.85	11.00	1.15	3.50	<.001
SF	2,057	50.00	10.00	49.07	10.88	0.93	2.85	<.01
RE	2,057	50.00	10.00	50.09	9.92	-0.09	-0.28	.78
MH	2,060	50.00	10.00	49.26	11.18	0.74	2.24	<.05

*Mean 2009 norms-based *T* score minus mean 1998 norms-based *T* score.

15

Reliability

The SF-36 is considered a reliable measure of health status based on years of empirical research (Garratt, Schmidt, Mackintosh, & Fitzpatrick, 2002; McDowell & Newell, 1996). This tradition continues with the SF-36v2 by retaining SF-36 item content and making past empirical work on the reliability of the SF-36 generalizable to the SF-36v2. As expected, the improvements made to SF-36v2 response choices have allowed for further improvement in the reliability and precision of the RP and RE health domain scales and, consequently, the reliability of the PCS and MCS measures, while the reliabilities of the other six health domain scales essentially remain unchanged.

This chapter presents reliability estimates for the SF-36v2 health domain scales and component summary measures and describes the methods used in calculating these estimates using data from the 2009 U.S. general population samples and the Medical Outcomes Survey (MOS; Stewart & Ware, 1992). Specifically, internal consistency and test-retest reliability data for the health domain scales and component summary measures are presented, as are standard errors of measurement for each scale and measure.

Interpreting Reliability Coefficients

Indices of reliability provide an indication of the extent to which 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 a respondent achieves the same scores 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 true, as opposed to being the result of chance or random error. For example, a reliability coefficient of .80 indicates

that 80% of the total measured variance is *true score variance*. It is suggested that scales used in group-level analyses should have a reliability coefficient of .70 or greater, while scales used in making decisions at the respondent level should have a reliability coefficient of .90 or greater (Nunnally & Bernstein, 1994). However, as discussed in Chapter 3, these minimum reliability levels should be viewed as suggested guidelines or recommendations rather than strict criteria. As an alternative to reporting reliability, the standard error of measurement (*SEM*) can be reported. This is particularly useful when interpreting individual scores (Anastasi, 1988).

Trends toward higher reliability coefficients for many of the newer health status surveys reflect both conceptual and methodological advances. First, the amount of information gained from each questionnaire item has increased because newer instruments tend to use response scales that have five or six choices, rather than only two. Thus, the scores for scales constructed from these items tend to be more reliable because each item yields more information. Second, in newer instruments, items in the same scale tend to define more homogeneous constructs and therefore yield more reliable scores. A *homogeneity coefficient* represents the average of a given scale's interitem correlations. As such, this coefficient indicates the internal consistency, or quality, of the measure, regardless of the number of items (Thissen & Wainer, 2001; Tyler & Fiske, 1968). Failure to consider homogeneity can be problematic because the calculation of alpha coefficients is influenced by the number of items. For example, items pertaining to mental health, physical symptoms, functional status, general health perceptions, and smoking were all included in one early health measure (Macmillan, 1957); these items were quite heterogeneous (i.e., low in internal consistency) and yielded a relatively low reliability coefficient (Ware, Johnston, Davies-Avery, & Brook, 1979). Therefore, it is not surprising that interpretation of this early health

measure was complicated. Had this instrument contained many more homogeneous items, it could have yielded a higher reliability coefficient, in which case the homogeneity coefficient would have provided a better estimate of internal consistency.

There are several methods that can be used to estimate reliability. For example, reliability can be estimated by correlating responses to items within the same scale or measure from a single administration (*internal consistency reliability*), by correlating scores from one administration with scores from another administration at some later point in time (*test-retest reliability*), or by correlating scores or otherwise examining the equivalence of individual answers across alternate forms of an instrument (*alternate forms reliability*; Nunnally & Bernstein, 1994).

The reliability of a score for a given scale depends on the number of items in the scale and the homogeneity of said items. As the demand for health status measurement tools has grown, different approaches have explored the trade-off between the length of a measure and its reliability and validity. Medical practitioners and clinical researchers find lengthy scales less practical for widespread use. To this end, single-item measures have been documented to correlate with long-form (parent) measures (Coates et al., 1987; Meyerboom-DeJong & Smith, 1990; Nelson, Landgraf, Hays, Kirk, et al., 1990; Nelson, Landgraf, Hays, Wasson, & Kirk, 1990; Nelson et al., 1987; Stewart & Ware, 1992) and are satisfactory for use in detecting moderate to large differences between groups of 150 or more patients. That said, single-item measures may be attractive but, from a practical point of view, would probably not detect meaningful differences at the individual respondent level because too much precision is lost. Short-form, multi-item measures that meet minimum psychometric standards while reducing respondent burden provide a compromise between single-item and long-form measures. For example, the SF-36 MH health domain scale, the five-item version of the Mental Health Inventory (MHI; Veit & Ware, 1983), has 84% fewer items than the full-length scale, with only a 7% drop in precision (McHorney, Ware, Rogers, Raczek, & Lu, 1992). Thus, a well-constructed short-form measure can provide a reasonable balance between the requirements of reliability and the demands of everyday use in clinical research and practice.

Internal Consistency Reliability

The internal consistency of a scale or measure refers to the degree to which its items measure the same

construct. Internal consistency is typically assessed by one or both of two measures: Cronbach's alpha coefficient and item-scale (or item-total) correlations. Both types of statistics were calculated for the eight SF-36v2 health domain scales. A somewhat different approach was taken to arrive at estimates of internal consistency for the PCS and MCS measures, which is discussed later in the following section of this chapter.

Data from the 2009 U.S. general population were used to estimate the internal consistency of the SF-36v2 health domain scales and component summary measures. Sampling procedures are documented in detail in Chapter 14 of this manual.

PCS and MCS Internal Consistency Estimates

Because the PCS and MCS measures are linear combinations of eight scales measuring distinct health constructs, it is necessary to take into account the reliability of each scale, as well as the covariances amongst them, when estimating reliability using the internal consistency method (Nunnally & Bernstein, 1994). Using the covariance matrix of the SF-36v2 scales in each sample, along with the physical and mental factor score coefficients from the 2009 U.S. general population (see Chapter 5 for the component summary measure scoring steps), reliabilities for the PCS and MCS measures were estimated using the following procedure:

1. Each off-diagonal covariance was multiplied by the product of its respective factor score coefficient, summed, and multiplied by two (i.e., two sides of the matrix) due to symmetry along the diagonal.
2. Observed score variance was calculated by multiplying each diagonal of the covariance matrix by the squared factor score coefficient.
3. Total score variance was calculated by summing the products of Steps 1 and 2.
4. True score variance was calculated by multiplying each diagonal entry (Step 2) by its respective scale reliability.

Upon completion of Steps 1 through 4, each component summary measure reliability coefficient was computed by subtracting the true score variance (Step 4) from the observed variance (Step 2), dividing the result by the total score variance (Step 3), and subtracting this result from one.

Table 15.1 presents SF-36v2 internal consistency coefficients for the PCS and MCS measures and the eight health domain scales, estimated in the 2009 U.S. general population study for respondents that provided complete data. PCS reliability coefficients of .96 and .97 were found

for the standard and acute forms, respectively, whereas a coefficient of .93 was obtained for the MCS measure from both forms. Thus, the PCS and MCS estimates met the recommended minimum standard of reliability for both group-level (.70) and respondent-level (.90) comparisons (Nunnally & Bernstein, 1994).

Table 15.1

Internal Consistency Reliability Estimates for the SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 U.S. General Population

Measure/Scale	Standard Form (N = 4,024–4,036)	Acute Form (N = 1,983–2,047)
Physical Component Summary	.96	.97
Mental Component Summary	.93	.93
Physical Functioning	.94	.95
Role-Physical	.96	.96
Bodily Pain	.87	.88
General Health	.82	.85
Vitality	.87	.87
Social Functioning	.84	.81
Role-Emotional	.93	.94
Mental Health	.87	.88

Note. For each measure and scale, alpha coefficients were computed using only those respondents with complete data for that measure or scale.

Tables 15.2 and 15.3 present estimated PCS and MCS reliabilities from the 2009 U.S. general population sample for age and gender subgroups, respondents with no reported chronic conditions (i.e., “healthy”), and those reporting one or more physical and/or mental conditions. All PCS and MCS coefficients were .90 or higher, again meeting the minimum standards for both group- and respondent-level comparisons. Across all subgroups on both forms, reliability estimates for the PCS and MCS measures are generally higher or the same as the reliability estimates for the eight health domain scales; however, several exceptions are noted on both forms. Overall, the evidence indicates the greater reliability of the component summary measures compared to that of the eight health domain scales, with the standard and acute form internal consistency reliability for the PCS and MCS measures ranging from .90 to .97 across the general population subgroups.

Health Domain Scale Alpha Coefficients

Table 15.1 presents SF-36v2 Cronbach’s alpha coefficients for the eight health domain scales, estimated for respondents that provided complete data in the 2009 U.S. general population study. For both the standard and acute forms, all coefficients exceeded the recommended minimum standard for group-level comparison of scores (.70; Nunnally & Bernstein, 1994). The reli-

ability coefficients ranged from .82 (GH) to .96 (RP) across the eight standard form scales; for the acute form, the coefficients ranged from .81 (SF) to .96 (RP).

Tables 15.2 and 15.3 present Cronbach’s alpha coefficients for the SF-36v2 standard and acute form component summary measures and health domain scales, respectively, estimated in the 2009 U. S. general population for age and gender subgroups, healthy respondents, and respondents indicating the presence of one or more physical and/or mental health conditions. Examining the results for the eight standard form health domain scales across all general population subgroups (Table 15.2), the vast majority of the reliability coefficients were in the .80s and .90s and, with few exceptions, all reliability estimates for the eight health domain scales exceeded the recommended minimum standard for group-level comparisons (.70). In fact, with only one exception, the coefficients for the PF, RP, and RE scales were .90 or higher, exceeding the recommended minimum standard for individual respondent-level comparisons. The standard form alpha coefficients that fell below the .70 threshold were found in the following subgroups: healthy (.64 on SF), heart attack (.42 and .31 on GH and SF, respectively), congestive heart failure (.63 and .66 on GH and SF, respectively), angina (.68 on GH), and HIV/AIDS (.64 and .43 on VT and SF, respectively).

Finally, Table 15.3 presents subgroup alpha coefficients for the SF-36v2 acute form. As with the standard form, the vast majority of the general population health domain scale reliability coefficients across all subgroups were in the .80s and .90s. With only two exceptions (SF coefficients in the heart attack [.65] and diabetes [.69] subgroups), all reliability estimates for the eight health domain scales exceeded the recommended minimum standard for group-level comparisons (.70).

Health Domain Scale Item-Scale Correlations

Another measure of internal consistency for the health domain scales is item-scale correlations; in other words, the correlation of each item with the scale for which it is scored. The correlations between the SF-36v2 items and the health domain scales in the 2009 U.S. general population for the standard and acute forms, corrected for overlap, are presented in Tables 14.10 and 14.11, respectively. Examination of these tables reveals that within each scale, correlations between items and their hypothesized scale exceeded the .40 standard for internal consistency (Helmstader, 1964) for both the standard and acute forms. Also, with the exception of one standard form item (PF Item 3a), items correlated higher with their parent scale than with any of the other scales.

Table 15.2

Internal Consistency Reliability Estimates for the SF-36v2 Standard (4-Week Recall) Form Component Summary Measures and Health Domain Scales, by Respondent Subgroup in the 2009 U.S. General Population

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
Age										
18–44	.94	.94	.94	.94	.85	.79	.81	.79	.94	.83
45–64	.96	.92	.95	.96	.89	.83	.87	.84	.95	.88
65+	.96	.92	.94	.96	.89	.83	.87	.87	.92	.85
Gender										
Male	.96	.93	.95	.96	.88	.80	.83	.79	.94	.86
Female	.96	.93	.95	.96	.89	.82	.84	.85	.94	.86
Condition										
Healthy ^a	.96	.93	.94	.93	.80	.75	.78	.64	.92	.79
Hypertension	.96	.93	.94	.96	.89	.81	.85	.84	.95	.88
Heart attack	.96	.94	.95	.94	.94	.42	.72	.31	.98	.76
Congenital heart failure	.96	.93	.94	.95	.91	.63	.77	.66	.93	.81
Angina	.96	.91	.95	.96	.91	.68	.81	.72	.93	.74
Other heart conditions	.96	.93	.95	.95	.89	.80	.87	.81	.93	.88
Diabetes	.96	.93	.94	.96	.90	.79	.87	.85	.94	.88
Cancer	.96	.93	.94	.96	.91	.81	.82	.77	.95	.84
COPD	.96	.93	.94	.97	.92	.81	.84	.88	.95	.83
Allergies	.96	.93	.95	.96	.89	.83	.87	.87	.94	.88
Rheumatoid arthritis	.96	.93	.94	.96	.93	.80	.87	.86	.95	.89
Osteoarthritis	.96	.93	.94	.96	.89	.82	.84	.87	.95	.90
Osteoporosis	.96	.93	.93	.96	.90	.83	.86	.88	.96	.89
Kidney disease	.96	.93	.94	.95	.91	.79	.81	.83	.93	.87
Liver disease	.96	.92	.95	.94	.95	.83	.86	.91	.94	.84
GERD	.96	.93	.94	.96	.89	.80	.88	.89	.94	.88
Stomach disease	.96	.93	.95	.97	.90	.79	.85	.89	.95	.84
IBS	.96	.93	.95	.96	.88	.79	.85	.90	.93	.86
Obesity	.96	.94	.94	.95	.89	.77	.81	.87	.93	.87
Stroke	.95	.94	.93	.94	.89	.78	.77	.73	.94	.85
HIV/AIDS	.95	.90	.94	.92	.89	.88	.64	.43	.97	.68
Anemia	.96	.93	.95	.95	.90	.80	.84	.85	.91	.89
Clinical depression	.95	.93	.95	.96	.91	.83	.82	.90	.94	.89
Alcohol/drug use	.96	.94	.94	.91	.90	.84	.85	.88	.91	.90
Chronic fatigue	.96	.93	.95	.96	.89	.76	.81	.89	.94	.89
Migraine	.96	.93	.95	.96	.90	.81	.86	.89	.95	.88
Sleep apnea	.96	.93	.96	.96	.92	.78	.82	.88	.95	.88
Chronic allergies	.96	.93	.95	.96	.91	.81	.86	.88	.94	.89
Seasonal allergies	.96	.93	.95	.96	.89	.83	.87	.86	.94	.87
Back problems	.96	.93	.94	.95	.86	.77	.82	.86	.95	.89
Trouble seeing	.96	.93	.94	.96	.90	.75	.81	.85	.94	.86
Trouble hearing	.96	.93	.95	.95	.90	.78	.84	.89	.94	.88
Any arthritis	.96	.93	.94	.96	.89	.80	.86	.85	.95	.89
Skin conditions	.96	.94	.95	.96	.90	.84	.84	.86	.93	.87
Asthma	.96	.93	.95	.96	.91	.83	.84	.84	.96	.85
Lung other than asthma	.96	.93	.95	.97	.91	.77	.82	.82	.95	.86
Ulcer	.96	.93	.95	.96	.93	.82	.79	.86	.96	.86
Depression	.96	.93	.94	.96	.90	.80	.80	.85	.93	.85
Anxiety	.96	.93	.95	.96	.91	.82	.81	.86	.93	.85
Severe headaches	.96	.93	.95	.96	.91	.81	.85	.87	.95	.90
Limited use of arm/leg	.96	.94	.93	.95	.88	.74	.82	.86	.96	.90

Note. Internal consistency reliabilities of the component summary measures were estimated for the 2009 U.S. general population using Nunnally and Bernstein's (1994) method that takes into account the reliability of each health domain scale, as well as the covariances amongst them. Internal consistency reliabilities of the health domain scales were estimated for the 2009 U.S. general population using Cronbach's alpha coefficient.

^aRespondents from the 2009 U.S. general population sample who reported never having been told they had any of 18 physical conditions or an alcohol or drug use disorder, and were not currently experiencing anxiety or depression.

Table 15.3

Internal Consistency Reliability Estimates for the SF-36v2 Acute (1-Week Recall) Form Component Summary Measures and Health Domain Scales, by Respondent Subgroup in the 2009 U.S. General Population

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
Age										
18–44	.95	.94	.92	.94	.83	.82	.86	.77	.94	.87
45–64	.97	.93	.95	.97	.91	.87	.89	.86	.95	.89
65+	.97	.91	.93	.96	.89	.86	.88	.84	.93	.86
Gender										
Male	.97	.93	.95	.97	.85	.85	.86	.84	.95	.87
Female	.97	.94	.94	.96	.90	.84	.88	.79	.94	.89
Condition										
Healthy ^a	.94	.94	.90	.93	.82	.78	.82	.70	.86	.83
Hypertension	.97	.93	.95	.96	.88	.84	.88	.80	.94	.89
Heart attack	.96	.93	.95	.97	.90	.83	.72	.65	.99	.90
Congenital heart failure	.96	.95	.95	.96	.89	.84	.89	.89	.97	.91
Angina	.96	.93	.96	.97	.94	.90	.89	.92	.97	.93
Other heart conditions	.97	.93	.95	.98	.90	.87	.88	.88	.96	.90
Diabetes	.96	.93	.95	.97	.84	.83	.87	.69	.96	.89
Cancer (except skin)	.96	.94	.92	.95	.84	.78	.86	.73	.98	.87
COPD	.96	.93	.92	.94	.89	.86	.86	.81	.94	.87
Allergies (nasal)	.97	.93	.95	.96	.91	.85	.88	.80	.95	.89
Rheumatoid arthritis	.96	.94	.94	.97	.90	.87	.90	.83	.96	.91
Osteoarthritis	.97	.94	.94	.96	.90	.84	.90	.90	.96	.92
Osteoporosis	.96	.94	.95	.97	.92	.89	.92	.89	.97	.92
Kidney disease	.96	.95	.93	.94	.84	.85	.89	.79	.96	.89
Liver disease	.97	.95	.95	.98	.82	.79	.84	.84	.97	.92
GERD	.97	.93	.95	.97	.89	.87	.89	.85	.95	.88
Stomach disease	.97	.94	.92	.97	.87	.86	.86	.84	.97	.87
IBS	.97	.93	.95	.97	.91	.84	.88	.90	.95	.88
Obesity	.97	.94	.95	.97	.91	.80	.86	.77	.95	.91
Stroke	.96	.93	.92	.97	.92	.84	.87	.81	.96	.89
HIV/AIDS ^b	—	—	—	—	—	—	—	—	—	—
Anemia	.97	.94	.96	.97	.91	.85	.88	.87	.95	.89
Clinical depression	.96	.94	.95	.96	.91	.82	.88	.85	.94	.90
Alcohol/drug use	.96	.94	.94	.97	.85	.86	.86	.87	.96	.92
Chronic fatigue syndrome	.97	.95	.94	.96	.88	.79	.86	.85	.96	.87
Migraine headaches	.96	.94	.93	.96	.85	.86	.89	.84	.96	.90
Sleep apnea	.97	.94	.94	.97	.87	.85	.85	.86	.95	.89
Chronic allergies	.97	.94	.94	.97	.89	.86	.87	.87	.95	.89
Seasonal allergies	.97	.94	.94	.96	.89	.85	.87	.80	.94	.88
Chronic back problems	.96	.94	.94	.96	.89	.86	.89	.89	.95	.91
Trouble seeing	.97	.94	.95	.97	.89	.86	.87	.83	.96	.89
Trouble hearing	.97	.94	.94	.96	.89	.89	.89	.89	.95	.91
Any arthritis	.97	.94	.94	.97	.89	.84	.90	.86	.96	.91
Skin conditions	.97	.92	.94	.96	.90	.82	.88	.87	.95	.86
Asthma	.97	.93	.95	.97	.92	.86	.91	.75	.95	.89
Lung other than asthma	.96	.93	.93	.96	.92	.88	.91	.85	.95	.84
Ulcer	.96	.95	.92	.95	.86	.84	.75	.89	.96	.90
Depression	.97	.93	.95	.96	.90	.83	.85	.81	.94	.86
Anxiety	.97	.94	.95	.97	.90	.84	.87	.83	.95	.87
Severe headaches	.96	.94	.95	.97	.88	.85	.89	.86	.96	.90
Limited use of arm/leg	.96	.94	.93	.96	.90	.81	.88	.84	.95	.88

Note. Internal consistency reliabilities of the component summary measures were estimated for the 2009 U.S. general population using Nunnally and Bernstein's (1994) method that takes into account the reliability of each health domain scale, as well as the covariances amongst them. Internal consistency reliabilities of the health domain scales were estimated for the 2009 U.S. general population using Cronbach's alpha coefficient.

^aRespondents from the 2009 U.S. general population sample who reported never having been told they had any of 18 physical conditions or an alcohol or drug use disorder, and were not currently experiencing anxiety or depression.

^bAlpha coefficients could not be calculated because the subsample has only one member.

Test-Retest Reliability

No formal study of test-retest reliability was conducted as part of the 2009 norming study. However, data from a subsample of participants who completed the same survey form twice during the data collection phase of the study were used to provide preliminary estimates of the test-retest reliability of the SF-36v2 and other instruments included in the study forms. (Note that data collected from the second administration of the survey forms to this subsample were *not* included in the main analyses conducted for the SF-36v2.)

Approximately the same number of respondents completed each of the four 2009 norming study forms twice, resulting in the availability of test-retest estimates for the SF-36v2 standard form (Study Forms A and B, combined $N = 147$) and the SF-36v2 acute form (Study Form C, $N = 45$). The mean time between administrations was 106.04 days for the standard form and 105.87 days for the acute form. Correlations of the scores from the first and second administrations for both the SF-36v2 standard and acute forms are presented in Table 15.4. Given a mean retest interval of 15 weeks, the resulting estimates of reliability were excellent: no estimates fell below .60 on either form and only one standard form scale (RE) fell below .70. Although these preliminary findings are quite promising, further study of the SF-36v2's test-retest reliability, using larger samples and a shorter retest interval (e.g., 4 weeks), is warranted.

Table 15.4

Test-Retest Reliability Estimates for the SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 U.S. General Population

Measure/Scale	Standard form ($N = 147$) ^a	Acute form ($N = 45$) ^b
Physical Component Summary	.88	.83
Mental Component Summary	.79	.68
Physical Functioning	.85	.85
Role-Physical	.78	.69
Bodily Pain	.71	.74
General Health	.87	.86
Vitality	.75	.66
Social Functioning	.70	.66
Role-Emotional	.61	.64
Mental Health	.76	.73

^aTime between survey administrations = 80–123 days, mean = 106.04 days, $SD = 5.93$ days.

^bTime between survey administrations = 80–121 days, mean = 105.87 days, $SD = 6.41$ days.

Standard Error of Measurement

The *standard error of measurement (SEM)* is an alternative way of expressing the reliability of a scale. Nunnally and Bernstein (1994) define the *SEM* as the following:

The expected standard deviation of scores for anyone taking a large number of parallel tests... Using [the *SEM*] implicitly assumes that the distribution of errors has the same shape and size for people at different points on the continuum of true scores. (pp. 239–240)

In other words, the *SEM* estimates how confident one can be in a score that a respondent obtains on a given scale or measure. Thus, the *SEM* is useful in interpreting respondent-level scores and in determining whether real differences have occurred in the scores obtained by the same person on two different occasions. This latter application is particularly helpful in monitoring changes in patient status over time.

The *SEM* is calculated using the health domain scale or component summary measure reliabilities with the following formula:

$$SEM = SD\sqrt{1 - r}$$

where *SD* is the standard deviation of the scale or measure and *r* is the reliability coefficient for said scale or measure.

The *SEMs* for the eight health domain scales and two component summary measures are presented in Table 15.5. Using these *SEM* values, one can compute intervals around each health domain scale and component summary measure score, with increasing levels of confidence that the respondent's true score falls within a given interval. Note that *SEM*-based confidence interval

Table 15.5

Standard Errors of Measurement (SEMs) for the SF-36v2 Standard (4-Week Recall) and Acute (1-Week Recall) Forms, 2009 U.S. General Population

Measure/Scale	Standard Form ($N = 4,024$ – $4,036$)	Acute Form ($N = 2,056$ – $2,061$)
Physical Component Summary	2.0	1.8
Mental Component Summary	2.7	2.8
Physical Functioning	2.5	2.2
Role-Physical	2.0	2.0
Bodily Pain	3.6	3.5
General Health	4.2	3.9
Vitality	3.6	3.6
Social Functioning	4.0	4.4
Role-Emotional	2.6	2.4
Mental Health	3.6	3.5

values for the 68%, 80%, 90%, and 95% confidence levels are presented in Chapter 7 as part of the discussion on norm-based interpretation of SF-36v2 results.

New Approaches to Evaluating the Reliability of Survey Instruments

This chapter has presented evidence that supports the reliability of the SF-36v2 using commonly employed statistical measures and interpretation thresholds that were developed using an approach grounded in *classical* psychometric theory. With the application of *modern*

psychometric methods, advances have been made in the manner in which objective measures can be developed, administered, and scored. As these modern approaches become more commonly employed, developers of psychometric instruments will need to think differently about the means by which the reliability and validity of tests, surveys, and other psychometric measures should be demonstrated. Consideration of different means for evaluating the reliability of measures that are constructed using modern psychometric methods, such as Rasch models and IRT, may eventually come to replace the more traditional measures currently used to evaluate the reliability of the SF-36v2.

16

Validity

Studies of measurement validity investigate the meanings of scores and whether they have their intended interpretations. Validity studies also increase the understanding of what a difference or a change in a score means. For example, when enough evidence has been accumulated to show that a given scale measures the intended health concept and does not measure other concepts, the scale is said to be *validated*. The process of validation, however, continues as long as new information is produced about the interpretation and meaning of the measure's scores.

This chapter presents findings from investigations examining the validity of the SF-36v2 when using the 2009 normative data. Evidence of its *construct validity* is presented in the form of data from studies involving factor analysis, item-scale correlations, interscale correlations, correlations of the health domain scales with the component summary measures and the SF-6D, and known-groups comparisons. *Criterion validity* is examined through the correlations of each health domain scale with other HRQOL measures, as well as with health- and work-related variables. Further evidence of criterion validity is provided in the form of data regarding the likelihood of future events (e.g., not working because of health, days in bed due to illness or injury) based on scale score ranges. Finally, *content validity* is examined through a comparison of the survey's coverage of health domains to the health domain coverage of other general health surveys, as well as through a discussion of the rationale for the selection of the domains covered by the Short Form instruments. Note that evidence of the SF-36v2's validity based on the 1998 norms and other studies can be found in the second edition of this manual (Ware et al., 2007) and amongst the more than 17,000 publications involving the Short Form instruments.

The methods used in the studies discussed in this chapter followed the guidelines recommended by the American Educational Research Association (AERA),

American Psychological Association (APA), and the National Council on Measurement in Education (NCME) for validating psychological and educational measures (AERA, APA, & NCME, 1999; see also Chapter 13). These same methods were used to study the validity of the SF-36 health domain scales (Ware, 1993) and component summary measures (Ware & Kosinski, 2001b; Ware, Kosinski, & Keller, 1994).

Types of Validity: An Overview

Like reliability, the validity of psychometric measures can be demonstrated in many ways. The manner in which a test is validated, or the types of evidence presented for the validity of a scale or measure, generally fall onto one of three categories: *construct validity*, *criterion-related validity*, and *content validity*. In this chapter, evidence of the SF-36v2's construct validity is examined through findings from *factor analyses*, *convergent* and *discriminant validation*, and *known-groups comparisons*. The method of known-groups comparison was employed to investigate the ability of the survey's standard and acute forms to distinguish between groups of respondents known to differ in physical and/or psychiatric conditions.

Criterion validity demonstrates that test scores are systematically related to one or more outcome criteria. This approach can be used when external evidence is available and suitable for use as a criterion against which the results of a given test can be compared. However, in order to judge a measure in terms of external evidence (known and independent), the investigator must know what the anticipated results should be. For example, the criterion validity of a general health measure is supported when (a) health status and resource use are negatively correlated, (b) age and physical health are negatively correlated (according to the theory that

physical function declines with increasing age), or (c) physical and mental health are each positively correlated with general health.

The two components of criterion-related validity—*concurrent validity* and *predictive validity*—were separately evaluated for the SF-36v2. In this chapter, evidence of concurrent validity is presented in the form of the survey's correlations or associations with other validated tests measuring the same constructs, or with non-test variables reflecting or representing those constructs, that were administered at approximately the same time. Meanwhile, evidence of predictive validity is presented in the form of its correlations and other associations with non-test variables (e.g., inability to work, health care utilization) that occur subsequent to survey administration.

Content validity, an indication of whether a given survey or scale offers an adequate sample of the construct purported to be measured, is a challenge in the field of general health surveys because of the breadth of health variables. Specifically, content validation requires the existence of a defining standard against which the content of a measure can be compared. Such standards can be based on well-accepted theoretical definitions, on published standards, or on interviews with individuals experiencing the types of health problems under study. When development of the SF-36 began more than 20 years ago, Ware (1987) published a set of standards for evaluating the content validity of comprehensive general health measures. These standards were applied when constructing and evaluating the SF-36v2 and are discussed later in this chapter.

Construct Validity

Construct validity refers to the extent to which a test or survey, or a scale within a test or survey, measures a specific construct or trait (Anastasi, 1988). There are several ways to determine the degree to which an instrument measures the construct that it purports to measure. For the SF-36v2, evidence of construct validity is found in studies involving factor analyses of the eight health domain scales, item-scale correlations, interscale correlations, and known-groups comparisons.

Factor Analyses

Factor analysis provided an empirical test of the SF-36v2's construct validity in relation to its hypothesized structure. Factor analysis also served a second purpose: In the absence of agreed upon criteria for scale validation, the validity of each scale was tested using

factor analytic methods. This methodology has been extensively used for testing the SF-36 measurement model (see Figure 2.1 in Chapter 2) and has provided the basis for the development and testing of the SF-36 and SF-36v2 PCS and MCS measures (McHorney, Ware, & Raczek, 1993; Ware & Kosinski, 2001b; Ware et al., 2007; Ware, Kosinski, & Keller, 1994, 1995).

Both versions of the SF-36 were constructed to represent the two major components of health—physical and mental—that had been confirmed in previous studies (Hays & Stewart, 1990; McHorney et al., 1993; Ware, Kosinski, Bayliss, et al., 1995; Ware et al., 1998; Ware, Kosinski, & Keller, 1994). As such, the two principal components were extracted from the correlations among the standard and acute form health domain scales and then rotated to an orthogonal simple structure. Note that the orthogonal solution has the advantage of permitting interpretation of correlations across components to estimate the factor content of each scale. Because the correlation between physical health and mental health is low (with the correlation between the PF and MH scales being generally less than .40), an orthogonal solution was also expected to reproduce interpretable components. In fact, the two-component solution did account for more than 80% of the reliable variance in health domain scale scores across numerous subgroups in both general and patient populations (Ware & Kosinski, 2001b) in the United States, as well as across the general populations of at least 10 other countries (see Ware et al., 1998).

The two components derived from the factor analytic studies of the SF-36 scales were interpreted on the basis of their correlations with the measure's health domain scales. Because the pattern of correlations across scales was consistent with expectations for the physical and mental dimensions of health, the two components were accordingly labeled *physical* and *mental*. If the two-dimensional structure had not been repeatedly confirmed or if the interpretation of the factors had been ambiguous, these components could not have been used as criteria in testing the validity of each scale.

Table 16.1 shows the factor loadings of the SF-36v2 standard (4-week recall) form's scales and Table 16.2 shows the factor loadings of the acute (1-week recall) form's scales, each based on both the 2009 and 1998 U.S. general population normative data. As hypothesized, the PF scale had the strongest association with the physical component of health and a weak correlation with the mental component of health. At the other extreme, the MH scale had the strongest association with the mental component of health and a weak association with the physical component of health. Overall, these psycho-

Table 16.1

Scale Validity and Correlations With Rotated Principal Components for the SF-36v2 Standard (4-Week Recall) Form, 2009 (N = 4,016) and 1998 (N = 6,742) U.S. General Populations

Scale	2009 SF-36v2 Standard Form Rotated Principal Components			1998 SF-36v2 Standard Form Rotated Principal Components		
	Physical	Mental	h^2	Physical	Mental	h^2
PF	.90	.20	.85	.88	.14	.80
RP	.87	.33	.87	.86	.29	.84
BP	.76	.34	.70	.74	.32	.65
GH	.56	.54	.60	.61	.51	.64
VT	.34	.77	.71	.35	.77	.71
SF	.50	.70	.74	.53	.67	.73
RE	.43	.70	.68	.45	.64	.61
MH	.12	.93	.88	.08	.93	.86
Variance explained						
Total	75%			74%		
Reliable	85%			84%		

Note. The proportion of each scale's total variance that can be explained by the two extracted components is equal to h^2 . Reliable variance explained is the sum of the eight values for both components, divided by the sum of the alpha values for the eight scales.

Table 16.2

Scale Validity and Correlations With Rotated Principal Components for the SF-36v2 Acute (1-Week Recall) Form, 2009 (N = 1,876) and 1998 (N = 7,683) U.S. General Populations

Scale	2009 SF-36v2 Acute Form Rotated Principal Components			1998 SF-36v2 Acute Form Rotated Principal Components		
	Physical	Mental	h^2	Physical	Mental	h^2
PF	.90	.21	.85	.89	.16	.82
RP	.88	.30	.86	.85	.33	.84
BP	.80	.32	.73	.77	.30	.68
GH	.68	.47	.69	.61	.49	.62
VT	.45	.73	.73	.36	.76	.71
SF	.45	.74	.75	.52	.68	.74
RE	.27	.79	.70	.42	.68	.63
MH	.17	.92	.88	.09	.93	.87
Variance explained						
Total	77%			75%		
Reliable	87%			85%		

Note. The proportion of each scale's total variance that can be explained by the two extracted components is equal to h^2 . Reliable variance explained is the sum of the eight values for both components, divided by the sum of the alpha values for the eight scales.

metric tests of construct validity largely confirmed the hypothesized factor content, based on previous research, of the SF-36v2's health domain scales.

As shown in Tables 16.1 and 16.2, the two rotated principal components accounted for more than 80% of the overall reliable variance across the eight health domain scales and for more than 60% of the reliable variance in each individual health domain scale for both forms (standard and acute) of the survey. Overall, the results presented in Tables 16.1 and 16.2 constitute strong evidence for the conceptualization of health that underlies the SF-36v2's construction and provide psy-

chometric data useful in the interpretation of each of its scales. Furthermore, the results clearly indicate that some scales principally measure the physical component of health (PF, RP, and BP), others principally measure the mental component of health (MH, RE, and SF), and still others (GH and VT) appear to be associated with both health components.

Significantly, the results presented in Tables 16.1 and 16.2 also provide confirmation that the SF-36v2's health domain scales for both the standard and acute forms, when using 2009 normative data, generally replicate the two-factor, higher order structure found when using

the 1998 normative data (Ware et al., 2007). Results were consistent for both standard and acute forms. The factors are interpreted as the hypothesized components of physical and mental health, based on the pattern of correlations observed with the health domain scales. For example, as with the 1998 norms, the PF scale loaded highest on the physical component and the MH scale loaded highest on the mental component. Furthermore, the magnitude and pattern of scale-to-component correlations across the survey scales replicated those found in previous factor analytic studies of the SF-36v2 (Ware et al., 2007). These two components accounted for more than 70% of the total variance and more than 80% of the reliable variance in the eight health domain scores across the SF-36v2's standard and acute forms.

Finally, note that the results summarized in Tables 16.1 and 16.2 greatly influenced the formulation of guidelines for the interpretation of the SF-36v2's eight health domain scales. However, as illustrated by the other validation tests presented in this chapter, it is the replication of these results across the clinical criteria used to define physical and mental morbidity that most contributes to the confidence that users can have in these guidelines, as they apply to outcomes research and clinical practice.

Convergent and Discriminant Validity

Convergent and discriminant validity are at the core of construct validation. Convergent validity is supported when different methods of measuring the same construct provide similar results. Discriminant validity is supported when a measure can differentiate its underlying construct from another construct. For example, in the Medical Outcomes Study (MOS; Stewart & Ware, 1992), measures of physical functioning, mobility, and satisfaction with physical abilities were expected to yield results that converged at least moderately with one another because they were all hypothesized to assess physical health (i.e., convergent validity). Conversely, a measure of physical functioning would not be expected to highly relate to a measure of depression or loneliness because different measures should yield different results (i.e., discriminant validity). Note that when more than one method of data collection and/or scale construction has been used to measure the same construct, they can be compared to test convergent validity; however, when different methods have been used to measure different constructs, both convergent and discriminant validity can be tested using the *multitrait-multimethod procedure* (Campbell & Fiske, 1959).

Item-scale correlations. Tests of item discriminant validity focus on the integrity of hypothesized item

groupings relative to the hypothesized health constructs. With regard to the SF-36v2, when the correlation between an item and its hypothesized health domain scale (i.e., a purported measure of a given construct) is significantly higher than the item's correlation with other health domain scales, its inclusion in that hypothesized item grouping is supported. To evaluate the discriminant validity of the SF-36v2, multitrait scaling techniques were employed. For item discriminant validity tests, a success was counted when an item correlated significantly higher (i.e., by two or more standard errors of the correlation coefficient) with its hypothesized health domain scale than with another scale. Then, the item discriminant validity success rate for each SF-36v2 form was computed by dividing the total number of successes by the total number of tests performed. For example, for the PF scale, 80 tests were performed (i.e., each of 10 items correlated with each of the 8 health domain scale scores).

Tables 14.10 and 14.11 present the correlations between items and scales (corrected for overlap) in the 2009 U.S. general population for the SF-36v2 standard and acute forms, respectively. Examination of these correlations reveals that, with one exception (standard form Item 3a), each item correlated highest with its hypothesized scale. Also, note that the lowest of such item correlations were .47 for the standard form and .51 for the acute form. Based on these results, the item discriminant success rate was 99.6% for the standard form and 100% for the acute form.

Interscale correlations. Further evidence of the construct validity of a given multiscale test or measure can be found in the relationship between each scale and every other scale in the test or measure. In terms of the SF-36v2, 2009 general population normative data were used to compute the correlations amongst the standard form (Table 16.3) and acute form (Table 16.4) health domain scales and component summary measures. (These same data also served as the bases for the factor analyses previously discussed.) As shown in Tables 16.3 and 16.4, the pattern and magnitude of the correlations amongst the eight health domain scales were generally comparable across the standard and acute forms. Also, note that the relationships between the health domain scales most closely associated with the physical component of health (PF, RP, BP, and GH) were generally stronger than their relationships between those scales most closely associated with the mental component of health (VT, SF, RE, and MH). However, this trend was less clear for the SF and RE scales than for the other scales. This might have been caused by the highly prevalent somatic comorbidity for respondents with mental health problems, which

Table 16.3

Correlations Between SF-36v2 Standard (4-Week Recall) Form Component Summary Measures and Health Domain Scales, 2009 U.S. General Population (N = 4,021–4,036)

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
PCS	1.00									
MCS	.14	1.00								
PF	.90	.26	1.00							
RP	.87	.39	.83	1.00						
BP	.81	.36	.64	.69	1.00					
GH	.69	.49	.57	.59	.58	1.00				
VT	.49	.71	.47	.53	.55	.63	1.00			
SF	.53	.75	.57	.67	.60	.56	.61	1.00		
RE	.41	.80	.54	.64	.48	.49	.54	.72	1.00	
MH	.23	.92	.34	.42	.42	.53	.69	.66	.67	1.00

Table 16.4

Correlations Between SF-36v2 Acute (1-Week Recall) Form Component Summary Measures and Health Domain Scales, 2009 U.S. General Population (N = 2,056–2,061)

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
PCS	1.00									
MCS	.09	1.00								
PF	.90	.23	1.00							
RP	.88	.32	.83	1.00						
BP	.81	.34	.68	.71	1.00					
GH	.73	.47	.64	.66	.63	1.00				
VT	.51	.72	.52	.55	.59	.70	1.00			
SF	.50	.76	.56	.62	.58	.60	.64	1.00		
RE	.30	.82	.47	.54	.44	.48	.53	.70	1.00	
MH	.23	.92	.36	.41	.44	.54	.75	.69	.69	1.00

can cause difficulty in distinguishing between physical and mental health attribution. Also, the SF items contain attribution to both physical and emotional problems.

Known-Groups Comparisons

The empirical validation of SF-36v2 health domain scales and component summary measures utilized an approach that very closely followed the logic and methods of the SF-36's validation studies. Moreover, this approach's tests were designed to closely parallel the intended uses of the survey. In addition to the methods previously discussed in this chapter, the method of construct validation referred to as *known-groups validity* (Kerlinger, 1973) was also used, just as in previous studies conducted in the MOS with the SF-36 and SF-12 (McHorney et al., 1993; Ware, Kosinski, Bayliss, et al., 1995; Ware, Kosinski, & Keller, 1996) and in studies conducted with the SF-36v2 using 1998 normative data (Ware et al., 2007). In the MOS, the validity of the SF-36 and SF-12 health domain scales and component summary measures was evaluated in terms of their ability to

discriminate between four mutually exclusive groups of patients known to differ in the severity of medical (i.e., physical) and psychiatric conditions. For example, it was expected that the SF-36 and SF-12 health domain scales and component summary measures assessing physical health status would be more valid in discriminating between groups of patients known to differ in the severity of a physical condition than the health domain scales and component summary measures assessing mental health status. These same tests were conducted to evaluate the empirical validity of the SF-36v2 health domain scales and component summary measures when using data from the 2009 U.S. general population sample.

The following sections summarize the results of the empirical validation of the 2009-normed SF-36v2 when replicating the original "four-group" tests that were used to demonstrate the validity of the SF-36 and 1998-normed SF-36v2 health domain scales and component summary measures in discriminating amongst groups of patients known to differ in the severity of physical and/or mental conditions.

Data sources and methods. The data analyzed in testing the validity of the SF-36v2 health domain scales and component summary measures through known-groups comparisons came from the 2009 U.S. general population (see Chapter 14 for sampling details). Data were collected via a survey that included a checklist of 40 chronic conditions. This checklist required respondents to indicate whether *a doctor had ever told them* that they had any of the following 26 conditions: hypertension, heart attack, congestive heart failure, angina, other heart condition, diabetes, cancer, chronic obstructive pulmonary disease (COPD), allergies, rheumatoid arthritis, osteoarthritis, osteoporosis, kidney disease, liver disease, gastroesophageal reflux disease (GERD), stomach disease, irritable bowel syndrome (IBS), obesity, stroke, HIV/AIDS, anemia, clinical depression, alcohol or drug use, chronic fatigue, migraines, or sleep apnea. Respondents were also asked whether they *currently* had any of the following 14 conditions: chronic allergies, seasonal allergies, back problems, vision problems, hearing problems, arthritis of any kind, skin conditions, asthma, lung problems other than asthma, ulcers, depression, anxiety, severe headaches, or limited use of an arm or leg.

From the responses to this chronic conditions checklist, four mutually exclusive groups were derived to test the validity of SF-12v2 health domain scales and component summary measures. The first group (*Well*) consisted of respondents who reported not having any of 18 specific physical conditions or 3 specific mental conditions from the checklist. The second group (*Physical Only*) consisted of those respondents who reported having one or more of the 18 specific physical conditions and none of the 3 specific mental conditions. The third group (*Mental Only*) consisted of those respondents who reported having one or more of the 3 specific mental conditions and none of the 18 specific physical conditions. Lastly, the fourth group (*Physical + Mental*) consisted of those respondents who reported at least one of the 18 specific physical conditions and at least one of the 3 specific mental conditions.

Analyses. When testing the validity of each SF-36v2 health domain scale and component summary measure in discriminating between groups differing in physical and/or mental health status, separate analyses were conducted on the standard form's and acute form's results. SF-6D results were also included in these analyses.

Tests of the validity of SF-36v2 health domain scales and component summary measures in discriminating between groups differing in *physical health* status consisted of comparisons of mean scores between the *Well* and the *Physical Only* groups and between the *Mental*

Only and the *Physical + Mental* groups. Note that the latter comparison tested the incremental impact of having a physical condition in addition to a mental condition. Tests of the validity of the health domain scales and component summary measures in discriminating between groups differing in *mental health* status consisted of comparisons of mean scores between the *Well* and the *Mental Only* groups and between the *Physical Only* and the *Physical + Mental* groups. Note that the latter comparison tested the incremental impact of having a mental condition in addition to a physical condition.

Unadjusted general linear models were used to estimate mean differences between pairs of clinical groups for each of the health domain scales and component summary measures. The resulting *F* statistic for each scale defines the ratio of between-groups (systematic) variance to within-groups (error) variance. In other words, the greater the *F* ratio, the greater amount of information (systematic variance) a scale provides about a given criterion, relative to error variance. To ensure standardization of the comparisons, the sample size was held constant across health domain scales and component summary measures. By analyzing identical samples across scales and measures for each comparison, the relative size of the *F* ratio thus reflects the relevance of each scale to the specific criterion measure.

Furthermore, *relative validity* (RV) coefficients were estimated for SF-36v2 health domain scales and component summary measures for each form by computing the ratio of pairwise *F* statistics (i.e., the *F* statistic for each comparison scale divided by the *F* statistic for the most valid scale [i.e., the scale with the highest *F* statistic]). The resulting coefficient estimates indicated, in proportional terms, how much less valid each scale was, relative to the most valid scale, as a measure of physical or mental health status.

Hypotheses. A strong theoretical foundation for generating hypotheses makes it easier to draw conclusions about measurement validity (Kerlinger, 1973). To this end, the expected results for a valid measure must be known in advance of each test. Accordingly, based on previous research on the SF-36 (McHorney et al., 1993; Ware, Kosinski, Bayliss, et al., 1995) and the SF-36v2 (Ware et al., 2007), the following results were hypothesized:

- The SF-36v2 health domain scales measuring physical functioning, role limitations due to physical health, bodily pain, and general health (PF, RP, BP, and GH) would be more valid in distinguishing between groups differing in the presence of a physical condition (*Well* versus

Physical Only and *Mental Only* versus *Physical + Mental*) and would be less valid than the mental health scales in distinguishing between groups differing in the presence of a mental condition.

- The SF-36v2 health domain scales measuring mental health, role limitations due to emotional problems, social functioning, and vitality (MH, RE, SF, and VT) would be more valid in distinguishing between groups differing in the presence of a mental condition (*Well* versus *Mental Only* and *Physical Only* versus *Physical + Mental*) and would be less valid than the physical health scales in distinguishing between groups differing in the presence of a physical condition.

Standard form results. Tables 16.5 through 16.8 present the results from the four-group tests involving the SF-36v2's standard form health domain scales, component summary measures, and SF-6D. First, Tables 16.5 and 16.6 summarize the results from the two validity tests regarding differences in physical health status. As hypothesized, the health domain scales and component summary measure that primarily assess physical health (PF, RP, BP, GH, and PCS) were more valid in discriminating between groups of respondents differing in the presence of a physical condition than were the health domain scales and component summary measure that primarily assess mental health (MH, RE, SF, VT, and MCS), as evidenced by the differences in the magnitude of *F* statistics and RV coefficients that were observed across the physical and mental scales. In the validity test involving relatively pure *physical* differences (Table 16.5), the PCS measure was the most valid (RV = 1.00), followed by the PF scale (RV = .72). The RVs for the RP, BP, and GH scales ranged from .49 to .67. Meanwhile, the MH and RE scales yielded small but significant group differences, as well as the smallest RV estimates (.02 and .15, respectively). The RV coefficients for VT and SF, the remaining mental health domain scales, were .21 to .23, respectively, and the RV coefficient for the SF-6D was .50.

In the validity test involving the incremental impact of a physical condition in addition to a mental condition (Table 16.6), the PCS measure was the most valid (RV = 1.00), followed by the BP scale (RV = .80). The RVs for the PF, RP, and GH scales ranged from .60 to .77. While the RV coefficients for the mental health domain scales and component summary measure (MH, RE, SF, VT, and MCS) were considerably lower (RVs < .43) than those for the physical health scales, they all significantly discriminated between the two groups of respondents. The RV coefficient for the SF-6D was .57.

Next, Tables 16.7 and 16.8 summarize the results from the two validity tests regarding differences in mental health status. As hypothesized, the health domain scales and component summary measure that primarily assess mental health (MH, RE, SF, VT, and MCS) were more valid in discriminating between groups of respondents differing in the presence of a mental condition than were the health domain scales and component summary measure that primarily assess physical health (PF, RP, BP, GH, and PCS). In both mental health validity tests (Tables 16.7 and 16.8), the MCS measure was the most valid (RVs = 1.00) discriminator, followed by the MH scale (RVs = .84 and .91, respectively). The RV coefficients for the RE, SF, and VT scales ranged from .43 to .58 in the validity test involving relatively pure *mental* health differences (Table 16.7) and from .48 to .61 in the validity test involving the incremental impact of a mental condition in addition to a physical condition (Table 16.8). In both validity tests, the RV coefficients for the physical health domain scales and component summary measure (PF, RP, BP, GH, and PCS) were considerably lower (RVs < .27) than those for the mental health scales; moreover, the PCS measure failed to significantly discriminate between groups in the validity test of relatively pure mental health differences (Table 16.7). The RV coefficients for the SF-6D were .57 (Table 16.7) and .51 (Table 16.8).

Acute form results. Tables 16.9 through 16.12 present the results from the four-group tests involving the SF-36v2's acute form health domain scales, component summary measures, and SF-6D. First, Tables 16.9 and 16.10 summarize the results from the two validity tests regarding differences in physical health status. As hypothesized, the physical health domain scales and component summary measure (PF, RP, BP, GH, and PCS) were more valid in discriminating between groups of respondents differing in the presence of a physical condition than were the mental health domain scales and component summary measure (MH, RE, SF, VT, and MCS). The PCS measure was the most valid (RVs = 1.00) in both physical validity tests, followed by the PF scale in the validity test involving relatively pure physical differences (RV = .81; Table 16.9) and the BP scale in the validity test involving the incremental impact of a physical condition in addition to a mental condition (RV = .79; Table 16.10). The RV coefficients for the RP, BP, and GH scales ranged from .53 to .58 in the validity test involving relatively pure physical health (Table 16.9), whereas the RV coefficients for the PF, RP, and GH scales ranged from .58 to .71 in the validity test involving the incremental impact of a physical condition in addition to a mental condition (Table 16.10).

Table 16.5

Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical Condition Groups, 2009 U.S. General Population

	Well ^a (N = 1,505)		Physical Only ^b (N = 1,522)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	54.47	6.47	46.22	9.96	8.25	698.37**	1.00
Mental Component Summary	52.97	7.19	52.61	7.76	0.36	1.62	–
Physical Functioning	54.17	6.82	47.06	9.90	7.11	506.20**	.72
Role-Physical	54.32	6.32	47.67	9.78	6.65	470.80**	.67
Bodily Pain	54.46	7.70	47.83	9.03	6.63	452.58**	.65
General Health	54.23	8.19	48.43	8.76	5.80	339.40**	.49
Vitality	53.83	8.69	49.92	8.56	3.91	148.90**	.21
Social Functioning	53.75	6.81	50.12	8.65	3.63	157.58**	.23
Role-Emotional	53.59	6.26	50.81	8.39	2.78	101.86**	.15
Mental Health	53.07	7.86	52.14	7.54	0.93	10.70*	.02
SF-6D	0.81	0.11	0.73	0.11	0.08	346.30**	.50

^aWell group consists of persons who reported not having any of 18 specific physical conditions or 3 specific mental conditions from a checklist of conditions.

^bPhysical Only group consists of persons reported having one or more of 18 specific physical conditions and none of 3 specific mental conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

* $p < .01$.

** $p < .001$.

Table 16.6

Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental and Physical Condition Groups, 2009 U.S. General Population

	Mental Only ^a (N = 236)		Physical + Mental ^b (N = 614)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	53.69	7.50	43.73	11.27	9.96	184.56*	1.00
Mental Component Summary	43.40	11.97	39.63	12.24	3.77	19.14*	.10
Physical Functioning	52.54	7.95	43.40	11.94	9.14	138.69*	.75
Role-Physical	51.73	8.87	42.23	12.07	9.50	141.91*	.77
Bodily Pain	50.76	8.89	41.71	11.09	9.05	148.05*	.80
General Health	49.67	10.16	41.79	10.69	7.88	111.64*	.60
Vitality	46.31	10.52	41.23	9.87	5.08	50.97*	.28
Social Functioning	47.92	10.52	40.39	12.62	7.53	77.52*	.42
Role-Emotional	46.99	11.79	39.90	13.12	7.09	61.22*	.33
Mental Health	43.81	11.39	40.22	11.71	3.59	19.09*	.10
SF-6D	0.70	0.13	0.61	0.13	0.09	104.45*	.57

^aMental Only group consists of persons who reported having one or more of 3 specific mental conditions and none of 18 specific physical conditions from a checklist of conditions.

^bPhysical+Mental group consists of persons who reported at least one of 18 specific physical conditions and at least one of 3 specific mental conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

* $p < .001$.

Meanwhile, the RV coefficients for the mental health domain scales and component summary measure (MH, RE, SF, VT, and MCS) were considerably lower than those for the physical health scales in both the relatively pure physical validity test (RVs < .20; Table 16.9) and the incremental physical validity test (RVs < .22; Table 16.10). The RV coefficients for the SF-6D were .44 (Table 16.9) and .36 (Table 16.10).

Next, Tables 16.11 and 16.12 summarize the results from the two validity tests regarding differences in mental health status. As hypothesized, the mental health domain scales and component summary measure (MH, RE, SF, VT, and MCS) were more valid in discriminating between groups of respondents differing in the presence of a mental condition than were the physical health domain scales and component summary measure (PF, RP,

Table 16.7

Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental Condition Groups, 2009 U.S. General Population

	Well ^a (N = 1,505)		Mental Only ^b (N = 236)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	54.47	6.47	53.69	7.50	0.78	3.46	–
Mental Component Summary	52.97	7.19	43.40	11.97	9.57	355.42*	1.00
Physical Functioning	54.17	6.82	52.54	7.95	1.63	13.52*	.04
Role-Physical	54.32	6.32	51.73	8.87	2.59	36.79*	.10
Bodily Pain	54.46	7.70	50.76	8.89	3.70	54.83*	.15
General Health	54.23	8.19	49.67	10.16	4.56	71.82*	.20
Vitality	53.83	8.69	46.31	10.52	7.52	175.11*	.49
Social Functioning	53.75	6.81	47.92	10.52	5.83	153.10*	.43
Role-Emotional	53.59	6.26	46.99	11.79	6.60	204.79*	.58
Mental Health	53.07	7.86	43.81	11.39	9.26	300.08*	.84
SF-6D	0.81	0.11	0.70	0.13	0.11	203.53*	.57

^aWell group consists of persons who reported not having any of 18 specific physical conditions or 3 specific mental conditions from a checklist of conditions.

^bMental Only group consists of persons who reported having one or more of 3 specific mental conditions and none of 18 specific physical conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

* $p < .001$.

Table 16.8

Comparison of SF-36v2 Standard (4-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical and Mental Condition Groups, 2009 U.S. General Population

	Physical Only ^a (N = 1,522)		Physical + Mental ^b (N = 614)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	46.22	9.96	43.73	11.27	2.49	24.42*	.03
Mental Component Summary	52.61	7.76	39.63	12.24	12.98	829.23*	1.00
Physical Functioning	47.06	9.90	43.40	11.94	3.66	51.36*	.06
Role-Physical	47.67	9.78	42.23	12.07	5.44	113.71*	.14
Bodily Pain	47.83	9.03	41.71	11.09	6.12	169.95*	.20
General Health	48.43	8.76	41.79	10.69	6.64	213.13*	.26
Vitality	49.92	8.56	41.23	9.87	8.69	398.55*	.48
Social Functioning	50.12	8.65	40.39	12.62	9.73	404.87*	.49
Role-Emotional	50.81	8.39	39.90	13.12	10.91	506.15*	.61
Mental Health	52.14	7.54	40.22	11.71	11.92	752.04*	.91
SF-6D	0.73	0.11	0.61	0.13	0.12	422.34*	.51

^aPhysical Only group consists of persons who reported having one or more of 18 specific physical conditions and none of 3 specific mental conditions from a checklist of conditions.

^bPhysical + Mental group consists of persons who reported at least one of 18 specific physical conditions and at least one of 3 specific mental conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

* $p < .001$.

BP, GH, and PCS). The MCS measure was the most valid (RVs = 1.00) in both mental health validity tests, followed by the RE scale in the validity test involving relatively pure mental health differences (RV = .92; Table 16.11) and the MH scale in the validity test involving the incremental impact of a mental condition in addition to a physical condition (RV = .85; Table 16.12). The RV coefficients for the MH, SF, and VT scales ranged from .39 to .70 in the validity test involving relatively pure mental health

(Table 16.11), whereas the RV coefficients for the RE, SF, and VT scales ranged from .46 to .76 in the validity test involving the incremental impact of a mental condition in addition to a physical condition (see Table 16.12). Meanwhile, the RV coefficients for the physical health domain scales and component summary measure (PF, RP, BP, GH, and PCS) were considerably lower than those for the mental health scales in both the relatively pure mental health validity test (RVs < .16; Table 16.11)

and the incremental mental health validity test ($RVs < .32$; Table 16.12). The RV coefficients for the SF-6D were .54 (Table 16.11) and .58 (Table 16.12).

Criterion Validity

Criterion validity offers an indication of the degree to which an individual's score on a given measure or performance on some task predicts his or her performance on another measure or activity, which serves as the criterion (Anastasi, 1988). A criterion may be something that is currently present (e.g., a score on another test, a diagnosis) or something that may happen in the future (e.g., rehospitalization of a discharged patient within the next 6 months). Validation against an existing criterion is referred to as *concurrent validity*, whereas validation against an event that may happen in the future is referred to as *predictive validity*.

Concurrent Validity

Evidence of the SF-36v2's concurrent validity is found through an examination of its relationships with other survey, validation, health care, and background variables that were assessed at the same time during the 2009 norming study using Study Forms A and B (standard form) and Form C (acute form). Correlations of the standard form component summary measures, health domain scales, SET, and SF-6D with other variables from Forms A and B thought, at least generally, to be conceptually related to SF-36v2 variables are presented in Table 16.13. Similarly, correlations of the acute form summary measures, scales, SET, and SF-6D with other variables from Form C thought to be conceptually related to SF-36v2 variables are presented in Table 16.14. Specific relationships were hypothesized between several SF-36v2 measures and external criterion variables, including: (a) VT and sleep problems; (b) MCS and/or MH and each of the three depression-related items (experiencing happiness/satisfaction with life, having interest/pleasure in doing things, feeling down/depressed), the two stress-related items (stress/pressure of daily living, extent stress/pressure has affected health), and excessive drinking; (c) GH and numerical health ratings, work performance issues, and the number and effect of chronic illnesses; (d) BP and the number and effect of chronic illnesses, as well as the number of inpatient and outpatient visits; (e) PCS and/or PF and the number and effect of chronic illnesses; (f) RP and/or RE and work performance issues; and (g) the SF-6D and ratings of overall quality of life, because it reflects almost all the health domains. Evidence of the

concurrent validity for each individual SF-36v2 component summary measure and health domain scale was demonstrated if the hypothesized correlation between the scale or measure and a given criterion variable met or exceeded Cohen's (1998) criterion for a large effect size for product moment correlations ($r \geq .50$).

Examination of the correlations found in Table 16.13 reveals different aspects of the relationships of each SF-36v2 standard form variable with specific validation, health care, and background criterion variables that were administered as part of the 2009 norming study. Overall, the correlations supported most of the hypothesized relationships between the SF-36v2 variables and conceptually related external variables that were assessed at the same time. However, GH, RP, and RE were not as strongly related to work-oriented variables as expected, while the chronic conditions variables were generally found to be related to *all* the SF-36v2 scales and measures. Also notable were weak relationships between MCS and MH scores and the number of occasions a respondent had 5 or more drinks. In addition, only weak to moderately strong relationships were found between the BP scale and inpatient and outpatient visits.

In addition, several other trends or patterns of correlations are worth noting. First, the relationships between each of the external variables and the responses to the SET item were relatively weak, as were the relationships between each SF-36v2 variable and the MOS Sleep-R Snoring and Optimal Sleep subscales. Second, as expected, the three depression-related validation items (rating of happiness/satisfaction, frequency of feeling little interest/pleasure, frequency of feeling down/depressed) and, to a somewhat lesser degree, the stress-related items (stress/pressure of daily living, extent stress/pressure has affected health) and the rating of overall job performance showed stronger relationships with *all* of the mental health variables (MCS, VT, SF, RE, and MH) than with the physical health variables (PCS, PF, RP, BP, and GH). Third, conversely, the physical health variables demonstrated stronger relationships with two of the general health criterion variables (number of chronic conditions ever told he/she had and the highest level of associated limitations) than did the mental health variables.

Fourth, the relationship between the SF-36v2 variables and each of the 6 MOS Sleep-R subscales generally fell below .50, with the lowest correlations across all scales and measures almost exclusively occurring with the Snoring and Optimal Sleep subscales. Fifth, in addition to overall quality of life, the SF-6D exhibited several strong relationships with the criterion variables, the strongest of which were average health

Table 16.9

Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical Condition Groups, 2009 U.S. General Population

	Well ^a (N = 750)		Physical Only ^b (N = 757)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	55.04	6.21	46.45	10.08	8.59	376.7**	1.00
Mental Component Summary	52.89	7.35	53.49	7.26	-0.60	2.46 ^d	-
Physical Functioning	54.76	5.65	47.37	9.84	7.39	304.47**	.81
Role-Physical	54.39	6.32	48.31	9.60	6.08	200.76**	.53
Bodily Pain	54.77	7.44	48.32	8.98	6.45	219.80**	.58
General Health	55.06	7.66	48.63	8.77	6.43	218.31**	.58
Vitality	54.11	8.64	50.26	8.78	3.85	70.14**	.19
Social Functioning	53.74	6.59	51.07	8.03	2.67	47.68**	.13
Role-Emotional	53.47	5.62	52.23	6.63	1.24	14.76**	.04
Mental Health	53.27	7.96	52.42	7.38	0.85	4.31*	.01
SF-6D	0.83	0.12	0.75	0.12	0.08	166.06**	.44

^aWell group consists of persons who reported not having any of 18 specific physical conditions or 3 specific mental conditions from a checklist of conditions.

^bPhysical Only group consists of persons who reported having one or more of 18 specific physical conditions and none of 3 specific mental conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

^dThe difference between the two comparison groups was not statistically significant (n/s).

* $p < .05$.

** $p < .001$.

Table 16.10

Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental and Physical Condition Groups, 2009 U.S. General Population

	Mental Only ^a (N = 110)		Physical + Mental ^b (N = 354)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	55.63	8.19	43.25	11.70	12.38	122.94**	1.00
Mental Component Summary	41.17	13.65	40.46	12.22	0.71	0.31 ^d	-
Physical Functioning	52.97	7.70	43.19	12.28	9.78	71.41**	.58
Role-Physical	52.83	6.94	42.26	12.11	10.57	87.46**	.71
Bodily Pain	51.94	9.56	41.78	10.34	10.16	96.57**	.79
General Health	50.76	9.17	41.26	10.25	9.50	86.84**	.71
Vitality	46.37	10.45	41.65	9.31	4.72	23.41**	.19
Social Functioning	46.96	11.76	40.57	12.40	6.39	26.23**	.21
Role-Emotional	43.84	13.96	40.50	13.61	3.34	5.73*	.05
Mental Health	43.32	11.72	40.99	11.00	2.33	4.19*	.03
SF-6D	0.71	0.12	0.62	0.13	0.09	44.40**	.36

^aMental Only group consists of persons who reported having one or more of 3 specific mental conditions and none of 18 specific physical conditions from a checklist of conditions.

^bPhysical + Mental group consists of persons who reported at least one of 18 specific physical conditions and at least one of 3 specific mental conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

^dThe difference between the two comparison groups was not statistically significant (n/s).

* $p < .05$.

** $p < .001$.

rating, relationships with both sleep problems indices, the three depression-related items (rating of happiness/satisfaction, frequency of feeling little interest/pleasure, frequency of feeling down/depressed), and the number of chronic conditions he/she now has or has ever had and the highest rating of their associated limitations.

Sixth, little to no relationship ($r < .14$; see Cohen, 1988) was found to exist between the drinking variable (number of occasions with 5+ drinks) and any of the SF-36v2 measures and scales. Seventh and lastly, with the exception of the SET item, the relationships between all the SF-36v2 variables and the rating of overall quality

Table 16.11

Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Mental Condition Groups, 2009 U.S. General Population

	Well ^a (N = 750)		Mental Only ^b (N = 110)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	55.04	6.21	55.63	8.19	-0.59	0.93 ^d	-
Mental Component Summary	52.89	7.35	41.17	13.65	11.72	216.58***	1.00
Physical Functioning	54.76	5.65	52.97	7.70	1.79	10.19**	.05
Role-Physical	54.39	6.32	52.83	6.94	1.56	6.69*	.03
Bodily Pain	54.77	7.44	51.94	9.56	2.83	14.96***	.07
General Health	55.06	7.66	50.76	9.17	4.30	33.45***	.15
Vitality	54.11	8.64	46.37	10.45	7.74	84.84***	.39
Social Functioning	53.74	6.59	46.96	11.76	6.78	92.89***	.43
Role-Emotional	53.47	5.62	43.84	13.96	9.63	198.48***	.92
Mental Health	53.27	7.96	43.32	11.72	9.95	152.06***	.70
SF-6D	0.83	0.12	0.71	0.12	0.12	116.88***	.54

^aWell group consists of persons who reported not having any of 18 specific physical conditions or 3 specific mental conditions from a checklist of conditions.

^bMental Only group consists of persons who reported having one or more of 3 specific mental conditions and none of 18 specific physical conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

^dThe difference between the two comparison groups was not statistically significant (n/s).

*p < .05.

**p < .01.

***p < .001.

Table 16.12

Comparison of SF-36v2 Acute (1-Week Recall) Form Health Domain Scales, Component Summary Measures, and Health Utility Index in Discriminating Physical and Mental Condition Groups, 2009 U.S. General Population

	Physical Only ^a (N = 757)		Physical + Mental ^b (N = 354)		Difference	F	RV ^c
	Mean	SD	Mean	SD			
Physical Component Summary	46.45	10.08	43.25	11.70	3.20	21.64*	.05
Mental Component Summary	53.49	7.26	40.46	12.22	13.03	484.50*	1.00
Physical Functioning	47.37	9.84	43.19	12.28	4.18	36.34*	.08
Role-Physical	48.31	9.60	42.26	12.11	6.05	79.66*	.16
Bodily Pain	48.32	8.98	41.78	10.34	6.54	114.42*	.24
General Health	48.63	8.77	41.26	10.25	7.37	150.54*	.31
Vitality	50.26	8.78	41.65	9.31	8.61	220.59*	.46
Social Functioning	51.07	8.03	40.57	12.40	10.50	282.27*	.58
Role-Emotional	52.23	6.63	40.5	13.61	11.73	368.01*	.76
Mental Health	52.42	7.38	40.99	11.00	11.43	411.39*	.85
SF-6D	0.75	0.12	0.62	0.13	0.13	280.22*	.58

^aPhysical Only group consists of persons who reported having one or more of 18 specific physical conditions and none of 3 specific mental conditions from a checklist of conditions.

^bPhysical + Mental group consists of persons who reported at least one of 18 specific physical conditions and at least one of 3 specific mental conditions from a checklist of conditions.

^cRelative validity (RV) of all scales against the scale with the highest F.

*p < .001.

of life were strong, with correlations ranging from $-.50$ (PF) to $-.67$ (GH).

Similar patterns of correlations were seen for the relationships between the SF-36v2 acute form variables and the criterion variables, as shown in Table 16.14.

The evidence presented in this manual on the concurrent validity of SF-36v2 measures and scales is based on

the QualityMetric 2009 Norming Study and should be regarded as supplementary to the vast literature on the validity of the SF-36 and SF-36v2. Further information, based on 1998 general population data, is presented in Ware et al. (2007). In light of the considerable amount of published and forthcoming studies on the validity of the SF-36 and SF-36v2, researchers and clinicians are advised

Table 16.13

Correlations of SF-36v2 Standard (4-Week Recall) Form Variables With Other Survey, Validation, Health Care, and Background Variables, 2009 U.S. General Population

Other Variables	N	SF-36v2 Variables											
		PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH	SET	SF-6D
MOS Sleep-R Standard (4-Week Recall) Form													
Sleep Problems Index I	2,001	.42	.59	.42	.44	.51	.56	.68	.55	.51	.61	-.21	.63
Sleep Problems Index II	1,993	.42	.60	.42	.45	.51	.57	.69	.56	.53	.62	-.21	.64
Sleep Disturbance	2,010	.36	.50	.37	.38	.45	.48	.52	.47	.47	.53	-.16	.55
Snoring	2,000	.16	.17	.14	.13	.21	.21	.20	.16	.17	.17	-.07	.21
Shortness of Breath/Headache	2,022	.35	.44	.39	.41	.39	.40	.41	.49	.46	.45	-.18	.49
Sleep Adequacy	2,022	.26	.47	.24	.26	.37	.43	.61	.37	.33	.49	-.16	.47
Sleep Somnolence	2,013	.45	.45	.44	.47	.45	.51	.60	.49	.44	.45	-.18	.56
Optimal Sleep	2,001	.23	.20	.21	.23	.22	.25	.28	.23	.18	.22	-.07	.25
Validation items													
Days of missed work due to illness/injury during past 4 weeks	2,083	-.23	-.17	-.22	-.29	-.19	-.20	-.13	-.25	-.21	-.16	.09	-.23
Rating of usual job performance in past 1-2 years	2,085	.17	.28	.13	.20	.20	.26	.30	.25	.22	.26	-.12	.29
Rating of overall job performance during past 4 weeks	2,079	.18	.43	.13	.26	.27	.33	.40	.37	.34	.39	-.16	.40
Days in bed due to illness/injury during past 4 weeks	4,002	-.38	-.41	-.40	-.44	-.41	-.38	-.38	-.49	-.47	-.37	.22	-.43
Rating of overall quality of life	4,022	-.51	-.57	-.50	-.53	-.52	-.67	-.60	-.58	-.53	-.60	.28	-.64
Rating of happiness/satisfaction with personal life during past 4 weeks	4,015	-.30	-.65	-.33	-.39	-.41	-.51	-.62	-.53	-.50	-.68	.24	-.59
Frequency of feeling little interest/pleasure in doing things during past 2 weeks	4,001	-.33	-.64	-.38	-.44	-.41	-.45	-.55	-.61	-.55	-.62	.21	-.58
Frequency of feeling down/depressed/hopeless during past 2 weeks	4,007	-.19	-.74	-.26	-.35	-.37	-.41	-.55	-.59	-.58	-.73	.18	-.58
Stress/pressure in daily living experienced during past 4 weeks	4,007	-.09	-.56	-.12	-.19	-.28	-.29	-.49	-.38	-.37	-.57	.16	-.44
Numerical health rating during past 4 weeks, on average	3,996	.54	.48	.51	.55	.52	.62	.53	.57	.49	.50	-.24	.60
Health care items													
BMI	3,937	-.25	-.07	-.25	-.18	-.20	-.24	-.17	-.13	-.12	-.09	.07	-.20
Number of outpatient visits during past 4 weeks	3,998	-.39	-.23	-.37	-.41	-.38	-.30	-.24	-.37	-.33	-.24	.10	-.36
Number of hospital stays during past 12 months	3,995	-.25	-.14	-.26	-.27	-.21	-.20	-.14	-.23	-.22	-.15	.04	-.23
Number of chronic conditions told he/she ever had	4,026	-.58	-.28	-.54	-.53	-.53	-.51	-.41	-.43	-.40	-.31	.15	-.52
Number of chronic conditions told he/she now has	3,145	-.55	-.51	-.52	-.58	-.59	-.57	-.57	-.62	-.52	-.51	.25	-.65
Highest rating of chronic conditions ever had	4,017	-.51	-.41	-.48	-.50	-.56	-.50	-.46	-.49	-.45	-.43	.19	-.57
Highest rating of chronic conditions now has	3,001	-.55	-.52	-.55	-.61	-.60	-.54	-.57	-.64	-.55	-.54	.24	-.67
Background items													
Number of occasions with 5+ drinks during past 4 weeks	1,964	-.02	-.12	-.03	-.03	-.05	-.13	-.04	-.12	-.09	-.13	.05	-.09
Extent stress/pressure has affected health during past 4 weeks	4,016	-.25	-.66	-.29	-.36	-.41	-.44	-.52	-.57	-.54	-.65	.18	-.58

Note. Negative correlations indicate inverse relationships between SF-36v2 scores and responses to criterion variables, reflecting the presence, severity, and/or frequency of thinking, feelings, behaviors, functional status, and/or treatment that would be considered problematic.

to search the literature for studies of validity of the SF-36 and SF-36v2 for their particular population and topic.

Predictive Validity

Correlations of a test or measure with conceptually related non-test variables that occur subsequent to the administration of said test or measure are commonly presented as evidence of predictive validity. As previously discussed, data were collected during the 2009 norming study in two waves, approximately 3 to 4 months apart, with a subsample of 607 study participants completing the same survey form in both waves (see Chapter 14). Note that approximately the same number of participants completed each of the four study forms. Also, data collected in the second wave from this subsample were not included in the main analyses; however, they were used to study the stability of the instruments included in the study forms (see Chapter 15) and the predictive validity of the SF-36v2. Thus, predictive validity was examined by comparing the subsample's observed SF-36v2 scores from the first administration of the 2009 survey form with the subsample's responses to non-SF-36v2 variables from the second administration of the same survey form. The non-test variables that were selected to investigate predictive validity were those from the 2009 study forms that were determined to be conceptually related to the health domain scales and component summary measures, clinically or socially important, and representative of plausible outcomes of the variations found in physical, social, and role functioning; pain; vitality; and mental health.

Table 16.15 summarizes information from Tables 9.5 and 9.14 that is relevant to evaluating the predictive validity of the SF-36v2 standard form when using the 2009 U.S. general population normative data. Due to the criteria that were used to select from the non-test variables available for study, the component summary measures were found to be relevant to each of the chosen variables. As a result, the data presented generally support the predictive validity of the SF-36v2 standard form component summary measures for the variables reported.

As shown in Table 16.15, a perfect relationship existed between decreasing MCS scores at baseline and increasing percentages of respondents reporting feeling down/depressed/hopeless and of those reporting having little interest/pleasure in doing things at reassessment (3–4 months later). Similar but less than perfect relationships were evident between baseline PCS scores and the percentage of respondents reporting one or more outpatient visits and of those reporting one or more bed days due to illness or injury at reassessment. Finally, baseline MCS scores did not appear to be closely associated with

reports of not working at a paying job because of health at reassessment; however, a relationship was noted between decreasing PCS scores at baseline and increasing percentages of respondents reporting not working due to health issues at reassessment.

Table 16.16 summarizes information from Tables 9.35 and 9.43 that is relevant to evaluating the predictive validity of the SF-36v2 acute form when using the 2009 U.S. general population normative data. As with the standard form data, the data presented in Table 16.16 generally support the predictive validity of the SF-36v2 acute form component summary measures for the variables reported.

Content Validity

Content validity refers to the sampling adequacy of the material (or domain) on which individuals are tested or surveyed (Nunnally & Bernstein, 1994). This type of validity is commonly used to evaluate achievement tests (Anastasi, 1988), but it is also important to assess the content validity of other types of measures, including health status surveys. For instruments like the SF-36v2, content validity can be evaluated in terms of the domains that are assessed (as reflected in its scales) and the extent to which important aspects of each individual domain are assessed (as reflected in the items).

Assessment of health status can involve the measurement of several different aspects of functioning. Table 16.17 compares the content of the SF-36v2 with the content of MOS measures that preceded it (both longer and shorter) and with the content of seven other widely used psychometric measures. As shown in this table, the concepts (domains) assessed vary from survey to survey, even amongst some of the more commonly used health status surveys. Table 16.17 also reveals that the SF-36v2 includes eight of the most frequently represented health concepts and identifies those concepts that are not measured by the SF-36v2 but are included in other measures (e.g., the Sickness Impact Profile [SIP], MOS Long Form, and Health Insurance Experiment [HIE] battery), such as sleep, cognitive functioning, and quality of life.

The health concepts or domains that are assessed by a given health status measure reflect several factors, including the survey developers' determination of what is important to include; the desired level of measurement precision; the particular purpose, application, or population for which the survey is designed; and the maximum time required to complete the survey. As noted in Chapter 1, development of the SF-36v2 stemmed from an interest in and a need for a comprehensive, short-form health

Table 16.15

Percentage of Respondents Reporting Subsequent (3–4 Months) Adverse Events by Baseline SF-36v2 Standard (4-Week Recall) Form Component Summary Measure T Scores

Baseline T-Score Range	% feeling down/ depressed/hopeless ^a by MCS	% having little interest/ pleasure in doing things ^b by MCS	% outpatient visits with health professional ^c by PCS	% bed days due to illness/injury ^d by PCS	% not working because of health ^e by PCS	by MCS
60+	0.0	2.6				63.2
55+	8.9 ^f	12.9 ^f	23.5	5.2	26.1	40.6 ^f
50–54.9	30.9	30.9	45.2	11.1	38.4	34.6
45–49.9	55.3	48.7	51.1	18.2	42.2	47.5
40–44.9	67.7	58.1	45.0	14.3	71.4	35.5
35–39.9	68.4	61.1	68.8	25.0	75.0	42.1
< 35	93.8	88.2	81.3 ^g	56.3 ^g	81.3 ^g	41.2
< 30			93.3	46.7	80.0	

^a% reporting feeling down, depressed, or hopeless *several, more than half, or nearly every day* during the 2 weeks preceding survey readministration.

^b% reporting experiencing little interest or pleasure in doing things *several, more than half, or nearly every day* during the 2 weeks preceding survey readministration.

^c% reporting one or more outpatient visits with a health professional during the 4 weeks preceding survey readministration.

^d% reporting one or more days in bed because of illness or injury during the 4 weeks preceding survey readministration.

^e% reporting not working at a paying job because of health at the time of survey readministration.

^fIncludes only those scoring in the 55–59.9 T-score range.

^gIncludes only those scoring in the 30–34.9 T-score range.

Table 16.16

Percentage of Respondents Reporting Subsequent (3–4 Months) Adverse Events by Baseline SF-36v2 Acute (1-Week Recall) Form Component Summary Measure T Scores

Baseline T-Score Range	% feeling down/ depressed/hopeless ^a by MCS	% having little interest/ pleasure in doing things ^b by MCS	% outpatient visits with health professional ^c by PCS	% not working because of health ^d by PCS
60+	4.8	0.0		
55+	12.5 ^e	13.5 ^e	23.8	26.3
50–54.9	18.0	28.2	41.2	32.4
45–49.9	21.1	30.0	42.9	52.4
40–44.9	82.4	55.6	76.9	84.6
35–39.9	75.0	91.7	40.0	90.0
< 35	95.2	81.0	60.0 ^f	100.0 ^f
< 30			87.5	75.0

^a% reporting feeling down, depressed, or hopeless *several, more than half, or nearly every day* during the 2 weeks preceding survey readministration.

^b% reporting experiencing little interest or pleasure in doing things *several, more than half, or nearly every day* during the 2 weeks preceding survey readministration.

^c% reporting one or more outpatient visits with a health professional during the 4 weeks preceding survey readministration.

^d% reporting not working at a paying job because of health at the time of survey readministration.

^eIncludes only those scoring in the 55–59.9 T-score range.

^fIncludes only those scoring in the 30–34.9 T-score range.

survey. The SF-36 was first made available in 1988 in a “developmental” form (Ware, 1988) and then in 1990 in the standard form (i.e., SF-36; Ware et al., 1993). Constructed to satisfy the minimum psychometric standards necessary for group comparisons, the eight health domains represented in the profiles of the SF-36, SF-36v2, and all of other the Short Form instruments were selected from the 40 domains included in the MOS (Stewart & Ware, 1992). Those chosen represent the health domains most frequently measured by other widely used health surveys and those believed to be most

affected by disease and health conditions (Ware, 1995; Ware et al, 1993).

Furthermore, the SF-36v2 items represent multiple operational indicators of health, including behavioral function and dysfunction, distress and well-being, objective reports and subjective ratings, and both favorable and unfavorable self-evaluations of general health status (Ware et al., 1993). Because all the Short Form surveys are generic measures of health, symptoms and problems that are specific to particular conditions are not included in any of the SF instruments (see Chapter 3). Note that

a discussion of the rationale for the selection of SF-36 health domain scale items is presented in Chapter 13, with a more detailed summary of the health phenomena captured by the health domain scales being presented in Table 13.1.

Each Short Form instrument was developed with a recognition of the trade-offs that exist between the breadth of the domains represented in the survey and the depth of the measurement these domains required. Such trade-offs are necessary to arrive at a useful,

psychometrically sound measure that is accepted by both patients and their health care providers. Despite its brevity and limited content coverage, research and feedback available to date indicate that the SF-36v2 provides a comprehensive, valid, and reliable assessment of the most important aspects of health status; is easily implemented; provides information useful for monitoring treatments and assessing the outcomes of said treatments; and is well-accepted by those receiving and providing health care services.

Table 16.17

Summary of Content of Widely Used General Health Surveys

Concept	QWB	SIP	HIE	NHP	MHIQ	COOP	Duke	MOS Long Form	SF-20	SF-36/ SF-36v2
Physical functioning	•	•	•	•	•	•	•	•	•	•
Social functioning	•	•	•	•	•	•	•	•	•	•
Role functioning	•	•	•	•		•	•	•	•	•
Psychological distress		•	•	•	•	•	•	•	•	•
Psychological well-being			•		•		•	•	•	•
Health perceptions			•	•		•	•	•	•	•
Pain			•	•		•	•	•	•	•
Energy/fatigue	•		•	•			•	•		•
Reported health transition						•		•		•
Symptoms/problems (specific)	•		•					•		
Sleep		•		•			•	•		
Cognitive functioning		•					•	•		
Sexual functioning								•		
Health distress			•					•		
Family functioning							•	•		
Self-esteem							•			
Eating		•								
Recreation/hobbies		•								
Communications		•								
Quality of life			•			•		•		

Note. Adapted from Ware, Kosinski, & Gandek (2000) and Ware et al. (2007)

QWB = Quality of Well-Being Scale (Patrick, Bush, & Chen, 1973)

SIP = Sickness Impact Profile (Bergner et al., 1981)

HIE = Health Insurance Experiment (Brook et al., 1979; Ware, Brook, et al., 1980)

NHP = Nottingham Health Profile (Hunt, McKenna, McEwen, Williams, & Papp, 1981)

MHIQ = McMaster Health Index Questionnaire (Chambers, 1988)

COOP = Dartmouth COOP Function Charts (Nelson, Landgraf, Hays, Kirk, et al., 1990)

Duke = Duke Health Profile (Parkerson, Broadhead, & Tse, 1990)

MOS Long Form = MOS 149-item Functional Status and Well-Being Survey (Stewart & Ware, 1992)

SF-20 = SF-20 Health Survey (Stewart, Hays, & Ware, 1988; Ware, Sherbourne, & Davies, 1992)

17

Statistical Power Analysis

Statistical power is the probability that a difference will be found when one exists. It is largely determined by features of the sample design, such as the size of the difference under study (effect size), sample size, number of groups being compared, and how comparison groups were formed. For example, larger differences are easier to detect than smaller ones, and differences of any size are easier to detect with larger samples. Moreover, comparisons of repeated measures between groups are generally more powerful when groups are randomly formed. When determining statistical power, sample size is often a more important factor than measurement error, which is why comparisons between large groups can be successfully performed with less reliable measures (e.g., in the .50–.70 range) than are required for comparisons involving individual scores (Nunnally, 1978).

The psychometric properties of the dependent measures also influence statistical power (Cohen, 1988). One such property is measurement reliability, because “noisy” measures have greater error variance relative to systematic variance, resulting in less statistical power. A scale’s variability influences statistical power because a less precise scale requires a larger sample size to be as effective as a more precise scale, as the detected effect size is reduced by the lack of measurement precision. Overall, better measures usually increase statistical power.

The purpose of this chapter is to assist researchers in determining the minimum sample sizes required to detect various levels of difference in SF-36v2 scores for the eight health domain scales and two component summary measures when employing the 2009 U.S. general population norms. Because the minimum sample size required depends on the type of study being conducted, minimum sample sizes are presented for both experimental and nonexperimental studies. It is recommended that the tables provided in this chapter be consulted when designing studies involving the SF-36v2. Also, researchers are advised to review the discussion of determining

minimally important differences (MID) found in Chapter 10 of this manual.

Statistical Power and *T* Scores

Tables 17.1 through 17.8 provide estimates of the sample sizes necessary to detect average group differences equal to 1, 2, 5, and 10 *T*-score points (i.e., 0.1, 0.2, 0.5, and 1 standard deviation, respectively) for each of the SF-36v2 standard (4-week recall) and acute (1-week recall) form health domain scales and component summary measures. These sample size estimates are based on formulas published by Cohen (1988) and on reliability and variance estimates from the 1998 U.S. general population. They are considered to be appropriate for use with data generated from the use of the 2009 algorithms. Because *T* scores represent a standardized distribution with a mean of 50 and a standard deviation (*SD*) of 10, the variability of each scale is constant and therefore does not influence the effect size that is being detected. Note that the reliability of each scale is still taken into consideration for the adjustment of the sample size.

One advantage of using the *T*-score metric in scoring the SF-36v2 is that differences can be assessed in a more standardized manner by its use of *SD* units, which are constant across scales. For example, with *T* scores, a 1-point average difference between two groups is equivalent to detecting a 0.1 *SD* average difference across the eight health domain scales and the two component summary measures.

Experimental Studies

Tables 17.1 through 17.3 present sample size estimates for two experimental study designs: two randomly formed groups with repeated assessments and two

randomly formed groups with only postintervention assessments. Note that the eight SF-36 health domain scales tend to correlate between .60 and .80 in retest studies (Brazier et al., 1992), making .60 an adequate, albeit conservative, standard considering the gains in reliability that have been reported for the SF-36v2 (Jenkinson, Stewart-Brown, Petersen, & Paice, 1999). As such, Tables 17.1 and 17.3 assume test-retest correlations of .60, while Table 17.2 assumes a correlation of .40. Table 17.1 presents sample size estimates for randomized two-group experiments with repeated measures

Table 17.1

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Between Postintervention Scores of Two Experimental Groups With Preintervention Scores as Covariates (Change Score ANCOVA, Retest Correlation = .60)

	Number of T-Score Points Difference			
	1	2	5	10
PCS	1,059	266	43	12
MCS	1,082	271	44	12
PF	1,071	268	44	12
RP	1,059	266	43	12
BP	1,118	280	46	12
GH	1,212	304	49	13
VT	1,184	297	48	13
SF	1,157	290	47	13
RE	1,082	271	44	12
MH	1,184	297	48	13

Note. Sample size requirements for each scale were adjusted by their respective measurement reliabilities. Estimates assume alpha = .05, two-tailed test, power = 80% (Cohen, 1988), and a test-retest correlation of .60.

Table 17.2

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Between Postintervention Scores of Two Experimental Groups With Preintervention Scores as Covariates (Change Score ANCOVA, Retest Correlation = .40)

	Number of T-Score Points Difference			
	1	2	5	10
PCS	1,368	343	57	15
MCS	1,397	350	58	15
PF	1,382	346	57	15
RP	1,368	343	57	15
BP	1,444	362	60	16
GH	1,565	392	65	17
VT	1,528	383	63	17
SF	1,493	374	62	16
RE	1,397	350	58	15
MH	1,528	383	63	17

Note. Sample size requirements for each scale were adjusted by their respective measurement reliabilities. Estimates assume alpha = .05, two-tailed test, power = 80% (Cohen, 1988), and a test-retest correlation of .40.

assuming a correlation of .60. As shown in the table, a 2-point difference (0.2 *SD* unit) in PF is detectable with 268 subjects per group, compared to 12 subjects per group to detect a difference of 10 points (1 *SD* unit). For intervention studies in which test-retest correlations are expected to be less than .60 for the experimental sample, sample size estimates were computed using a correlation of .40 and are presented in Table 17.2.

Table 17.3 presents sample size estimates for comparisons between two experimental groups with only postintervention outcomes assessed. Comparisons of the sample sizes found in Tables 17.3 and 17.1 reveal that gains in statistical power can be made when using a repeated measures experimental design, relative to one with only postintervention measures. Specifically, compared to a repeated measures design, about 10% more subjects are required to detect the same score difference when using a postintervention measure design. For example, 1,672 subjects are required to detect the smallest PF *T*-score point difference (0.1 *SD* unit) when using a postintervention measures design (Table 17.3), compared with just 1,071 subjects when using a repeated measures design (Table 17.1).

Nonexperimental Studies

Tables 17.4 through 17.8 present sample size estimates for three nonexperimental comparisons involving the SF-36v2 health domain scales and component summary measures: (a) comparisons between two self-selected groups with pre- and postintervention survey administra-

Table 17.3

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Between Two Experimental Groups, Postintervention Scores Only (ANOVA, Retest Correlation = .60)

	Number of T-Score Points Difference			
	1	2	5	10
PCS	1,655	414	67	18
MCS	1,690	423	69	18
PF	1,672	419	68	18
RP	1,655	414	67	18
BP	1,747	437	71	18
GH	1,894	474	77	20
VT	1,849	463	75	19
SF	1,807	452	73	19
RE	1,690	423	69	18
MH	1,849	463	75	19

Note. Sample size requirements for each scale were adjusted by their respective measurement reliabilities. Estimates assume alpha = .05, two-tailed test, power = 80% (Cohen, 1988), and a test-retest correlation of .60.

Table 17.4

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Between Two Self-Selected Groups, Repeated Measures Design (Retest Correlation = .60)

	Number of T-Score Points Difference			
	1	2	5	10
PCS	1,315	330	55	15
MCS	1,343	337	56	15
PF	1,329	334	55	15
RP	1,315	330	55	15
BP	1,388	348	57	16
GH	1,504	378	62	17
VT	1,469	369	61	17
SF	1,435	360	59	16
RE	1,343	337	56	15
MH	1,469	369	61	17

Note. Sample size requirements for each scale were adjusted by their respective measurement reliabilities. Estimates assume alpha = 0.05, two-tailed test, power = 80% (Cohen, 1988), and a test-retest correlation of .60.

Table 17.5

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Between Two Self-Selected Groups, Repeated Measures Design (Retest Correlation = .40)

	Number of T-Score Points Difference			
	1	2	5	10
PCS	1,986	498	81	22
MCS	2,029	509	83	22
PF	2,008	503	82	22
RP	1,986	498	81	22
BP	2,097	526	86	23
GH	2,273	570	93	25
VT	2,220	556	91	24
SF	2,169	544	89	24
RE	2,029	509	83	22
MH	2,220	556	91	24

Note. Sample size requirements for each were adjusted by their respective measurement reliabilities. Estimates assume alpha = .05, two-tailed test, power = 80% (Cohen, 1988), and a test-retest correlation of .40.

tions (Tables 17.4 and 17.5), (b) repeated measures over time for a single group (Tables 17.6 and 17.7), and (c) comparisons between a group mean score and a fixed score, such as general population norms (Table 17.8).

The sample size estimates for nonexperimental, two-group studies with repeated measures (correlation = .60; see Table 17.4) assume that score differences will be analyzed to maximize the internal validity of the study design. Note that differences between the sample sizes presented in Tables 17.4 and 17.1 illustrate the power gained from using an experimental versus a nonexperimental two-group comparison, with nearly a 20% gain in power for the smallest scale score difference (1 point). Just as for experimental intervention studies,

Table 17.6

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Over Time Within One Group (Retest Correlation = .60)

	Number of T-Score Points Difference			
	1	2	5	10
PCS	657	165	27	8
MCS	671	169	28	8
PF	664	167	28	8
RP	657	165	27	8
BP	694	174	29	8
GH	752	189	31	9
VT	735	184	30	8
SF	718	180	30	8
RE	671	169	28	8
MH	735	184	30	8

Note. Sample size requirements for each scale were adjusted by their respective measurement reliabilities. Estimates assume alpha = .05, two-tailed test, power = 80% (Cohen, 1988), and a test-retest correlation of .60.

sample size estimates for self-selected groups were also computed using a reliability of .40 for studies in which test-retest correlations are expected to be less than .60. These estimates are presented in Table 17.5.

Table 17.6 presents the sample sizes required to detect score differences over time within one group, assuming a test-retest correlation of .60. As the table illustrates, these sample sizes are smaller than the sample sizes required for the other study designs. However, the results obtained from comparing scores over time within one group are more difficult to interpret than the results from study designs comparing scores between two groups receiving different interventions (Cook & Campbell, 1979). As with the other study designs, sample size estimates were also computed to detect differences in SF-36v2 scores over time within one group using an expected test-retest reliability of .40 and are presented in Table 17.7.

Table 17.8 presents estimates of the sample sizes required to compare average health domain scale and component summary measure scores to a fixed norm, such as the general population. As illustrated in Table 17.8, a difference of 5 *T*-score points (0.5 *SD* unit) on the PF scale can be detected with just 34 subjects in the sample, compared with 210 subjects needed for a detectable difference of 2 *T*-score points (0.2 *SD* unit).

Statistical Power and Scale Measurement Properties

The sample size recommendations provided in this chapter are based on the assumption that the SF-36v2 scales and measures can be analyzed using standard

Table 17.7

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Over Time Within One Group (Retest Correlation = .40)

	Number of T-Score Points Difference			
	1	2	5	10
PCS	993	249	41	11
MCS	1,015	254	42	11
PF	1,004	252	41	11
RP	993	249	41	11
BP	1,048	263	43	11
GH	1,137	285	46	12
VT	1,110	278	45	12
SF	1,084	272	44	12
RE	1,015	254	42	11
MH	1,110	278	45	12

Note. Sample size requirements for each scale were adjusted by their respective measurement reliabilities. Estimates assume alpha = .05, two-tailed test, power = 80% (Cohen, 1988), and a test-retest correlation of .40.

methods for continuous data. Such sample size recommendations were also provided in the SF-36 manual (Ware, Snow, Kosinski, & Gandek, 1993) and an earlier edition of the SF-36v2 manual (Ware et al., 2007). Note that the assumptions underlying these recommendations were questioned by some statisticians (Julious, George, & Campbell, 1995), who suggested that sample size estimation should be based on the assumption that the scales are ordered categorical variables, particularly for scales with relatively few levels (possible discrete values) like the SF-36's role-functioning scales (see also Machin & Fayes, 1998; Prieto, Alonso, & Anto, 1996; Walters, 2004). Such alternative methods for sample size calculations have been

Table 17.8

Sample Sizes Needed to Detect SF-36v2 Standard (4-Week Recall) or Acute (1-Week Recall) Form SD-Unit Differences Between a Group Mean and a Fixed Norm

	Number of T-Score Points Difference			
	1	2	5	10
PCS	828	208	34	9
MCS	846	212	35	9
PF	837	210	34	9
RP	828	208	34	9
BP	874	219	36	10
GH	947	238	39	10
VT	925	232	38	10
SF	904	227	37	10
RE	846	212	35	9
MH	925	232	38	10

Note. Sample size requirements for each scale were adjusted by their respective measurement reliabilities. Estimates assume alpha = .05, two-tailed test, and power = 80% (Cohen, 1988).

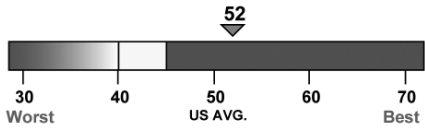
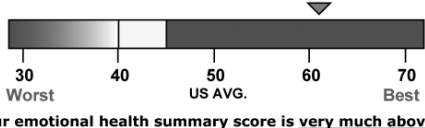
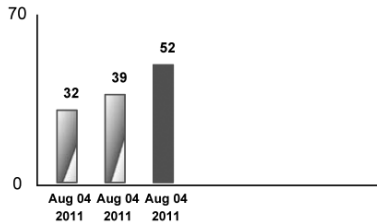
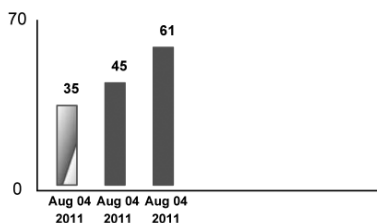
recommended for scales with less than seven levels and/or large floor or ceiling effects (Walters, 2004). However, in regard to the SF-36v2, all scales have more than seven levels and floor and ceiling effects have been substantially reduced compared to the SF-36. Furthermore, subsequent comparative studies of data analysis using methods for continuous data and bootstrap methods for categorical data have shown that categorical data methods provide results that are very similar to standard methods with regard to sample size estimates, standard errors, confidence intervals, and statistical tests (Walters & Campbell, 2004, 2005). These results support the approach to sample size estimation presented here.

Appendix A

Sample SF-36v2 Individual Respondent Reports

Appendix A.1 Sample SF-36v2 Member Report

SF-36v2® Health Survey: United States (English)
Member Report Report for *Jim Smith*
August 04, 2011

YOUR RESULTS	YOUR HISTORY
<p>Survey Date: August 04, 2011 Mode: eForm Age: 41 Gender: Male Conditions: None</p> <p>PHYSICAL HEALTH SUMMARY</p>  <p>Your physical health summary score is about average, even taking into account the margin of error.</p> <p>MENTAL HEALTH SUMMARY</p>  <p>Your emotional health summary score is very much above average, even taking into account the margin of error.</p>	<p>Your PCS Score History:</p>  <p>Your MCS Score History:</p> 
<p>Well below average Below Average At or above average</p>	

PROGRESS	INTERPRETATION												
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Date</th> <th>Physical Health</th> <th>Mental Health</th> </tr> </thead> <tbody> <tr> <td>Current: 08/04/2011</td> <td>52</td> <td>61</td> </tr> <tr> <td>Previous: 08/04/2011</td> <td>39</td> <td>45</td> </tr> <tr> <td>Change (current/previous):</td> <td>Better</td> <td>Better</td> </tr> </tbody> </table> <ul style="list-style-type: none"> Your physical and mental health summary scores changed significantly compared to the last time the survey was taken. Be sure to mention this to your doctor. 	Date	Physical Health	Mental Health	Current: 08/04/2011	52	61	Previous: 08/04/2011	39	45	Change (current/previous):	Better	Better	<p>Based on your answers about health in the past 4 weeks, our research shows that</p> <p>Compared to the general population ...</p> <p>Physically, your ...</p> <ul style="list-style-type: none"> pain is much less functioning is better than most performance of work, home or school activities is the same or better <p>Emotionally ...</p> <ul style="list-style-type: none"> bothered less than most performance of work, home and school activities is limited less <p>Overall, your ...</p> <ul style="list-style-type: none"> rating of your health is much better participation in social activities is less limited energy level is much higher <p>Compared to other men of similar age...</p> <ul style="list-style-type: none"> your physical health appears to be about the same your emotional health appears to be better;
Date	Physical Health	Mental Health											
Current: 08/04/2011	52	61											
Previous: 08/04/2011	39	45											
Change (current/previous):	Better	Better											
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>Date</th> <th>Physical Health</th> <th>Mental Health</th> </tr> </thead> <tbody> <tr> <td>Current: 08/04/2011</td> <td>52</td> <td>61</td> </tr> <tr> <td>Baseline: 08/04/2011</td> <td>32</td> <td>35</td> </tr> <tr> <td>Change (current/baseline):</td> <td>Better</td> <td>Better</td> </tr> </tbody> </table> <ul style="list-style-type: none"> Your physical and mental health summary scores changed significantly compared to the first time the survey was taken. Be sure to mention this to your doctor. <p><i>The margin of error can cause small changes in scores. This survey focuses on changes of 5 points or more. However you should always report to your doctor any changes that are important to you.</i></p>	Date	Physical Health	Mental Health	Current: 08/04/2011	52	61	Baseline: 08/04/2011	32	35	Change (current/baseline):	Better	Better	
Date	Physical Health	Mental Health											
Current: 08/04/2011	52	61											
Baseline: 08/04/2011	32	35											
Change (current/baseline):	Better	Better											

Note: This survey is not a diagnostic tool. It is intended to supplement clinical decision making. Visit us at www.amihealthy.com for more information.

Appendix A.2 Sample SF-36v2 Member Report with PIQ-6



SF-36v2® Health Survey with PIQ-6™


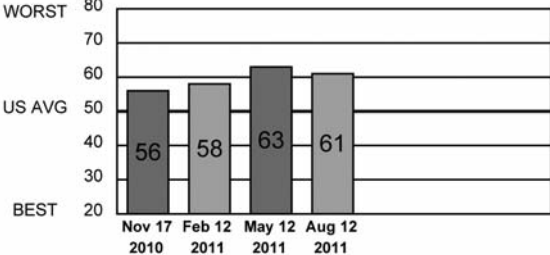
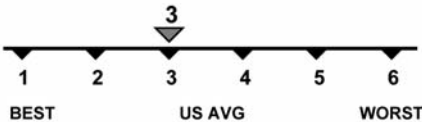
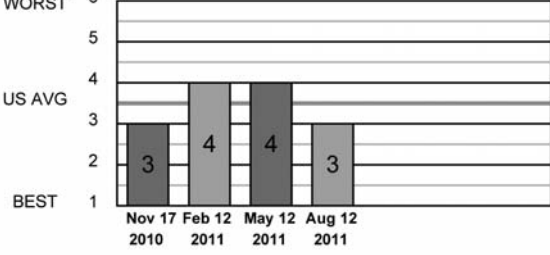
Member Report

Member Report for *Jim Smith*



Back to Member Report

September 23, 2011

PATIENT RESULTS	PATIENT HISTORY										
<p>Survey Date: August 12, 2011 Mode: eForm Age: 33 Gender: Male</p> <p>PAIN IMPACT 61</p>  <p>20 30 40 50 60 70 80 BEST US AVG WORST</p> <p>You have substantial pain impact, taking into account the margin of error.</p>	<p>Pain Impact Score History</p>  <table border="1"> <tr><th>Date</th><th>Score</th></tr> <tr><td>Nov 17 2010</td><td>56</td></tr> <tr><td>Feb 12 2011</td><td>58</td></tr> <tr><td>May 12 2011</td><td>63</td></tr> <tr><td>Aug 12 2011</td><td>61</td></tr> </table>	Date	Score	Nov 17 2010	56	Feb 12 2011	58	May 12 2011	63	Aug 12 2011	61
Date	Score										
Nov 17 2010	56										
Feb 12 2011	58										
May 12 2011	63										
Aug 12 2011	61										
<p>PAIN SEVERITY 3</p>  <p>1 2 3 4 5 6 BEST US AVG WORST</p> <p>Your Pain Severity is mild, taking into account the margin of error.</p>	<p>Pain Severity Score History</p>  <table border="1"> <tr><th>Date</th><th>Score</th></tr> <tr><td>Nov 17 2010</td><td>3</td></tr> <tr><td>Feb 12 2011</td><td>4</td></tr> <tr><td>May 12 2011</td><td>4</td></tr> <tr><td>Aug 12 2011</td><td>3</td></tr> </table>	Date	Score	Nov 17 2010	3	Feb 12 2011	4	May 12 2011	4	Aug 12 2011	3
Date	Score										
Nov 17 2010	3										
Feb 12 2011	4										
May 12 2011	4										
Aug 12 2011	3										

Note: This survey is not a diagnostic tool. It is intended to supplement clinical decision making. Visit us at www.amihealthy.com for more information.

Appendix B


Sample SF-36v2 Group-Level Reports



Appendix B.1 Sample SF-36v2 SF Comparison for Total Sample Report

Demo Study 02 - SF Comparison for Total Sample - All Timepoint Values

Report Type: [SF Comparison for Total Sample](#)
 Survey: SF-36v2® Health Survey

Report Date: 2/22/2011



Print this report  Instructional Guide 

Report Criteria

Date Range: 1/11/2011 - 2/10/2011

Timepoint: All Timepoints

Demographic Profile

Sample Size: 296

Male (*): 53% (78)

Female (*): 47% (70)

Mean Age: 55

Age Range: 17 - 90

*Of those reporting gender (n=148)

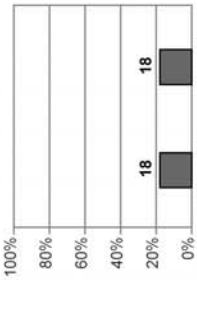
Comments

Scores for Total Sample

Component	Score
PCS	48.3
MCS	50.8
PF	48
RP	48.38
BP	49.48
GH	49.3
VT	50.49
SF	49.76
RE	48.66
MH	51

Sample: 48.3 50.8 48 48.38 49.48 49.3 50.49 49.76 48.66 51

First Stage Positive Depression Screening: % at Risk

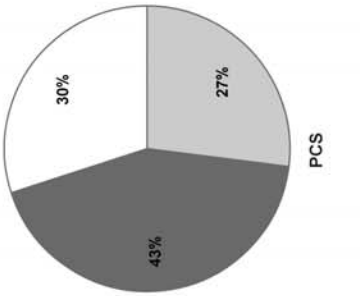


Component	% at Risk
PCS	18
MH	18

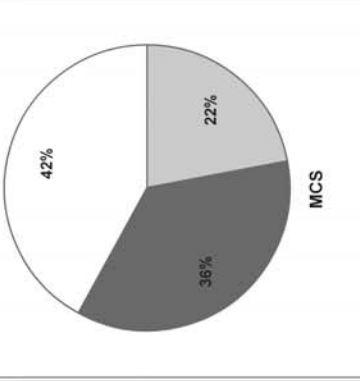
Abbreviation

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

% Sample whose Scores are Above, At or Below the General Population Norm

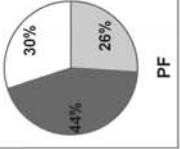


PCS

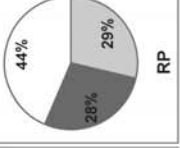


MCS

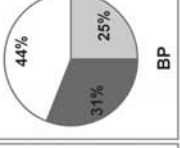
PCS Scale Scores:



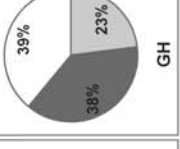
PF



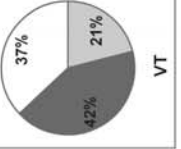
RP



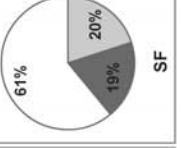
BP



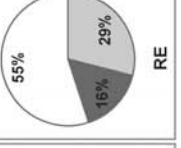
GH



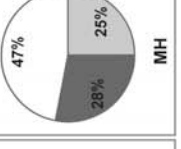
VT



SF

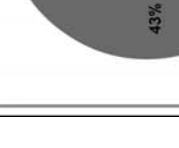


RE

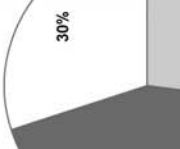


MH

MCS Scale Scores:



PCS


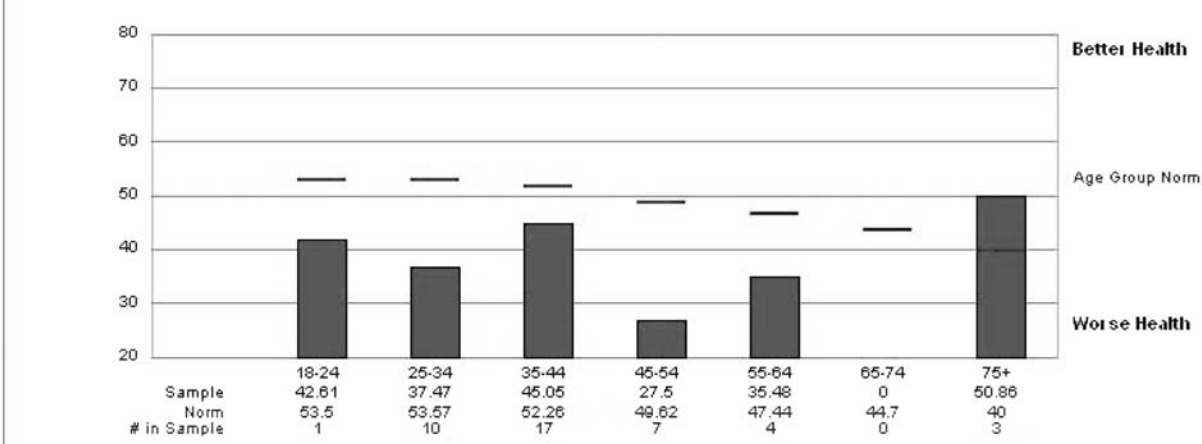
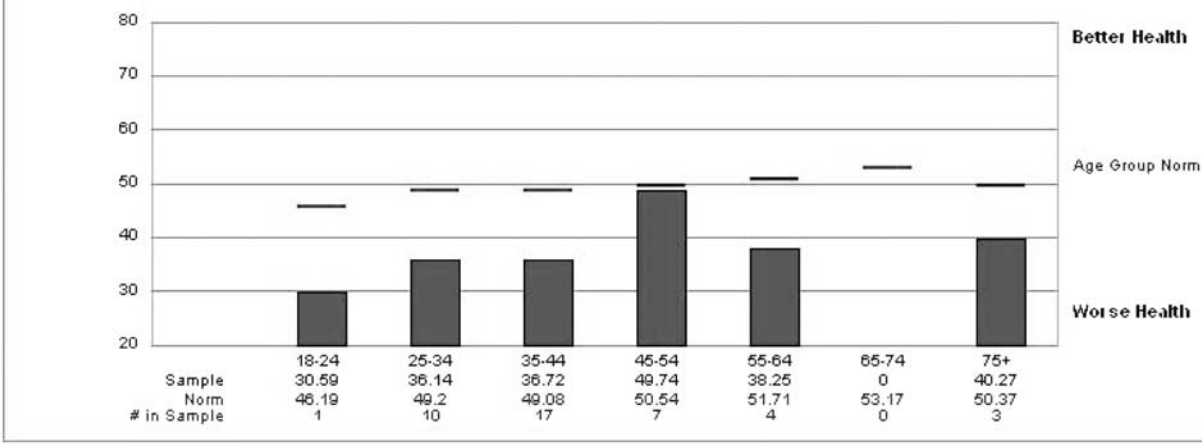


MCS

Note: Totals may not equal 100%, due to statistical rounding.

Copyright © 1999-2011 QualityMetric Incorporated

Appendix B.2 Sample SF-36v2 Scores by Age Group Report

Scores by Age Group Report		Report Date: March 18, 2010																																
Report Type: <u>Scores by Age Group Report</u>																																		
Survey: <u>SF-36v2 Health Survey Standard Version (Advanced Reporting)</u>																																		
Report Criteria Date Range: 01/01/2005 - 05/02/2010 Group: uugroup Site: All Sites Sample Includes: All Surveys	Demographic Profile Sample Size: 43 Male (*): 51% (22) Female (*): 49% (21) Mean Age: 42 Age Range: 12 - 108 *Of those reporting gender (n=43)	Comments SF-36v2 Population Age Sample report - March 2010.																																
PCS Scores by Age Group																																		
 <table border="1" style="width: 100%; margin-top: 10px;"> <thead> <tr> <th>Age Group</th> <th>Sample</th> <th>Norm</th> <th># in Sample</th> </tr> </thead> <tbody> <tr> <td>18-24</td> <td>42.61</td> <td>53.5</td> <td>1</td> </tr> <tr> <td>25-34</td> <td>37.47</td> <td>53.57</td> <td>10</td> </tr> <tr> <td>35-44</td> <td>45.05</td> <td>52.26</td> <td>17</td> </tr> <tr> <td>45-54</td> <td>27.5</td> <td>49.62</td> <td>7</td> </tr> <tr> <td>55-64</td> <td>35.48</td> <td>47.44</td> <td>4</td> </tr> <tr> <td>65-74</td> <td>0</td> <td>44.7</td> <td>0</td> </tr> <tr> <td>75+</td> <td>50.86</td> <td>40</td> <td>3</td> </tr> </tbody> </table>			Age Group	Sample	Norm	# in Sample	18-24	42.61	53.5	1	25-34	37.47	53.57	10	35-44	45.05	52.26	17	45-54	27.5	49.62	7	55-64	35.48	47.44	4	65-74	0	44.7	0	75+	50.86	40	3
Age Group	Sample	Norm	# in Sample																															
18-24	42.61	53.5	1																															
25-34	37.47	53.57	10																															
35-44	45.05	52.26	17																															
45-54	27.5	49.62	7																															
55-64	35.48	47.44	4																															
65-74	0	44.7	0																															
75+	50.86	40	3																															
MCS Scores by Age Group																																		
 <table border="1" style="width: 100%; margin-top: 10px;"> <thead> <tr> <th>Age Group</th> <th>Sample</th> <th>Norm</th> <th># in Sample</th> </tr> </thead> <tbody> <tr> <td>18-24</td> <td>30.59</td> <td>46.19</td> <td>1</td> </tr> <tr> <td>25-34</td> <td>36.14</td> <td>49.2</td> <td>10</td> </tr> <tr> <td>35-44</td> <td>36.72</td> <td>49.08</td> <td>17</td> </tr> <tr> <td>45-54</td> <td>49.74</td> <td>50.54</td> <td>7</td> </tr> <tr> <td>55-64</td> <td>38.25</td> <td>51.71</td> <td>4</td> </tr> <tr> <td>65-74</td> <td>0</td> <td>53.17</td> <td>0</td> </tr> <tr> <td>75+</td> <td>40.27</td> <td>50.37</td> <td>3</td> </tr> </tbody> </table>			Age Group	Sample	Norm	# in Sample	18-24	30.59	46.19	1	25-34	36.14	49.2	10	35-44	36.72	49.08	17	45-54	49.74	50.54	7	55-64	38.25	51.71	4	65-74	0	53.17	0	75+	40.27	50.37	3
Age Group	Sample	Norm	# in Sample																															
18-24	30.59	46.19	1																															
25-34	36.14	49.2	10																															
35-44	36.72	49.08	17																															
45-54	49.74	50.54	7																															
55-64	38.25	51.71	4																															
65-74	0	53.17	0																															
75+	40.27	50.37	3																															
Abbreviation PCS = Physical Component Summary MCS = Mental Component Summary PF = Physical Functioning RP = Role Physical GH = General Health BP = Bodily Pain VT = Vitality SF = Social Functioning RE = Role Emotional MH = Mental Health		Print this report Instructional Guide																																
Copyright © 1999-2010 QualityMetric Incorporated																																		


Appendix B.3 Sample SF-36v2 Scores by Gender Report


Demo Study 02 - Scores By Gender - All Timepoint Values

Report Type: Scores By Gender

Survey: SF-36v2® Health Survey

Report Date: 2/22/2011



Print this report  [Instructional Guide](#)

Report Criteria

Date Range: 1/11/2011 - 2/10/2011

Timepoint: All Timepoints

Demographic Profile

Sample Size: 296

Male (*): 53% (78)

Female (*): 47% (70)

Mean Age: 55

Age Range: 17 - 90

*Of those reporting gender (n=148)

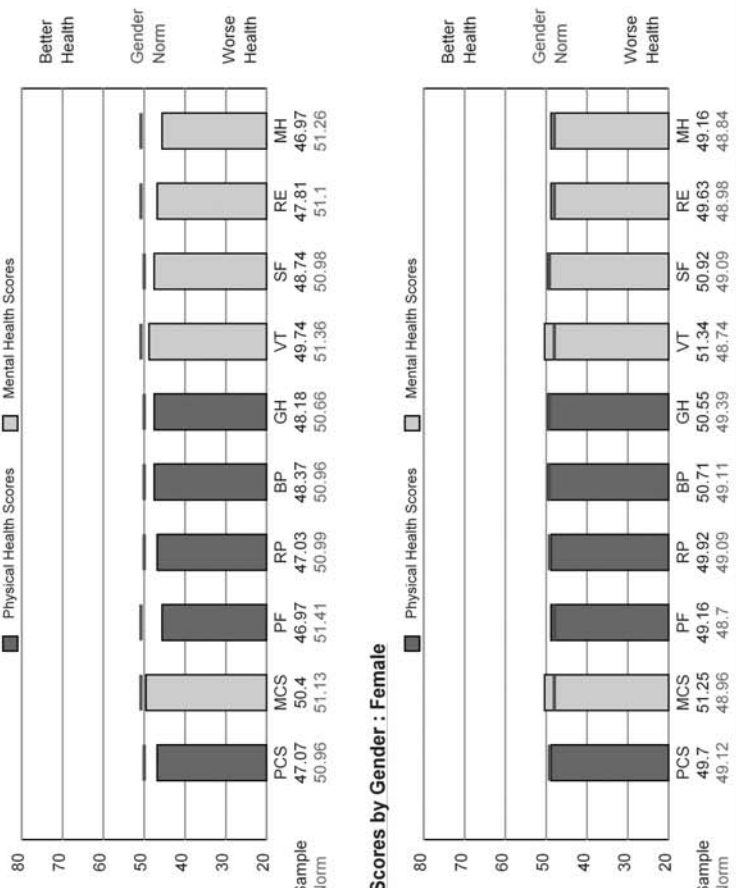
Comments

Scores by Gender : Male

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
Sample	47.07	50.4	46.97	47.03	48.37	48.18	49.74	48.74	47.81	46.97
Norm	50.96	51.13	51.41	50.99	50.96	50.66	51.36	50.98	51.1	51.26

Scores by Gender : Female

	PCS	MCS	PF	RP	BP	GH	VT	SF	RE	MH
Sample	49.7	51.25	49.16	49.92	50.71	50.55	51.34	50.92	49.63	49.16
Norm	49.12	48.96	48.7	49.09	49.11	49.39	48.74	49.09	49.98	48.84

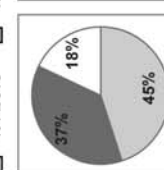


PCS and MCS summary by Gender

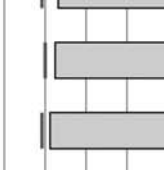
% of Respondents that score Above, At or Below the normal range for their profile.

% Above % At % Below

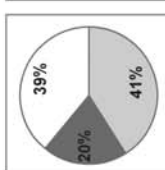
PCS - Male



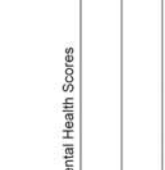
MCS - Male



PCS - Female

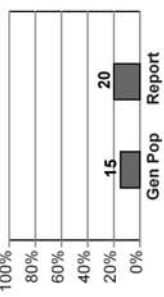


MCS - Female

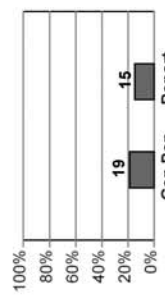


First Stage Positive Depression Screening: % at Risk

Male



Female



Note: Totals may not equal 100%, due to statistical rounding.

Abbreviation

BP = Bodily Pain

VT = Vitality

SF = Social Functioning

RE = Role Emotional

MH = Mental Health

Copyright © 1999-2011 QualityMetric Incorporated

Appendix B.4 Sample SF-36v2 Data Quality Evaluation Report

QualityMetric SF-36v2 Data Quality Evaluation Report - Demo Study EZ01-1234

[Print this report](#) | [View Scoring Summary Report](#) | [View MDE Report](#)

Project Information:

Project Name..... Demo Study EZ01-1234
 MDE Method Used..... Maximum Data Recovery
 Total Number of Records in Project..... 294

This evaluation includes the Health Transition (HT) item.

Data Quality Indicators appear to be satisfactory.

Data Quality Indicators:		Satisfactory	Norms
1. Completeness of Data..... Items with 5% or more missing values: NONE	98.1%	YES	90
2. Responses within Range..... Items with 5% or more out-of-range values: NONE	100.0%	YES	100
3. Consistent Responses.....	92.5%	YES	90
4. Estimable Scale Scores.....			
Estimable without MDE	91.5%	YES	90
Estimable with Half-Scale MDE	98.4%		
Estimable with Full MDE	99.4%		
5. Item Internal Consistency Items that <u>failed</u> internal consistency test: GH04	97.1%	YES	90
6. Discriminant Validity Items that <u>failed</u> discriminant validity test: GH02 MH03	99.2%	YES	80
7. Reliable Scales Scales that <u>failed</u> reliability criteria: NONE	100.0%	YES	100

Definition of Data Quality Indicators:

1. Percentage of completed responses (within range) divided by the total possible number of responses (items*N). This calculation includes the Health Transition (HT) item.
2. Percentage of item responses within the range of response codes printed on the questionnaire. This calculation includes the Health Transition (HT) item.
3. Percentage of subjects with no inconsistent responses on the Response Consistency Index (score = 0).
4. Percentage of subjects for whom all scales are computable with and without application of the SF-MDE
5. Percentage of items that correlated (corrected for overlap) 0.40 or higher with their hypothesized scale.
6. Percentage of items that correlated significantly higher with their hypothesized scale than with competing scales. score.
7. Percentage of scales with Cronbach's Alpha coefficients greater than or equal to 0.70.

Appendix B.5 Sample SF-36v2 Summary Report of Scale and Summary Measure Scores

QualityMetric SF-36v2 Summary Report of Scale and Summary Measure Scores - ABC-1234

[Print this report](#) | [View DOE Report](#) | [View MDE Report](#)

Project Information:

Project Name..... ABC-1234
 MDE Method Used..... Maximum Data Recovery

Table 1. SF-36 V2 Scale and Summary Measure Scores, Norm-Based Scoring

	<u>Scales</u>								<u>Summaries</u>	
	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS
Mean	49.07	49.06	49.70	51.76	50.84	50.44	49.30	51.62	49.40	51.24
25th Percentile	44.15	43.68	42.64	48.43	46.66	47.31	45.72	46.94	43.96	46.83
50th Percentile (median)	53.71	54.16	51.51	53.19	52.60	57.34	56.17	53.48	52.96	53.91
75th Percentile	57.54	57.16	55.55	59.37	58.54	57.34	56.17	58.72	56.50	57.66
Standard Deviation	10.73	10.80	9.71	10.72	9.68	10.63	10.67	9.40	10.60	9.82
Min	19.26	21.23	21.68	18.95	22.89	17.23	14.39	14.24	13.21	14.05
Max	57.54	57.16	62.00	66.50	70.42	57.34	56.17	63.95	65.70	69.62
N	150	149	150	150	149	149	149	148	149	148

Appendix B.6 Sample SF-36v2 Missing Score Estimation Report

QualityMetric SF-36v2 Missing Data Estimation Report - ABC-1234

[Print this report](#) | [View DOE Report](#) | [View Scoring Summary Report](#)

Project Information:

Project Name..... ABC-1234
 MDE Method Used..... Maximum Data Recovery

Table 1. SF-36V2 Missing Data Estimation Report

	Scales									Summaries	
	Total Summary	PF	RP	BP	GH	VT	SF	RE	MH	PCS	MCS
Total Number of Records in Project	150	150	150	150	150	150	150	150	150	150	150
Total Number of Records Scored with Complete Data	116	135	144	150	144	142	146	142	138	116	116
Total Number of Records Scored with Half-Scale MDE	144	150	148	150	148	149	149	146	148	144	144
Total Number of Records Scored with Full MDE	148	150	149	150	150	149	149	149	148	149	148

[Print this report](#) | [View DOE Report](#) | [View Scoring Summary Report](#)

Appendix C

SF-36v2 Score Estimation Using Item Response Theory (IRT)

This appendix describes how item response theory (IRT) can be used to estimate a score on the Physical Functioning (PF) scale when only a few of the PF items have been answered. Generally speaking, IRT provides a statistical model of the relationship between a respondent's answer to a multiple-choice question and his or her overall score on the construct being measured. With regard to health assessment, the latent (i.e., true but unobservable) health status variable is the independent variable, which determines the probability of respondents choosing each of the questionnaire response categories.

Figure C.1 illustrates predictions based on an IRT model—the *partial credit model* (Masters, 1982; Masters & Wright, 1997)—for the three response choices offered with Item 3d (i.e., limitations in climbing flights of stairs). In this figure, the horizontal axis represents physical functioning as it would be measured by an ideal instrument (i.e., latent physical functioning). In line with the standard *T*-score metric, this axis was calibrated such that physical functioning has a mean score of 50 and a

standard deviation of 10 in the U.S. general population. (Note that many other IRT applications set the mean to 0 and the standard deviation to 1.) The three curves shown in Figure C.1 show the probability of selecting each response choice, at each level of physical functioning. For example, according to the model, a respondent with the score of 50 (i.e., the U.S. population average) has a .65 probability of choosing *not limited*, a .34 probability of choosing *limited a little*, and a .01 probability of choosing *limited a lot*. For respondents functioning at higher levels (i.e., above a *T* score of 70), the probability of choosing *not limited* approaches 1.0, or 100%. Conversely, for those with very low levels of physical functioning (i.e., below a *T* score of 20), the probability of choosing *limited a lot* approaches 1.0, or 100%. The curves illustrated in Figure C.1, referred to as *item characteristic curves* or *option characteristic curves*, define item characteristics that hold true regardless of the population's health status. Item parameters for the 10 PF items are reported in Table C.1.

Figure C.1 Item Characteristic Curves for Physical Functioning Item 3d

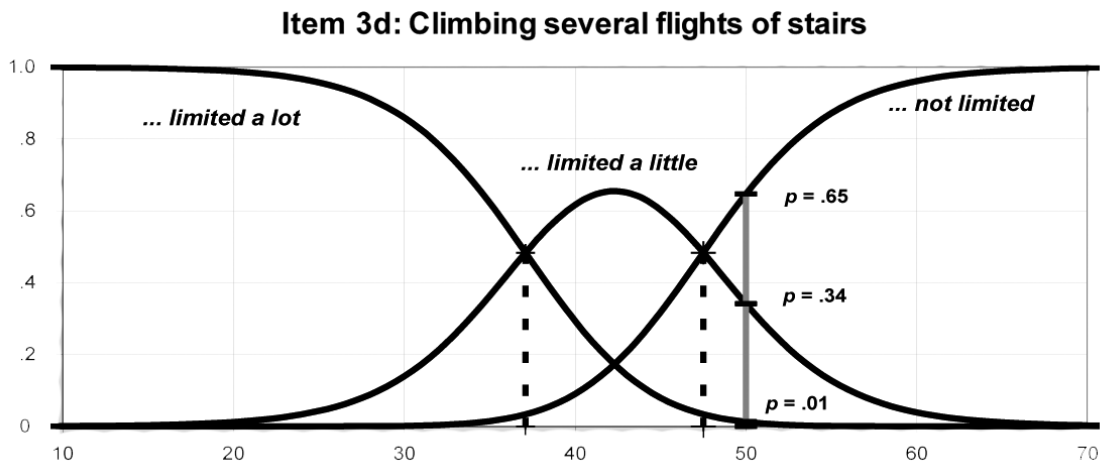


Table C.1*Partial Credit Model Item Response Theory Parameters for the SF-36v2 Physical Functioning Items*

Item	Content	T-Score Metric ^a			Standard Normal Metric ^b		
		Slope	Threshold		Slope	Threshold	
			1	2		1	2
3a	Vigorous activities	0.256	43.9	56.0	2.56	-0.61	0.60
3b	Moderate activities	0.256	33.7	44.0	2.56	-1.63	-0.60
3c	Lift/carry groceries	0.256	31.0	41.1	2.56	-1.90	-0.90
3d	Climb several flights of stairs	0.256	37.1	47.5	2.56	-1.29	-0.25
3e	Climb one flight of stairs	0.256	30.7	40.1	2.56	-1.93	-1.00
3f	Bend/kneel/stoop	0.256	34.0	46.9	2.56	-1.60	-0.31
3g	Walking more than a mile	0.256	33.0	46.2	2.56	-1.70	-0.38
3h	Walking several hundred yards	0.256	35.9	41.4	2.56	-1.41	-0.86
3i	Walking one hundred yards	0.256	30.2	36.7	2.56	-1.98	-1.33
3j	Bathing or dressing	0.256	25.4	31.4	2.56	-2.46	-1.87

^aMean = 50, *SD* = 10^bMean = 0, *SD* = 1

Furthermore, the probabilities from these item characteristic curves can be combined to estimate the probability of any pattern of item responses, for any given level of health (see Thissen & Orlando, 2001). Figure C.2 illustrates how this principle can be used to estimate the physical functioning of respondents with different combinations of answers. In this figure, the top three graphs show the curves for three different PF items. (Note that the layout and interpretation of these graphs are the same as for Figure C.1.) As shown in the figure, Item 3d (climbing stairs) is the most difficult item because its thresholds are higher than (i.e., to the right of) those for the other two items. In comparison, Item 3i (walking one block) is shown to be the easiest item. To illustrate, respondents with a physical functioning *T* score of 30 or lower are the most likely to select *limited a lot* in response to Items 3d and 3f and *limited a little* in response to Item 3i. Also, respondents are less likely to choose *limited a little* in response to Item 3i, relative to the other items. In general, it has been observed that middle response categories have narrower ranges on easy items as compared to more difficult items.

In practice, the latent physical functioning of any given respondent is unknowable; however, the physical functioning level underlying each pattern of responses can be estimated. For example, a score can be estimated for a respondent who chose *limited a little* in response to Items 3d and 3f and *not limited* in response Item 3i. The probability of this answer combination can be calculated at each level of latent physical functioning by multiplying the values derived from Figure C.2's three trace lines (i.e., the solid black lines in the upper three graphs). The result

of this process is the black line depicted in Figure C.2's final (bottom) graph. In other words, if a respondent's physical functioning is viewed as a parameter one wants to estimate, this line represents the *likelihood function* for the latent physical functioning of this observed answer combination. As shown in the graph, it is clear that a respondent giving these three answers would most likely have a latent physical functioning score of approximately 43. In principle, any subset of items that fits the model can be used to get an unbiased estimate (i.e., an estimate without systematic error) of latent physical functioning. Note that the QualityMetric Health Outcomes™ Scoring Software 5.0 (Saris-Baglama et al., 2011) uses a modified version of this approach (weighted maximum likelihood [WML] estimation; Warm, 1989) to estimate the scores for respondents with missing responses to PF items (when full MSE is specified).

Although both the IRT scores and the traditional (sum) scores are norm-based, with a mean of 50 and a standard deviation of 10, these scores are not directly comparable. Thus, IRT-based scores must be calibrated to the sum score metric. With the partial credit model, a one-to-one relationship exists between the sum score and the IRT score estimate (Andersen, 1977); thus, the transformation can be established using the IRT model. The relation between the two scoring metrics is illustrated in Figure C.3. In case of missing data, score estimation proceeds in the following manner: (a) estimate an IRT score based on the items answered, the item parameters found in Table C.1, and the WML estimation method; and (b) calibrate the IRT score to the sum score metric using the graph shown in Figure C.3.

Figure C.2 Item Characteristic Curves for Physical Functioning Items 3d, 3f, and 3i

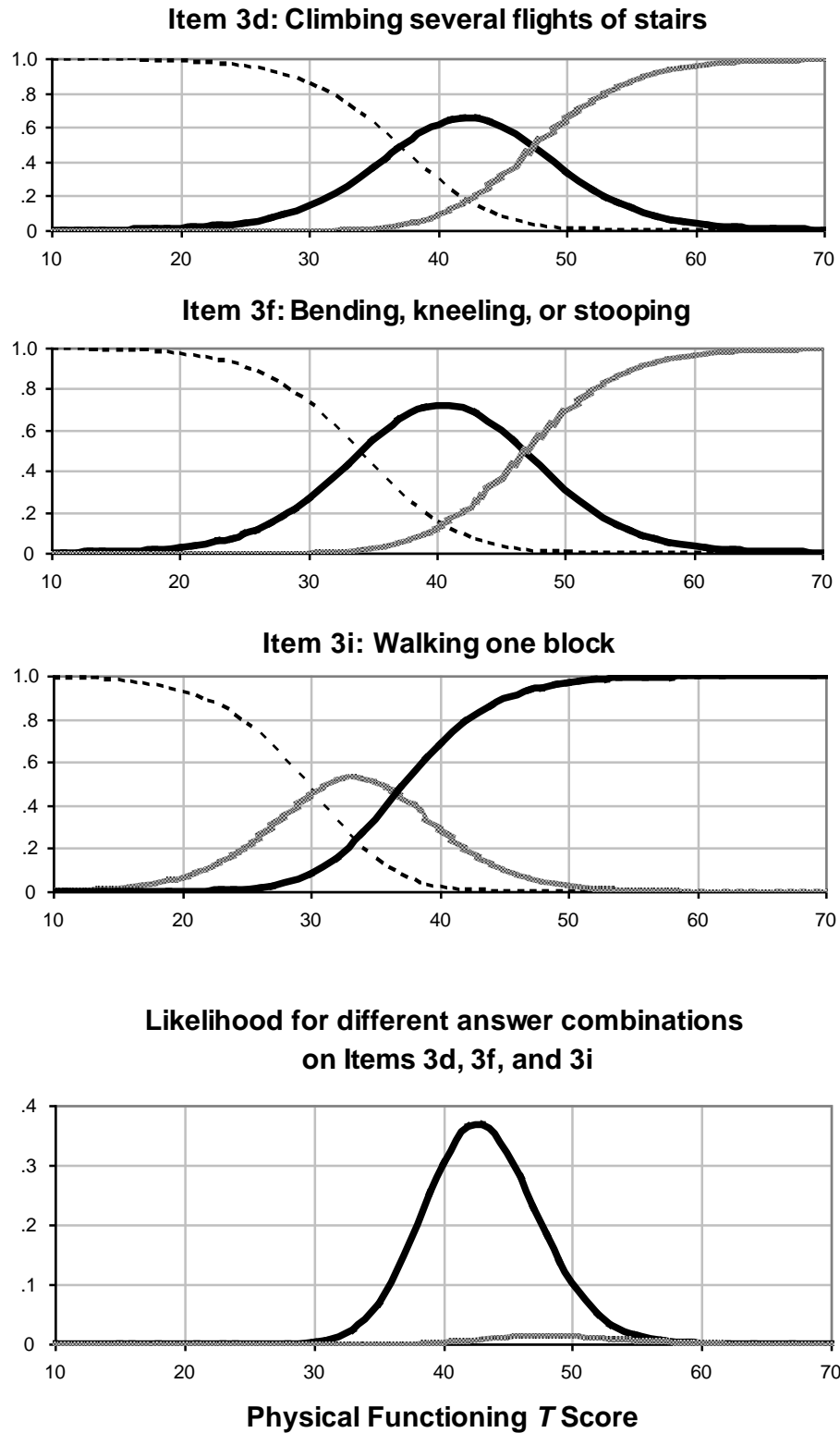
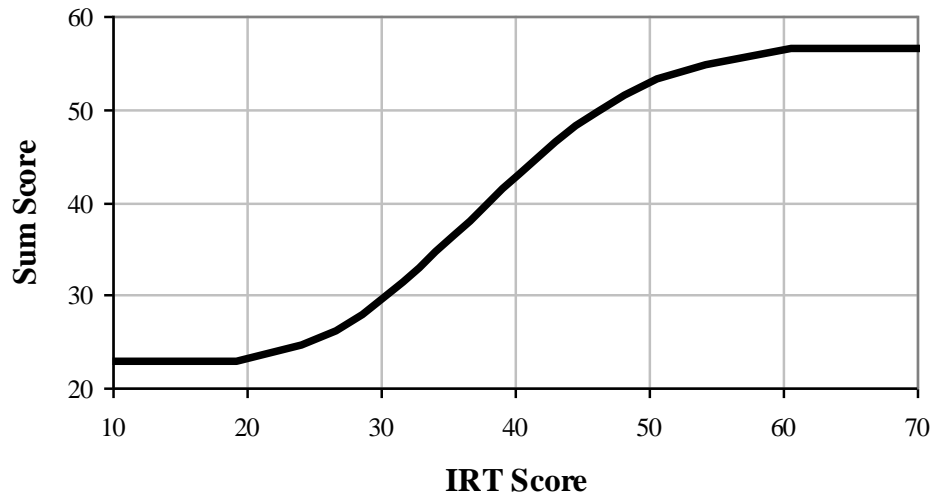


Figure C.3 Relation Between Physical Functioning IRT and Sum Scores



References

- Aaronson, N. K., Acquadro, C., Alonso, J., Apolone, G., Bucquet, D., Bullinger, M., ... Ware, J. E., Jr. (1992). International Quality of Life Assessment (IQOLA) Project. *Quality of Life Research, 1*, 349–351.
- Abramson, M. J., Schattner, R. L., Sulaiman, N. D., Birch, K. E., Simpson, P. P., Del Colle, E. A., ... Thien, F. C. (2010). Do spirometry and regular follow-up improve health outcomes in general practice patients with asthma or COPD? A cluster randomised controlled trial. *Medical Journal of Australia, 193*, 104–109.
- Acquadro, C., Berzon, R., Dubois, D., Leidy, N. K., Marquis, P., Revicki, D., & Rothman, M. (2003). Incorporating the patient's perspective into drug development and communication: An ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. *Value in Health, 6*, 522–531.
- Afdhal, N. H., Dieterich, D. T., Pockros, P. J., Schiff, E. R., Shiffman, M. L., Sulkowski, M. S., ... Bowers, P. J. (2004). Epoetin alfa maintains Ribavirin dose in HCV-infected patients: A prospective, double-blind, randomized controlled study. *Gastroenterology, 126*, 1302–1311.
- American College of Physicians, Health and Public Policy Committee. (1988). Comprehensive functional assessment for elderly patients. *Annals of Internal Medicine, 109*, 70–72.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.
- Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan Publishing Company.
- Andersen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42*, 69–81.
- Angst, F., Aeschlimann, A., & Stucki, G. (2001). Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. *Arthritis and Rheumatism, 45*, 384–391.
- Aoki, F. Y., Fleming, D. M., Griffin, A. D., Lacey, L. A., & Edmundson, S. (2000). Impact of zanamivir treatment on productivity, health status, and healthcare resource use in patients with influenza. *Pharmacoeconomics, 2*, 187–195.
- Apolone, G., De Carli, G., Brunetti, M., & Garattini, S. (2001). An evaluation of the EMEA recommendations on the use of quality of life measures in drug approval. *Pharmacoeconomics, 19*, 187–195.
- Apolone, G., Mosconi, P., & Ware, J. E., Jr. (1997). *Questionario sullo stato di salute SF-36. Manuale d'uso e guida all'interpretazione dei risultati* [SF-36 health status questionnaire. User's manual and guide to the interpretation of results]. Milano: Guerini & Associati Editore.
- Avlund, K., Kreiner, S., & Schultz-Larsen, K. (1993). Construct validation and the Rasch model: Functional capacity of healthy elderly people. *Scandinavian Journal of Social Medicine, 21*, 233–246.
- Baldwin, C. M., Grant, M., Wendel, C., Hornbrook, M. C., Herrinton, L. J., McMullen, C., & Krouse, R. S. (2009). Gender differences in sleep disruption and fatigue on quality of life among persons with ostomies. *Journal of Clinical Sleep Medicine, 5*, 335–343.
- Becker, J., Saris-Baglama, R. N., Kosinski, M., Williams, B., & Bjorner, J. B. (2005). *The Pain Impact Questionnaire (PIQ-6): A user's guide*. Lincoln, RI: QualityMetric Incorporated.

- Bennett, R. M., Schein, J., Kosinski, M. R., Hewitt, D. J., Jordan, M., & Rosenthal, N. R. (2005). Impact of fibromyalgia pain on health-related quality of life before and after treatment with tramadol/acetaminophen. *Arthritis and Rheumatism*, *53*, 519–527.
- Berdit, M., & Williamson, J. E. (1973). *Function limitation scale for measuring health outcome in health status indexes*. Chicago: Hospital Research and Educational Trust.
- Bergner, M., Bobbitt, R. A., Carter, W. B., & Gilson, B. S. (1981). The Sickness Impact Profile: Development and final revision of a health status measure. *Medical Care*, *19*, 787–805.
- Bergner, M., Bobbitt, R. A., Kressel, S., Pollard, W. E., Gilson, B. S., & Morris, J. R. (1976). The Sickness Impact Profile: Conceptual formulation and methodology for the development of a health status measure. *International Journal of Health Services*, *6*, 393–415.
- Berki, S. E., & Ashcraft, M. L. (1979). On the analysis of ambulatory utilization: An investigation of the roles of need, access, and price as predictors of illness and preventive visits. *Medical Care*, *17*, 1163–1181.
- Berwick, D. M., Murphy, J. M., Goldman, P. A., Ware, J. E., Jr., Barsky, A. J., & Weinstein, M. C. (1991). Performance of a five-item mental health screening test. *Medical Care*, *29*, 169–176.
- Bird, D., Oldenburg, B., Cassimatis, M., Russell, A., Ash, S., Courtney, M. D., ... Friedman, R. H. (2010). Randomised controlled trial of an automated, interactive telephone intervention to improve type 2 diabetes self-management (Telephone-Linked Care Diabetes Project): Study protocol. *BMC Public Health*, *10*, 599.
- Bjorner, J. B., Damsgaard, M. T., Watt, T., Bech, P., Rasmussen, N. K., Kristensen, T. S., ... Thunedborg, K. (1997). *Dansk manual til SF-36. Et spørgeskema om helbredsstatus* [Danish manual for the SF-36. A questionnaire about health]. Copenhagen: Lif.
- Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003a). Calibration of an item pool for assessing the burden of headaches: An application of item response theory to the Headache Impact Test (HIT). *Quality of Life Research*, *12*, 913–933.
- Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003b). The feasibility of applying item response theory to measures of migraine impact: A re-analysis of three clinical studies. *Quality of Life Research*, *12*, 887–902.
- Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2003c). Using item response theory to calibrate the Headache Impact Test (HIT) to the metric of traditional headache scales. *Quality of Life Research*, *12*, 981–1002.
- Bjorner, J. B., Kosinski, M., & Ware, J. E., Jr. (2005). Computerized adaptive testing and item banking. In P. M. Fayers & R. D. Hays (Eds.), *Assessing quality of life* (2nd ed., pp. 95–112). Oxford, England: Oxford University Press.
- Bjorner, J. B., Kreiner, S., Ware, J. E., Jr., Damsgaard, M. T., & Bech, P. (1998). Differential item functioning in the Danish translation of the SF-36. *Journal of Clinical Epidemiology*, *51*, 1189–1202.
- Bjorner, J. B., Wallenstein, G. V., Martin, M. C., Lin, P., Blaisdell-Gross, B., Piech, C. T., & Mody, S. H. (2007). Interpreting score differences in the SF-36 Vitality scale: Using clinical conditions and functional outcomes to define the minimally important difference. *Current Medical Research and Opinion*, *23*, 731–739.
- Bjorner, J. B., & Ware, J. E., Jr. (1998). Using modern psychometric methods to measure health outcomes. *Medical Outcomes Trust Monitor*, *3*, 11–16.
- Bliven, B. D., Kaufman, S. E., & Spertus, J. A. (2001). Electronic collection of health-related quality of life data: Validity, time benefits, and patient preference. *Quality of Life Research*, *10*, 15–22.
- Bombardier, C., Melfi, C. A., Paul, J., Green, R., Hawker, G., Wright, J., & Coyte, P. (1995). Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. *Medical Care*, *33*(Suppl. 4), AS131–AS144.
- Bombardier, C., Ware, J. E., Jr., Russell, I. J., Larson, M., Chalmers, A., & Read, J. L. (1986). Auranofin therapy and quality of life in patients with rheumatoid arthritis: Results of a multicenter trial. *American Journal of Medicine*, *81*, 565–578.
- Brazier, J., Roberts, J., & Deverill, M. (2002). The estimation of a preference-based measure of health from the SF-36. *Journal of Health Economics*, *21*, 271–292.
- Brazier, J., Usherwood, T., Harper, R., & Thomas, K. (1998). Deriving a preference-based single index from the UK SF-36 Health Survey. *Journal of Clinical Epidemiology*, *51*, 1115–1128.
- Brazier, J. E., Harper, R., Jones, N. M. B., O’Cathain, A., Thomas, K. J., Usherwood, T., & Westlake, L. (1992). Validating the SF-36 health survey questionnaire: New outcome measure for primary care. *British Medical Journal*, *305*, 160–164.
- Brazier, J. E., & Roberts, J. (2004). The estimation of a preference-based measure of health from the SF-12. *Medical Care*, *42*, 851–859.

- Brook, R. H., Fink, A., Koscoff, J., Linn, L. S., Watson, W. E., Davies, A. R., ... Delbanco, T. L. (1987). Educating physicians and treating patients in the ambulatory setting: Where are we going and how will we know when we arrive? *Annals of Internal Medicine*, *107*, 392–398.
- Brook, R. H., Ware, J. E., Jr., Davies-Avery, A., Stewart, A. L., Donald, C. A., Rogers, W. H., ... Johnston, S. A. (1979). Overview of adult health measures fielded in Rand's health insurance study. *Medical Care*, *17*(Suppl. 7), iii–x, 1–131.
- Brook, R. H., Ware, J. E., Jr., Rogers, W. H., Keeler, E. B., Davies, A. R., Donald, C. A., ... Newhouse, J. P. (1983). Does free care improve adults' health? Results from a randomized controlled trial. *New England Journal of Medicine*, *309*, 1426–1434.
- Bullinger, M., Alonso, J., Apolone, G., Lepège, A., Sullivan, M., Wood-Dauphinee, S., ... Ware, J. E., Jr. (1998). Translating health status questionnaires and evaluating their quality: The International Quality of Life Assessment Project approach. *Journal of Clinical Epidemiology*, *51*, 913–923.
- Buskirk, T. D., & Stein, K. D. (2008). Telephone vs. mail survey gives different SF-36 quality-of-life scores among cancer survivors. *Journal of Clinical Epidemiology*, *61*, 1049–1055.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A. M., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.
- Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2: Manual for administration and scoring* (2nd ed.). Minneapolis, MN: University of Minnesota Press.
- Cameron, G. D. (1954). The Canadian Sickness Survey 1950–1951: Its implications for the practising physician. *Canadian Medical Association Journal*, *71*, 613–615.
- Camilleri-Brennan, J., & Steele, R. J. (2002). Objective assessment of morbidity and quality of life after surgery for low rectal cancer. *Colorectal Disorders*, *4*, 61–66.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.
- Caro, J. J., Sr., Caro, I., Caro, J., Wouters, F., & Juniper, E. F. (2001). Does electronic implementation of questionnaires used in asthma alter responses compared to paper implementation? *Quality of Life Research*, *10*, 683–691.
- Carreon, L. Y., Glassman, S. D., Campbell, M. J., & Anderson, P. A. (2010). Neck Disability Index, Short Form-36 Physical Component Summary, and pain scales for neck and arm pain: The minimum clinically important difference and substantial clinical benefit after cervical spine fusion. *Spine Journal*, *10*, 469–474.
- Chambers, L. W. (1988). The McMaster Health Index Questionnaire: An update. In S. R. Walker & R. M. Rosser (Eds.), *Quality of life: Assessment and application* (pp. 113–131). Lancaster, England: MTP Press Limited.
- Chassany, O., Sagnier, P., Marquis, P., Fullerton, S., & Aaronson, N. (2002). Patient-reported outcomes: The example of health-related quality of life—A European guidance document for the improved integration of health-related quality of life assessment in the drug regulatory process. *Drug Information Journal*, *36*, 209–238.
- Cheng, M. B., & Ferrante, F. M. (2006). Health-related quality of life in sacroiliac syndrome: A comparison to lumbosacral radiculopathy. *Regional Anesthesia and Pain Medicine*, *31*, 422–427.
- Cluff, L. E. (1981). Chronic disease, function, and the quality of care. *Journal of Chronic Diseases*, *34*, 299–304.
- Coates, A., Gebiski, V., Bishop, J. F., Jeal, P. N., Woods, R. L., Snyder, R., ... Forbes, J. F. (1987). Improving the quality of life during chemotherapy for advanced breast cancer: A comparison of intermittent and continuous treatment strategies. *New England Journal of Medicine*, *317*, 1490–1495.
- Codman, E. A. (1990). The product of a hospital. 1914. *Archives of Pathology and Laboratory Medicine*, *114*, 1106–1111.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Colangelo, K. J., Pope, J. E., & Peschken, C. (2009). The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36. *Journal of Rheumatology*, *36*, 2231–2237.
- Comrey, A. L. (1973). *A first course in factor analysis*. New York: Academic Press.
- Crespi, C. M., Smith, S. K., Petersen, L., Zimmerman, S., & Ganz, P. A. (2010). Measuring the impact of cancer: A comparison of non-Hodgkin lymphoma and breast cancer survivors. *Journal of Cancer Survivorship: Research and Practice*, *4*, 45–58.

- Croog, S. H., Levine, S., Testa, M. A., Brown, B., Bulpitt, C. J., Jenkins, C. D., ... Williams, G. H. (1986). The effects of antihypertensive therapy on the quality of life. *New England Journal of Medicine*, *314*, 1657–1664.
- Daut, R. L., Cleeland, C. S., & Flannery, R. C. (1983). Development of the Wisconsin Brief Pain Questionnaire to assess pain in cancer and other diseases. *Pain*, *17*, 197–210.
- Davies, A. R. (2000). Using health measures: Clinical practice [Video study guide]. *Understanding health outcomes: An educational series on CD-ROM, Series one—Health status: Concepts, measures, and applications*. Woodbridge, NJ: HealthStat Productions, Inc.
- Davies, A. R., & Kram, B. (Eds.) (2002). Monitoring outcomes: Patients with sleep disorders [Video study guide]. *Understanding health outcomes: An educational series on CD-ROM, Series two—Measuring specific conditions*. Woodbridge, NJ: HealthStat Productions, Inc.
- Davies, A. R., Sherbourne, C. D., Peterson, J. R., & Ware, J. E., Jr. (1988). *Scoring manual: Adult health status and patient satisfaction measures used in RAND's Health Insurance Experiment* (Publication No. N-2190-HHS). Santa Monica, CA: The RAND Corporation.
- Davies, A. R., & Ware, J. E., Jr. (1981). *Measuring health perceptions in the Health Insurance Experiment* (Publication No. R-2711-HHS). Santa Monica, CA: The RAND Corporation.
- Deniston, O. L., & Jette, A. A. (1980). A functional status assessment instrument: Validation in an elderly population. *Health Services Research*, *15*, 21–34.
- Derogatis, L. R. (1986). The Psychosocial Adjustment to Illness Scale (PAIS). *Journal of Psychosomatic Research*, *30*, 77–91.
- Detmar, S. B., Muller, M. J., Schornagel, J. H., Wever, L. D. V., & Aaronson, N. K. (2002). Health-related quality-of-life assessments and patient-physician communication: A randomized controlled trial. *Journal of the American Medical Association*, *288*, 3027–3034.
- Deyo, R. A., & Patrick, D. L. (1989). Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Medical Care*, *27*, S254–S268.
- DiCocco, L., & Apple, D. (1958). Health needs and opinions of older adults. *Public Health Reports*, *3*, 479–487.
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- Donald, C. A., & Ware, J. E., Jr. (1984). The measurement of social support. In J. R. Greenley (Ed.), *Research in community and mental health* (pp. 325–370). Greenwich, CT: JAI Press.
- Donald, C. A., Ware, J. E., Jr., Brook, R. H., & Davies-Avery, A. (1978). *Conceptualization and measurement of health for adults in the Health Insurance Study. Vol. IV: Social health* (Report No. R-1987/4-HEW). Santa Monica, CA: The RAND Corporation.
- Dupuy, H. J. (1973). *Developmental rationale, substantive, derivative, and conceptual relevance of general well-being*. Draft working paper. Washington, DC: National Center for Health Statistics.
- Dupuy, H. J. (1984). The Psychological General Well-Being (PGWB) Index. In N. K. Wenger, M. E. Mattson, C. D. Furberg, & J. Elinson (Eds.), *Assessment of quality of life in clinical trials of cardiovascular therapies* (pp. 170–183). New York: Le Jacq Publishing Company.
- Eisen, M., Donald, C. A., Ware, J. E., Jr., & Brook, R. H. (1980). *Conceptualization and measurement of health for children in the Health Insurance Study* (Report No. R-2313-HEW). Santa Monica, CA: The RAND Corporation.
- Ellwood, P. M. (1988). Shattuck lecture—Outcomes management: A technology of patient experience. *New England Journal of Medicine*, *318*, 1549–1556.
- Elston, J., Honan, W., Powell, R., Gormley, J., & Stein, K. (2010). Do metronomes improve the quality of life in people with Parkinson's disease? A pragmatic, single-blind, randomized cross-over trial. *Clinical Rehabilitation*, *24*, 523–532.
- Fayers, P., Aaronson, N. K., Bjordal, K., Curran, D., & Groenvold, M. (1999). *EORTC QLQ-C30 scoring manual* (2nd ed.). Brussels, Belgium: European Organization for Research on the Treatment of Cancer.
- Fernandez-Fairen, M., Sala, P., Ramirez, H., & Gil, J. (2007). A prospective randomized study of unilateral versus bilateral instrumented posterolateral lumbar fusion in degenerative spondylolisthesis. *Spine*, *32*, 395–401.
- Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. Berlin, Germany: Springer-Verlag.

- Fitzgibbons, R. J., Jr., Giobbie-Hurder, A., Gibbs, J. O., Dunlop, D. D., Reda, D. J., McCarthy, M., Jr., ... Jonasson, O. (2006). Watchful waiting vs. repair of inguinal hernia in minimally symptomatic men: A randomized clinical trial. *Journal of the American Medical Association*, *295*, 285–292.
- Fleischman, J. A., Cohen, J. W., Manning, W. G., & Kosinski, M. (2006). Using SF-12 health status measures to improve predictions of medical expenditures. *Medical Care*, *44*(Suppl. 5), I54–I63.
- Forbes, A., While, A., Mathes, L., & Griffiths, P. (2006). Health problems and health-related quality of life in people with multiple sclerosis. *Clinical Rehabilitation*, *20*, 67–78.
- Fowler, F. J., Jr. (1984). *Survey research methods*. Beverly Hills, CA: Sage Publications.
- Fowler, F. J., Jr., Wennberg, J. E., Timothy, R. P., Barry, M. J., Mulley, A. G., Jr., & Hanley, D. (1988). Symptom status and quality of life following prostatectomy. *Journal of the American Medical Association* *259*, 3018–3022.
- Fukuhara, S., Suzukamo, Y., Bito, S., & Kurokawa, K. (2001). *Manual of SF-36 Japanese version 1.2*. Tokyo: Public Health Research Foundation.
- Gandek, B., Sinclair, S. J., Kosinski, M., & Ware, J. E., Jr. (2004). Psychometric evaluation of the SF-36 Health Survey in Medicare managed care. *Health Care Financing Review*, *25*(4), 5–25.
- Gandek, B., & Ware, J. E., Jr. (1998a). Methods for validating and norming translations of health status questionnaires: The IQOLA Project approach. *Journal of Clinical Epidemiology*, *51*, 953–959.
- Gandek, B., & Ware, J. E., Jr. (Eds.) (1998b). Translating functional health and well-being: International Quality of Life Assessment (IQOLA) Project studies of the SF-36 Health Survey [Special issue]. *Journal of Clinical Epidemiology*, *51*(11).
- Garratt, A., Schmidt, L., Mackintosh, A., & Fitzpatrick, R. (2002). Quality of life measurement: Bibliographic study of patient assessed health outcome measures. *British Medical Journal*, *324*, 1417–1419.
- Geigle, R., & Jones, S. B. (1990). Outcomes measurement: A report from the front. *Inquiry*, *27*, 7–13.
- Gersh, E., Arnold, C., & Gibson, S. J. (2011). The relationship between the readiness for change and clinical outcomes in response to multidisciplinary pain management. *Pain Medicine*, *12*, 165–172.
- Girard, F., Chouinard, P., Boudreault, D., Poirier, C., Richard, C., Ruel, M., & Ferraro, P. (2006). Prevalence and impact of pain on the quality of life of lung transplant recipients: A prospective observational study. *Chest*, *130*, 1535–1540.
- Goldberg, D. P., & Hillier, V. (1979). A scaled version of the General Health Questionnaire. *Psychological Medicine*, *9*, 139–145.
- Gottschalk, A., & Flocke, S. A. (2005). Time spent in face-to-face patient care and work outside the examination room. *Annals of Family Medicine*, *3*, 488–493.
- Granger, C. V., Hamilton, B. B., Linacre, J. M., Heine-mann, A. W., & Wright, B. D. (1993). Performance profiles of the functional independence measure. *American Journal of Physical Medicine and Rehabilitation*, *72*, 84–89.
- Greenfield, D. M., Walters, S. J., Coleman, R. E., Hancock, B. W., Snowden, J. A., Shalet, S. M., ... Ross, R. J. M. (2010). Quality of life, self-esteem, fatigue, and sexual function in young men after cancer: A controlled cross-sectional study. *Cancer*, *116*, 1592–1601.
- Groenvold, M., Bjorner, J. B., Klee, M. C., & Kreiner, S. (1995). Test for item-bias in a quality of life measure. *Journal of Clinical Epidemiology*, *48*, 805–816.
- Guttman, L. A. (1944). A basis for scaling qualitative data. *American Sociological Review*, *9*, 139–150.
- Guyatt, G. H., Berman, L. B., Townsend, M., Pugsley, S. O., & Chambers, L. W. (1987). A measure of quality of life in chronic lung disease. *Thorax*, *42*, 773–778.
- Guyatt, G. H., Osoba, D., Wu, A. W., Wyrwich, K. W., & Norman, G. R. (2002). Methods to explain the clinical significance of health status measures. *Mayo Clinic Proceedings*, *77*, 371–383.
- Gwaltney, C. J., Shields, A. L., & Shiffman, S. (2008). Equivalence of electronic and paper-and-pencil administration of patient-reported outcome measures: A meta-analytic review. *Value in Health*, *11*, 322–333.
- Haffer, S. C., Bowen, S. E., Shannon, E. D., & Fowler, B. M. (2003). Assessing beneficiary health outcomes and disease management initiatives in Medicare. *Disease Management and Health Outcomes*, *11*, 111–124.
- Haley, S. M., McHorney, C. A., & Ware, J. E., Jr. (1994). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): I. Unidimensionality and reproducibility of the Rasch item scale. *Journal of Clinical Epidemiology*, *47*, 671–684.
- Hall, J. A., Epstein, A. M., & McNeil, B. J. (1989). Multidimensionality of health status in an elderly population: Construct validity of a measurement battery. *Medical Care*, *27*(3, Suppl.), S168–S177.
- Hanmer, J., Lawrence, W. F., Anderson, J. P., Kaplan, R. M., & Fryback, D. G. (2006). Report of nationally representative values for the noninstitutionalized U.S. adult population for seven health-related quality-of-life scores. *Medical Decision Making*, *26*, 391–400.

- Hanscom, B., Lurie, J. D., Homa, K., & Weinstein, J. N. (2002). Computerized questionnaires and the quality of survey data. *Spine*, *27*, 1797–1801.
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed., rev.). Chicago: University of Chicago Press.
- Harris, M. L., & Harris, C. W. (1971). A factor analytic interpretation strategy. *Educational and Psychological Measurement*, *31*, 589–606.
- Hatcher, G. (1956). Symposium on the Canadian Sickness Survey: Summary and implications. *Canadian Journal of Public Health*, *47*, 378–382.
- Hawn, M. T., Itani, K. M., Giobbie-Hurder, A., McCarthy, M., Jr., Jonasson, O., & Neumayer, L. A. (2006). Patient-reported outcomes after inguinal herniorrhaphy. *Surgery*, *140*, 198–205.
- Hawthorne, G., Kaye, A. H., Gruen, R., Houseman, D., & Bauer, I. (2011). Traumatic brain injury and quality of life: Initial Australian validation of the QOLIBRI. *Journal of Clinical Neuroscience*, *18*, 197–202.
- Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The RAND 36-Item Health Survey 1.0. *Health Economics*, *2*, 217–227.
- Hays, R. D., & Stewart, A. L. (1990). The structure of self-reported health in chronic disease patients. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*, 22–30.
- Helmstadter, G. C. (1964). *Principles of psychological measurement*. New York: Appleton-Century-Crofts.
- Hing, E., Cherry, D. K., & Woodwell, D. A. (2005). National Ambulatory Medical Care Survey: 2003 summary. *Advanced Data from Vital and Health Statistics*, No. 365. Hyattsville, MD: National Center for Health Statistics. Retrieved from <http://www.cdc.gov/nchs/data/ad/ad365.pdf>
- Hirsch, J. D., Lee, S. J., Terkeltaub, R., Khanna, D., Singh, J., Sarkin, A. . . Kavanaugh, A. (2008). Evaluation of an instrument assessing influence of gout on health-related quality of life. *Journal of Rheumatology*, *35*, 2406–2414.
- Hornbrook, M. C., & Goodman, M. J. (1995). Assessing relative health plan risk with the RAND-36 Health Survey. *Inquiry*, *32*, 56–74.
- Hudson, M., Thombs, B. D., Steele, R., Panopalis, P., Newton, E., & Baron, M. (2009). Quality of life in patients with systemic sclerosis compared to the general population and patients with other chronic conditions. *Journal of Rheumatology*, *36*, 768–772.
- Hunt, S. M., McKenna, S. P., McEwen, J., Williams, J., & Papp, E. (1981). The Nottingham Health Profile: Subjective health status and medical consultations. *Social Sciences in Medicine*, *15*(3, Pt.1), 221–229.
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.
- Jenkinson, C., Layte, R., Wright, L., & Coulter, A. (1996). *The U.K. SF-36: An analysis and interpretation manual*. Oxford, England: University of Oxford.
- Jenkinson, C., & Stewart-Brown, S. (1999). Reply [Letter to the editor]. *Journal of Epidemiology and Community Health*, *53*, 651–652.
- Jenkinson, C., Stewart-Brown, S., Petersen, S., & Paice, C. (1999). Assessment of the SF-36 Version 2 in the United Kingdom. *Journal of Epidemiology and Community Health*, *53*, 45–50.
- Jette, A. M. (1980). The Functional Status Index: Reliability of a chronic disease evaluation instrument. *Archives of Physical Medicine and Rehabilitation*, *61*, 395–401.
- Jette, A. M. (1987). The Functional Status Index: Reliability and validity of a self-report functional disability measure. *Journal of Rheumatology*, *14*(Suppl. 15), 15–19.
- Jette, A. M., Davies, A. R., Cleary, P. D., Calkins, D. R., Rubenstein, L. V., Fink, A., . . . Delbanco, T. L. (1986). The Functional Status Questionnaire: Reliability and validity when used in primary care. *Journal of General Internal Medicine*, *1*, 143–149.
- Jörngården, A., Wettergen, L., & von Essen, L. (2006). Measuring health-related quality of life in adolescents and young adults: Swedish normative data for the SF-36 and the HADS, and the influence of age, gender, and method of administration. *Health and Quality of Life Outcomes*, *4*, 91.
- Julious, S. A., George, S., & Campbell, M. J. (1995). Sample sizes for studies using the Short Form 36 (SF-36). *Journal of Epidemiology and Community Health*, *49*, 642–644.
- Kantz, M. E., Harris, W. J., Levitsky, K., Ware, J. E., Jr., & Davies, A. R. (1992). Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. *Medical Care*, *30*(Suppl. 5), MS240–MS252.
- Kaplan, R. M. (1989). Health outcome models for policy analysis. *Health Psychology*, *8*, 723–735.
- Kaplan, R. M., & Anderson, J. P. (1988). A general health policy model: Update and applications. *Health Services and Research*, *23*, 203–235.
- Katz, J. N., Larson, M. G., Phillips, C. B., Fossel, A. H., & Liang, M. H. (1992). Comparative measurement sensitivity of short and longer health status instruments. *Medical Care*, *30*, 917–925.

- Katz, S. (Ed.). (1987). The Portugal conference: Measuring quality of life and functional status in clinical and epidemiological research. *Journal of Chronic Diseases, 40*, 459–650.
- Katz, S., Downs, T. D., Cash, H. R., & Grotz, R. C. (1970). Progress in development of the index of ADL. *Gerontologist, 10*(1), 20–30.
- Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., & Jaffe, M. W. (1963). Studies of illness in the aged. The Index of ADL: A standardized measure of biological and psychosocial function. *Journal of the American Medical Association, 185*, 914–919.
- Keller, S. D., Bayliss, M. S., Ware, J. E., Jr., Hsu, M. A., Damiano, A. M., & Goss, T. F. (1997). Comparison of responses to SF-36 Health Survey questions with one-week and four-week recall periods. *Health Services Research, 32*, 367–384.
- Keller, S. D., & Ware, J. E., Jr. (1995). Interpretation strategies for health status scores. *Medical Outcomes Trust Bulletin, 3*(5), 2–3.
- Keller, S. D., Ware, J. E., Jr., Gandek, B., Aaronson, N. K., Alonso, J., Apolone, G., ... Wood-Dauphinee, S. (1998). Testing the equivalence of translations of widely used response choice labels: Results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology, 51*, 933–944.
- Kerlinger, F. N. (1973). *Foundations of behavioral research*. New York: Holt, Rinehart, and Winston.
- Kim, D. S., Sim, Y.-J., Jeong, H. J., & Kim, G. C. (2010). Effect of active resistive exercise on breast cancer-related lymphedema: A randomized controlled trial. *Archives of Physical Medicine and Rehabilitation, 91*, 1844–1848.
- Ko, C. Y., Rusin, L. C., Schoetz, D. J., Collier, J. A., Murray, J. J., Roberts, P. L., & Moreau, L. (2002). Using quality of life scores to help determine treatment: Is restoring bowel continuity better than an ostomy? *Colorectal Disease, 4*, 41–47.
- Kosinski, M., Bayliss, M., Bjorner, J. B., & Ware, J. E., Jr. (2000). Improving estimates of SF-36 Health Survey scores for respondents with missing data. *Monitor, Fall*, 8–10.
- Kosinski, M., Bayliss, M. S., Turner-Bowker, D. M., & Fortin, E. W. (2004). *Asthma Control Test: A user's guide*. Lincoln, RI: QualityMetric Incorporated.
- Kosinski, M., Bjorner, J. B., Ware, J. E., Jr., Sullivan, E., & Straus, W. L. (2006). An evaluation of a patient-reported outcomes found computerized adaptive testing was efficient in assessing osteoarthritis impact. *Journal of Clinical Epidemiology, 59*, 715–723.
- Kosinski, M., Schein, J. R., Vallow, S. M., Ascher, S., Harte, C., Shikar, R., ... Orsanger, G. (2005). An observational study of health-related quality of life and pain outcomes in chronic low back pain patients treated with fentanyl transdermal system. *Current Medical Research and Opinion, 21*, 849–862.
- Kosinski, M., Zhao, S. Z., Dedhiya, S., Osterhaus, J. T., & Ware, J. E., Jr. (2000). Determining minimally important changes in generic and disease-specific health-related quality of life questionnaires in clinical trials of rheumatoid arthritis. *Arthritis and Rheumatism, 43*, 1478–1487.
- Kurtin, P. S., Davies, A. R., Meyer, K. B., DeGiacomo, J. M., & Kantz, M. E. (1992). Patient-based health status measures in outpatient dialysis: Early experiences in developing an outcomes assessment program. *Medical Care, 30*(5, Suppl.), MS136–MS149.
- Landgraf, J. M., Abetz, L., & Ware, J. E., Jr. (1999). *The CHQ: A user's manual*. Boston: The Health Institute.
- Lanman, T. H., & Hopkins, T. J. (2004). Early findings in a pilot study of anterior cervical interbody fusion in which recombinant human bone morphogenetic protein-2 was used with poly(L-lactide-co-D,L-lactide) bioabsorbable implants. *Neurosurgical Focus, 16*(3), E6.
- Laslett, L. L., Burnet, S. P., Jones, J. A., Redmond, C. L., & McNeil, J. D. (2007). Musculoskeletal morbidity: The growing burden of shoulder pain and disability and poor quality of life in diabetic outpatients. *Clinical and Experimental Rheumatology, 25*, 422–429.
- Lauridsen, H. H., Hartvigsen, J., Manniche, C., Korsholm, L., & Grunnet-Nilsson, N. (2006). Responsiveness and minimal clinically important difference for pain and disability instruments in low back pain patients. *BMC Musculoskeletal Disorders, 7*, 82.
- Lembcke, P. A. (1952). Measuring the quality of medical care through vital statistics based on hospital service areas: 1. Comparative study of appendectomy rates. *American Journal of Public Health, 42*, 276–286.
- Lepège, A., Ecosse, E., Pouchot, J., Coste, J., & Perneger, T. (2001). *Le questionnaire MOS SF-36: Manuel de l'utilisateur et guide d'interprétation des scores* [The MOS SF-36 questionnaire: User's manual and guide for interpreting scores]. Paris: Estem.
- Liang, J. (1986). Self-reported physical health among aged adults. *Journal of Gerontology, 41*, 248–260.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 5–55.
- Linder, J. A., & Singer, D. E. (2003). Health-related quality of life of adults with upper respiratory tract infections. *Journal of General Internal Medicine, 18*, 802–807.

- Lohr, K. N. (1989). Advances in health status assessment: Overview of the conference. *Medical Care*, 27, 1–11.
- Lohr, K. N. (1992). Applications of health status assessment measures in clinical practice: Overview of the Third Conference on Advances in Health Status Assessment. *Medical Care*, 30, 1–14.
- Lohr, K. N., & Ware, J. E., Jr. (1987). Proceedings of the Advances in Health Assessment Conference. *Journal of Chronic Disease*, 40(Suppl. 1), 1S–193S.
- Lungenhausen, M., Lange, S., Maier, C., Schaub, C., Trampisch, H. J., & Endres, H. G. (2007). Randomised controlled comparison of the Health Survey Short Form (SF-12) and the Graded Chronic Pain Scale (GCPS) in telephone interviews versus self-administered questionnaires. Are the results equivalent? *BMC Medical Research Methodology*, 7, 50.
- Lurie, N., Ward, N. B., Shapiro, M. F., & Brook, R. H. (1984). Termination from MediCal—Does it affect health? *New England Journal of Medicine*, 311, 480–484.
- Lyons, R. A., Wareham, K., Lucas, M., Price, D., Williams, J., & Hutchings, H. A. (1999). SF-36 scores vary by method of administration: Implications for study design. *Journal of Public Health Medicine*, 21, 41–45.
- Machin, D., & Fayers, P. (1998). Sample sizes for randomized trials measuring quality of life. In M. J. Staques, R. D. Hays, & P. Fayers (Eds.), *Quality of life assessment in clinical trials: Methods and practice* (pp. 37–50). Oxford, England: Oxford University Press.
- Macmillan, A. M. (1957). The health opinion survey: Techniques for estimating prevalence of psychoneurotic and related types of disorder in communities. *Psychological Reports*, 3(Monograph suppl. 7), 377–387.
- Manning, W. G., Newhouse, J. P., & Ware, J. E., Jr. (1982). The status of health in demand estimation: Beyond excellent, good, fair, and poor. In V. R. Fuchs (Ed.), *Economic aspects of health* (pp. 143–184) (Publication No. R-2696-HHS). Chicago: University of Chicago Press & The RAND Corporation.
- Marsh, J. L., McKinley, T., Dirschl, D., Pick, A., Haft, G., Anderson, D. D., & Brown T. (2010). The sequential recovery of health status after tibial plafond fractures. *Journal of Orthopaedic Trauma*, 24, 499–504.
- Martin, B. I., Levenson, L. M., Hollingworth, W., Kliot, M., Heagerty, P. J., Turner, J. A., & Jarvik, J. G. (2005). Randomized clinical trial of surgery versus conservative therapy for carpal tunnel syndrome. *BMC Musculoskeletal Disorders*, 6, 2.
- Maruish, M. E. (2002). *Psychological testing in the age of managed behavioral health care*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Maruish, M. E. (Ed.). (2004a). *The use of psychological testing for treatment planning and outcomes assessment. Volume 1: General considerations* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Maruish, M. E. (Ed.). (2004b). *The use of psychological testing for treatment planning and outcomes assessment. Volume 2: Instruments for children and adolescent* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Maruish, M. E. (Ed.). (2004c). *The use of psychological testing for treatment planning and outcomes assessment. Volume 3: Instruments for adults* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Mason, J. H., Anderson, J. J., & Meenan, R. F. (1988). A model of health status for rheumatoid arthritis: A factor analysis of the Arthritis Impact Measurement Scales. *Arthritis and Rheumatism*, 31, 714–720.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–173.
- Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–122). Berlin, Germany: Springer.
- McCune, C. A., Ravine D., Carter, K., Jackson, H. A., Hutton, D., Hedderich, J., ... Worwood, M. (2006). Iron loading and morbidity among relatives of HFE C282Y homozygotes identified either by population genetic testing or presenting as patients. *Gut*, 55, 554–562.
- McDermott, W. (1981). Absence of indicators of the influence of its physicians on a society's health: Impact of physician care on society. *American Journal of Medicine*, 70, 833–843.
- McDowell, I., & Newell, C. (1987). *General health measures. Measuring health: A guide to rating scales and questionnaires*. New York: Oxford University Press.
- McDowell, I., & Newell, C. (1996). *Measuring health: A guide to rating scales and questionnaires* (2nd ed.). New York: Oxford University Press.
- McHorney, C. A. (1997). Generic health measurement: Past accomplishments and a measurement paradigm for the 21st century. *Annals of Internal Medicine*, 127(8, Pt. 2), 743–750.
- McHorney, C. A., & Cohen, A. S. (2000). Equating health status measures with item response theory: Illustrations with functional status items. *Medical Care*, 38, II43–II59.

- McHorney, C. A., Haley, S. M., & Ware, J. E., Jr. (1997). Evaluation of the MOS SF-36 Physical Function Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology*, *50*, 451–461.
- McHorney, C. A., Kosinski, M., & Ware, J. E., Jr. (1994). 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*, *32*, 551–567.
- McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: Are available health status surveys adequate? *Quality of Life Research*, *4*, 293–307.
- McHorney, C. A., Ware, J. E., Jr., Lu, J. F., & Sherbourne, C. D. (1994). The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care*, *32*, 40–66.
- McHorney, C. A., Ware, J. E., Jr., & Raczek, A. E. (1993). The MOS 36-item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, *31*, 247–263.
- McHorney, C. A., Ware, J. E., Jr., Rogers, W., Raczek, A. E., & Lu, J. F. (1992). 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*, *30*(Suppl. 5), MS253–MS265.
- Medical Outcomes Trust. (1994). *Scoring Exercise for the SF-36 Health Survey*. Boston, MA: Medical Outcomes Trust.
- Meyer, K. B., Espindle, D. M., DeGiacomo, J. M., Jenuleson, C. S., Kurtin, P. S., & Davies, A. R. (1994). Monitoring dialysis patients' health status. *American Journal of Kidney Diseases*, *24*, 267–279.
- Meyerboom-DeJong, B., & Smith, R. J. A. (1990). Studies with the Dartmouth COOP charts in general practice: Comparison with the Nottingham Health Profile and the General Health Questionnaire. In M. Lipkin (Ed.), *Functional status measurement in primary care* (pp. 132–149). New York: Springer-Verlag.
- Millhollon, M., & Murray, K. (2001). *Microsoft Word version 2002 inside out*. Redmond, WA: Microsoft Press.
- Montgomery, E. A., & Paranjpe, A. V. (1985). *A report card on HMOs 1980–1984*. Menlo Park, CA: The Henry J. Kaiser Family Foundation.
- Morfeld, M., Bullinger, M., Nantke, J., & Brähler, E. (2005). [The version 2.0 of the SF-36 Health Survey: Results of a population-representative study.] *Sozial- und Präventivmedizin*, *50*, 292–300.
- Mosteller, F., Ware, J. E., Jr., & Levine, S. (1989). Finale panel: Comments on the conference on advances in health status assessment. *Medical Care*, *27*(3, Suppl.), S282–S294.
- Motallebzadeh, R., Bland, J. M., Markus, H. S., Kaski, J. C., & Jahangiri, M. (2006). Health-related quality of life outcome after on-pump versus off-pump coronary artery bypass graft surgery: A prospective randomized study. *Annals of Thoracic Surgery*, *82*, 615–619.
- Mullin, P. A., Lohr, K. N., Bresnahan, B. W., & McNulty, P. (2000). Applying cognitive design principles to formatting HRQOL instruments. *Quality of Life Research*, *9*, 13–27.
- Muthen, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *29*, 177–185.
- National Committee for Quality Assurance. (2004). *HEDIS 2004 (Vol. 6): Specifications for the Medicare Health Outcomes Survey*. Washington, DC: Author.
- Nelson, E., Conger, B., Douglass, R., Gephart, D., Kirk, J., Page, R., ... Zubkoff, M. (1983). Functional health status levels of primary care patients. *Journal of the American Medical Association*, *249*, 3331–3338.
- Nelson, E. C., & Berwick, D. M. (1987). The measurement of health status in clinical practice. *Medical Care*, *27*(3, Suppl.), S77–S90.
- Nelson, E. C., Landgraf, J. M., Hays, R. D., Kirk, J. W., Wasson, J. H., Keller, A., & Zubkoff, M. (1990). The COOP function charts: A system to measure patient function in physicians' offices. In M. Lipkin (Ed.), *Functional status measurement in primary care* (pp. 97–131). New York: Springer-Verlag.
- Nelson, E. C., Landgraf, J. M., Hays, R. D., Wasson, J. H., & Kirk, J. W. (1990). The functional status of patients: How can it be measured in physicians' offices? *Medical Care*, *28*, 1111–1126.
- Nelson, E. C., Wasson, J., Kirk, J. G., Keller, A., Clark, D., Dietrich, A. J., ... Zubkoff, M. (1987). Assessment of function in routine clinical practice: Description of the COOP Chart method and preliminary findings. *Journal of Chronic Diseases*, *40*(Suppl. 1), 55S–69S.
- Newhouse, J. P., & The Insurance Experiment Group. (1993). *Free for all? Lessons from the RAND Health Insurance Experiment*. Cambridge, MA: Harvard University Press.
- Nicholson, B., Ross, E., Sasaki, J., & Weil, A. (2006). Randomized trial comparing polymer-coated extended-release morphine sulfate to controlled-release oxycodone HCL in moderate to severe nonmalignant pain. *Current Medical Research and Opinions*, *22*, 1503–1514.

- Norman, G. R., Sloan, J. A., & Wywich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*, 582–592.
- Nortvedt, M. W., Riise, T., Myhr, K. M., & Nyland, H. I. (2000). Performance of the SF-36, SF-12, and RAND-36 summary scales in a multiple sclerosis population. *Medical Care*, *38*, 1022–1028.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). Blacklick, OH: McGraw-Hill College.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- O'Brien, B. J., Spath, M., Blackhouse, G., Severens, J. L., Dorian, P., & Brazier, J. (2003). A view from the bridge: Agreement between the SF-6D utility algorithm and the Health Utilities Index. *Health Economics*, *12*, 975–981.
- Ochiai, S., Hagino, T., Tonotsuka, H., & Haro, H. (2010). Health-related quality of life in patients with an anterior cruciate ligament injury. *Archives of Orthopaedic Trauma Surgery*, *130*, 397–399.
- Ogles, B. M., Lambert, M. J., & Fields, S. A. (2002). *Essentials of outcomes assessment*. New York: John Wiley & Sons.
- Okamoto, L. J., Noonan, M., DeBoisblanc, B. P., & Kellerman, D. J. (1996). Fluticasone propionate improves quality of life in patients with asthma requiring oral corticosteroids. *Annals of Allergy, Asthma and Immunology*, *76*, 455–461.
- Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment*, *12*, 354–359.
- Parkerson, G. R., Jr., Broadhead, W. E., & Tse, C. K. (1990). The Duke Health Profile: A 17-item measure of health and dysfunction. *Medical Care*, *28*, 1056–1072.
- Parkerson, G. R., Jr., Gehlbach, S. H., Wagner, E. H., James, S. A., Clapp, N. E., & Muhlbaier, L. H. (1981). The Duke-UNC Health Profile: An adult health status instrument for primary care. *Medical Care*, *19*, 806–828.
- Patrick, D. L., Bush, J. W., & Chen, M. M. (1973). Methods for measuring levels of well-being for a health status index. *Health Services Research*, *8*, 228–245.
- Patrick, D. L., & Chiang, Y.-P. (2000). Measurement of health outcomes in treatment effectiveness evaluations: Conceptual and methodological challenges. *Medical Care*, *38*(9, Suppl. II), II14–II25.
- Patrick, D. L., & Deyo, R. A. (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care*, *27*(3, Suppl.), S217–S232.
- Patrick, D. L., & Erickson, P. (1988). Assessing health-related quality of life for clinical decision-making. In S. Walker (Ed.), *Quality of life: Assessment and application* (pp. 9–49). London: MTP Press.
- Patterson, C., Langan, C. E., McKaig, G. A., Anderson, P. M., Maclaine, G. D. H., Rose, L. B., ... Campbell, M. J. (2000). Assessing patient outcomes in acute exacerbations of chronic bronchitis: The Measure Your Medical Outcome Profile (MYMOP), Medical Outcomes Study 6-item general health survey (MOS-6A) and EuroQol (EQ-5D). *Quality of Life Research*, *9*, 521–527.
- Perkins, J. J., & Sanson-Fisher, R.W. (1998). An examination of self- and telephone-administered modes of administration for the Australian SF-36. *Journal of Clinical Epidemiology*, *51*, 969–973.
- Perry, K. T., Freedland, S. J., Hu, J. C., Phelan, M. W., Kristo, B., Gritsch, A. H., ... Schulam, P. G. (2003). Quality of life, pain, and return to normal activities following laparoscopic donor nephrectomy versus open mini-incision donor nephrectomy. *Journal of Urology*, *169*, 2018–2021.
- Pickard, A. S., Johnson, J. A., & Feeny, D. H. (2004). Responsiveness of generic health-related quality of life measures in stroke. *Quality of Life Research*, *13*, 1–13.
- Poole, K., & Mason, H. (2005). Disability in the upper extremity and quality of life in hand-arm vibration syndrome. *Disability and Rehabilitation*, *27*, 1373–1380.
- Prieto, L., Alonso J., & Anto, J. M. (1996). Estimating sample sizes for studies using the SF-36 Health Survey [Letter to the editor]. *Journal of Epidemiology and Community Health*, *50*, 473–474.
- QualityMetric Incorporated. (2010). *U.S. Patent No. 7765113B2*. Washington, DC: U.S. Patent and Trademark Office.
- Raczek, A. E., Ware, J. E., Jr., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., ... Sullivan, M. (1998). Comparison of Rasch and summated rating scales constructed from SF-36 Physical Functioning items in seven countries: Results from the IQOLA Project. International Quality of Life Assessment. *Journal of Clinical Epidemiology*, *51*, 1,203–1,214.
- Ravens-Sieberer, U., Erhart, M., Wetzel, R., Krügel, A., & Brambosch, A. (2008). Phone respondents reported less mental health problems whereas mail interviewees gave higher physical health ratings. *Journal of Clinical Epidemiology*, *61*, 1,056–1,060.
- Razvi, S., Ingoe, L. E., McMillan, C. V., & Weaver, J. U. (2005). Health status in patients with sub-clinical hypothyroidism. *European Journal of Endocrinology*, *152*, 713–717.

- Read, J. L., Quinn, R. J., & Hoefler, M. A. (1987). Measuring overall health: An evaluation of three important approaches. *Journal of Chronic Diseases, 40*(Suppl. 1), 7S–26S.
- Reeve, B. (2004). The National Cancer Institute's conference on improving health outcomes measurement. *International Society for Quality of Life Newsletter, 9*(2), 8.
- Regensteiner, J. G., Ware, J. E., Jr., McCarthy, W. J., Zhang, P., Forbes, W. P., Heckman, J., & Hiatt, W. R. (2002). Effect of cilostazol on treadmill walking, community-based walking ability, and health-related quality of life in patients with intermittent claudication due to peripheral arterial disease: Meta-analysis of six randomized trials. *Journal of the American Geriatric Society, 50*, 1939–1946.
- Revicki, D. A., & Cella, D. F. (1997). Health status assessment for the twenty-first century: Item response theory, item banking, and computer adaptive testing. *Quality of Life Research, 6*, 595–600.
- Ricci, A. R., Yue, J. J., Taffet, R., Catalano, J. B., DeFalco, R. A., & Wilkens, K. J. (2004). Less invasive stabilization system for treatment of distal femur fractures. *American Journal of Orthopedics, 33*, 250–255.
- Russell, A. S., Conner-Spady, B., Mintz, A., Mallon, C., & Maksymowych, W. (2003). The responsiveness of generic health status measures as assessed in patients with rheumatoid arthritis receiving infliximab. *Journal of Rheumatology, 30*, 941–947.
- Ryan, J. M., Corry, J. R., Attewell, R., & Smithson, M. J. (2002). A comparison of an electronic version of the SF-36 General Health questionnaire to the standard paper version. *Quality of Life Research, 11*, 19–26.
- Saleh, K. J., Radosevich, D. M., Kassim, R. A., Moussa, M., Dykes, D., Bottolfson, H., ... Robinson, H. (2002). Comparison of commonly used orthopaedic outcome measures using palm-top computers and paper surveys. *Journal of Orthopaedic Research, 20*, 1146–1151.
- Saris-Baglana, R. N., DeRosa, M. A., Raczek, A. E., Bjorner, J. B., Turner-Bowker, D. M., & Ware, J. E., Jr. (2007). *The SF-10 Health Survey for Children: A user's guide*. Lincoln, RI: QualityMetric Incorporated.
- Saris-Baglana, R. N., Dewey, C. J., Chisholm, G. B., Kosinski, M., Bjorner, J. B., & Ware, J. E., Jr. (2004). *SF Health Outcomes Scoring Software user's guide*. Lincoln, RI: QualityMetric Incorporated.
- Saris-Baglana, R. N., Dewey, C. J., Chisholm, G. B., Plumb, E., Kosinski, M., Bjorner, J. B., & Ware, J. E., Jr. (2007). *QualityMetric Health Outcomes Scoring Software 2.0 user's guide*. Lincoln, RI: QualityMetric Incorporated.
- Saris-Baglana, R. N., Dewey, C. J., Chisholm, G. B., Plumb, E., King, J., Rasicot, P. ... Ware, J. E., Jr. (2011). *QualityMetric Health Outcomes Scoring Software 5.0 user's guide*. Lincoln, RI: QualityMetric Incorporated.
- Schag, C. A. C., Heinrich, R. L., Aadland, R. L., & Ganz, P. A. (1990). Assessing problems of cancer patients: Psychometric properties of the Cancer Inventory of Problem Situations. *Health Psychology, 9*, 83–102.
- Schroeder, S. A. (1987). Outcome assessment 70 years later: Are we ready? *New England Journal of Medicine, 316*, 160–162.
- Scientific Advisory Committee of the Medical Outcomes Trust. (1995). Instrument review criteria. *Medical Outcomes Trust Bulletin, 3*, 1–4.
- Scientific Advisory Committee of the Medical Outcomes Trust. (1996). Development of scientific review criteria. *Clinical Therapeutics, 18*, 979–992.
- Scientific Advisory Committee of the Medical Outcomes Trust. (2002). Assessing health status and quality-of-life instruments: Attributes and review criteria. *Quality of Life Research, 11*, 193–205.
- Sekhon, S., Pope, J., & Baron, M. (2010). The minimally important difference in clinical practice for patient-centered outcomes including health assessment questionnaire, fatigue, pain, sleep, global visual analog scale, and SF-36 in scleroderma. *Journal of Rheumatology, 37*, 591–598.
- Shapiro, M. F., Ware, J. E., Jr., & Sherbourne, C. D. (1986). Effects of cost sharing on seeking care for serious and minor symptoms: Results of a randomized controlled trial. *Annals of Internal Medicine, 104*, 246–251.
- Sherbourne, C. D. (1992). Social functioning: Social activity limitations measure. In A. L. Stewart & J. E. Ware, Jr. (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 173–181). Durham, NC: Duke University Press.
- Sherbourne, C. D., Stewart, A. L., & Wells, K. B. (1992). Role functioning measures. In A. L. Stewart & J. E. Ware, Jr. (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 205–219). Durham, NC: Duke University Press.
- Sidorov, J., Shull, R. D., Girolami, S., & Mensch, D. (2003). Use of the Short Form 36 in a primary care based disease management program for patients with congestive heart failure. *Disease Management, 6*, 111–117.
- Silagy, C. A., Griffin, A. D., Lacey, L. A., & Edmundson, S. (1998). Impact of Zanamivir on productivity, health status and healthcare resource use in patients with influenza. *Clinical Infectious Disease, 27*, 926.

- Silver, G. A. (1990). Paul Anthony Lembcke, MD, MPH: A pioneer in medical care evaluation. *American Journal of Public Health, 80*, 342–348.
- Smith, G. R., Jr., Monson, R. A., & Ray, D. C. (1986). Patients with multiple unexplained symptoms: Their characteristics, functional health, and health care utilization. *Archives of Internal Medicine, 146*, 69–72.
- Snyder, M. K., & Ware, J. E., Jr. (1974). *A study of twenty-two hypothesized dimensions of patient attitudes regarding medical care* (NTIS No. MHC 74 10, PB 239 518/AS 65). Springfield, VA: National Technical Information Service.
- Spiegel, B. M., Younossi, Z. M., Hays, R. D., Revicki, D., Robbins, S., & Kanwal, F. (2005). Impact of hepatitis C on health related quality of life: A systematic review and quantitative assessment. *Hepatology, 41*, 790–800.
- Spratt, K. F. (2009). Patient-level minimal clinically important difference based on clinical judgment and minimally detectable measurement difference: A rationale for the SF-36 Physical Function scale in the SPORT intervertebral disc herniation cohort. *Spine, 34*, 1722–1731.
- Stevic, M. O., Haffer, S. C., Cooper, J., Adams, R., & Michael, J. (2000). How healthy are our seniors? Baseline results from the Medicare Health Outcomes Survey. *Journal of Clinical Outcomes Management, 7*, 39–42.
- Stewart, A. L., Greenfield, S., Hays, R. D., Wells, K., Rogers, W. H., Berry, S. D., ... Ware, J. E., Jr. (1989). Functional status and well-being of patients with chronic conditions: Results from the Medical Outcomes Study. *Journal of the American Medical Association, 262*, 907–913.
- Stewart, A. L., Hays, R. D., & Ware, J. E., Jr. (1988). The MOS Short-Form General Health Survey: Reliability and validity in a patient population. *Medical Care, 26*, 724–735.
- Stewart, A. L., Hays, R. D., & Ware, J. E., Jr. (1992). Methods of validating MOS health measures. In A. L. Stewart & J. E. Ware, Jr. (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 309–324). Durham, NC: Duke University Press.
- Stewart, A. L., & Kamberg, C. J. (1992). Physical functioning measures. In A. L. Stewart & J. E. Ware, Jr. (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 86–101). Durham, NC: Duke University Press.
- Stewart, A. L., & Ware, J. E., Jr. (Eds.). (1992). *Measuring functioning and well-being: The Medical Outcomes Study approach*. Durham, NC: Duke University Press.
- Stewart, A. L., Ware, J. E., Jr., & Brook, R. H. (1981). Advances in the measurement of functional status: Construction of aggregate indexes. *Medical Care, 19*, 473–488.
- Stewart, A. L., Ware, J. E., Jr., Brook, R. H., & Davies-Avery, A. (1978). *Conceptualization and measurement of health for adults in the Health Insurance Study. Volume II: Physical health in terms of functioning* (Publication No. R-1987/2-HEW). Santa Monica, CA: The RAND Corporation.
- Stewart, A. L., Ware, J. E., Jr., Sherbourne, C. D., & Wells, K. B. (1992). Psychological distress/well-being and cognitive functioning measures. In A. L. Stewart & J. E. Ware, Jr. (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 102–142). Durham, NC: Duke University Press.
- Strand, V., Scott, D. L., Emery, P., Kalden, J. R., Smolen, J. S., Cannon, G. W., ... Crawford, B. (2005). Physical function and health related quality of life: Analysis of 2-year data from randomized, controlled studies of leflunomide, sulfasalazine, or methotrexate in patients with active rheumatoid arthritis. *Journal of Rheumatology, 32*, 590–601.
- Strand, V., Tugwell, P., Bombardier, C., Maetzel, A., Crawford, B., Dorrier, C., ... Wells, G. (1999). Function and health-related quality of life: Results from a randomized controlled trial of leflunomide versus methotrexate or placebo in patients with active rheumatoid arthritis. *Arthritis and Rheumatism, 42*, 1870–1878.
- Suris, A., Borman, P. D., Lind, L., & Kashner, T. M. (2007). Aggression, impulsivity, and health functioning in a veteran population: Equivalency and test-retest reliability of computerized and paper-and-pencil administrations. *Computers in Human Behavior, 23*, 97–110.
- Sullivan, M., Karlsson, J., & Ware, J. E., Jr. (1994). *SF-36 Hälsoenkät: Svensk manual och tolkningsguide* [SF-36 Health Survey: Swedish manual and interpretation guide]. Gothenburg, Sweden: Sahlgrenska University Hospital.
- Taft, C., Karlsson, J., & Sullivan, M. (2000). Assessing changes in SF-36 version 2.0: Results from a Swedish population survey. *Quality of Life Research, 9*, 305.
- Taft, C., Karlsson, J., & Sullivan, M. (2001). Do SF-36 summary component scores accurately summarize subscale scores? *Quality of Life Research, 10*, 395–404.
- Tarlov, A. R. (1983). Shattuck lecture—The increasing supply of physicians, the changing structure of the health-services system, and the future practice of medicine. *New England Journal of Medicine, 308*, 1235–1244.

- Tarlov, A. R., Ware, J. E., Jr., Greenfield, S., Nelson, E. C., Perrin, E., & Zubkoff, M. (1989). The Medical Outcomes Study: An application of methods for monitoring the results of medical care. *Journal of the American Medical Association*, *262*, 925–930.
- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum Associates.
- Thissen, D., & Wainer, H. (2001). True score theory: The traditional method. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 23–72). Mahwah, NJ: Lawrence Erlbaum Associates.
- Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.
- Turner-Bowker, D. M., Bartley, P. J., & Ware, J. E., Jr. (2002). *SF-36 Health Survey & SF bibliography* (3rd ed.). Lincoln, RI: QualityMetric Incorporated.
- Turner-Bowker, D. M., DeRosa, M. A., & Ware, J. E., Jr. (2007). SF-36 Health Survey. In S. Boslaugh (Ed.), *Encyclopedia of epidemiology* (pp. 967–972). Thousand Oaks, CA: Sage Publications.
- Tyler, T. A., & Fiske, D. W. (1968). Homogeneity indices and text length. *Educational and Psychological Measurement*, *28*, 767–777.
- U.S. Department of Health and Human Services, Agency for Health Care Policy and Research. (1999). *Health outcomes measures in assessing treatment effectiveness: Measurement models, validation, and interpretation of effects*. AHCPR Health Outcomes Methodology Symposium, Airlie House, VA.
- U.S. Department of Health and Human Services, Food and Drug Administration. (2006). *Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims* (draft guidance). Rockville, MD: Author.
- U.S. Department of Health and Human Services, Food and Drug Administration. (2009). *Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims*. Rockville, MD: Author.
- U.S. Department of Health and Human Services, National Center for Health Statistics (1976). *Health: United States*. Washington, DC: Government Printing Office.
- Valdez, R. B., Ware, J. E., Jr., Manning, W. G., Brook, R. H., Rogers, W. H., Goldberg, G. A., & Newhouse, J. P. (1989). Prepaid group practice effects on the utilization of medical services and health outcomes for children: Results from a controlled trial. *Pediatrics*, *83*, 168–180.
- van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. Berlin, Germany: Springer.
- Veit, C. T., & Ware, J. E., Jr. (1983). The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology*, *51*, 730–742.
- Wachtel, T., Piette, J., Mor, V., Stein, M., Fleishman, J., & Carpenter, C. (1992). Quality of life in persons with human immunodeficiency virus infection: Measurement by the Medical Outcomes Study instrument. *Annals of Internal Medicine*, *116*, 129–137.
- Wagner, A. K., Gandek, B., Aaronson, N. K., Acquadro, C., Alonso, J., Apolone, G., ... Ware, J. E., Jr. (1998). Cross-cultural comparisons of the content of SF-36 translations across 10 countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology*, *51*, 925–932.
- Wagner, A. K., Ehrenberg, B. L., Tran, T. A., Bungay, K. M., Cynn, D. J., & Rogers, W. H. (1997). Patient-based health status measurement in clinical practice: A study of its impact on epilepsy patients' care. *Quality of Life Research*, *6*, 329–341.
- Wainer, H., Dorans, N. J., Flaugher, R., Mislevy, R. J., Thissen, D., Eignor, D., ... Steinberg, L. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Walker, D. R., Landis, D. L., Stern, P. M., & Vance, R. P. (2003). Disease management positively affects patient quality of life. *Managed Care Interface*, *16*, 56–60.
- Walters, S. J. (2004). Sample size and power estimation for studies with health-related quality of life outcomes: A comparison of four methods using the SF-36. *Health and Quality of Life Outcomes*, *2*, 26.
- Walters, S. J., & Brazier, J. E. (2003). What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. *Health and Quality of Life Outcomes*, *1*, 4.
- Walters, S. J., & Campbell, M. J. (2004). The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36). *Health and Quality of Life Outcomes*, *2*, 70.
- Walters, S. J., & Campbell, M. J. (2005). The use of bootstrap methods for estimating sample size and analysing health-related quality of life outcomes. *Statistics Medicine*, *24*, 1075–1102.

- Wang, Y. T., Taylor, L., Pearl, M., & Chang, L. (2004). Effects of Tai Chi exercise on physical and mental health of college students. *American Journal of Chinese Medicine*, *32*, 453–459.
- Ware, J. E., Jr. (1976a). Scales for measuring general health perceptions. *Health Services and Research*, *11*, 396–415.
- Ware, J. E., Jr. (1976b). *The conceptualization and measurement of health for policy-relevant research in medical care delivery* (Publication No. P-5599). Santa Monica, CA: The RAND Corporation.
- Ware, J. E., Jr. (1986). The assessment of health status. In L. H. Aiken & D. Mechanic (Eds.), *Applications of social sciences to clinical medicine and health policy* (pp. 204–228). New Brunswick, NJ: Rutgers University Press.
- Ware, J. E., Jr. (1987). Standards for validating health measures: Definition and content. *Journal of Chronic Diseases*, *40*, 473–480.
- Ware, J. E., Jr. (1988). *How to score the revised MOS Short-Form health scales*. Boston: Institute for the Improvement of Medical Care and Health, New England Medical Center.
- Ware, J. E., Jr. (1990). Measuring patient function and well-being: Some lessons from the Medical Outcomes Study. In K. A. Heitgoff & K. N. Lohr (Eds.), *Effectiveness and outcomes in health care: Proceedings of an invitational conference by the Institute of Medicine Division of Health Care Services* (pp. 107–119). Washington, DC: National Academy Press.
- Ware, J. E., Jr. (1993). Measuring patients' views: The optimum outcome measure. *British Medical Journal*, *306*, 1429–1430.
- Ware, J. E., Jr. (1995). The status of health assessment 1994. *Annual Review of Public Health*, *16*, 327–354.
- Ware, J. E., Jr. (2000). SF-36 Health Survey update. *Spine*, *25*, 3130–3139.
- Ware, J. E., Jr. (2003). Conceptualization and measurement of health-related quality of life: Comments on an evolving field. *Archives of Physical Medicine and Rehabilitation*, *84*(Suppl. 2), S43–S51.
- Ware, J. E., Jr. (2004). The SF-36 Health Survey: An update. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment. Volume 3: Instruments for adults* (3rd ed., pp. 693–718). Mahwah, NJ: Lawrence Erlbaum Associates.
- Ware, J. E., Jr. (2008). Improvements in short-form measures of health status: Introduction to a series. *Journal of Clinical Epidemiology*, *61*, 1–5.
- Ware, J. E., Jr., Bayliss, M. S., Rogers, W. H., Kosinski, M., & Tarlov, A. R. (1996). Differences in 4-year health outcomes for elderly and poor, chronically ill patients treated in HMO and fee-for-service systems. Results from the Medical Outcomes Study. *Journal of the American Medical Association*, *276*, 1039–1047.
- Ware, J. E., Jr., Bjorner, J. B., & Kosinski, M. (1999). Dynamic health assessments: The search for more practical and more precise outcomes measures. *Quality of Life Newsletter*, *21*, 11–13.
- Ware, J. E., Jr., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care*, *38*(Suppl. 9), II73–II82.
- Ware, J. E., Jr., Brook, R. H., Davies, A. R., & Lohr, K. N. (1981). Choosing measures of health status for individuals in general populations. *American Journal of Public Health*, *71*, 620–625.
- Ware, J. E., Jr., Brook, R. H., Davies-Avery, A., Williams, K. N., Stewart, A. L., Rogers, W. H., ... Johnston, S. A. (1980). *Conceptualization and measurement of health for adults in the Health Insurance Study. Volume I: Model of health and methodology* (Publication No. R-1987/1-HEW). Santa Monica, CA: The RAND Corporation.
- Ware, J. E., Jr., Brook, R. H., Rogers, W. H., Keeler, E. B., Davies, A. R., Sherbourne, C. D., ... Newhouse, J. P. (1986). Comparison of health outcomes at a health maintenance organisation with those of fee-for-service care. *Lancet*, *1*, 1017–1022.
- Ware, J. E., Jr., Davies-Avery, A., & Brook, R. H. (1980). *Conceptualization and measurement of health for adults in the Health Insurance Study. Volume VI: Analysis of relationships among health status measures* (Publication No. R-1987/6-HEW). Santa Monica, CA: The RAND Corporation.
- Ware, J. E., Jr., Davies-Avery, A., & Donald, C. (1978). *Conceptualization and measurement of health for adults in the Health Insurance Study. Volume V: General health perceptions* (Publication No. R-1987/5-HEW). Santa Monica, CA: The RAND Corporation.
- Ware, J. E., Jr., Gandek, B., Sinclair, S. J., & Kosinski, M. (2004). *Measuring and improving health outcomes: An SF-36 primer for the Medicare Health Outcomes Survey*. Waltham, MA: Health Assessment Lab and QualityMetric Incorporated.

- Ware, J. E., Jr., Johnston, S. A., Davies-Avery, A., & Brook, R. H. (1979). *Conceptualization and measurement of health for adults in the Health Insurance Study. Volume III: Mental health* (Publication No. R-1987/3-HEW). Santa Monica, CA: The RAND Corporation.
- Ware, J. E., Jr., & Karmos, A. (1976a). *Development and validation of scales to measure perceived health and patient role propensity. Volume II: Final report*. Springfield, VA: National Technical Information Service.
- Ware, J. E., Jr., & Karmos, A. H. (1976b). Scales for measuring general health perceptions. *Health Services Research, 11*, 396–415.
- Ware, J. E., Jr., & Keller, S. D. (1996). Interpreting general health measures. In B. Spilker (Ed.), *Quality of life and pharmacoeconomics in clinical trials* (2nd ed., pp. 445–460). Philadelphia: Lippincott-Raven Publishers.
- Ware, J. E., Jr., & Kosinski, M. (1996). *SF-36 Health Survey (version 2.0)* [Technical note, September 20]. Boston: Health Assessment Lab.
- Ware, J. E., Jr., & Kosinski, M. (2001a). Interpreting SF-36 summary health measures: A response. *Quality of Life Research, 10*, 405–413.
- Ware, J. E., Jr., & Kosinski, M. (2001b). *SF-36 physical and mental health summary scales: A manual for users of version 1* (2nd ed.). Lincoln, RI: QualityMetric Incorporated.
- Ware, J. E., Jr., Kosinski, M., Bayliss, M. S., McHorney, C. A., Rogers, W. H., & Raczek, A. (1995). Comparison of methods for the scoring and statistical analysis of SF-36 health profile and summary measures: Summary of results from the Medical Outcomes Study. *Medical Care, 33*(Suppl. 4), AS264–AS279.
- Ware, J. E., Jr., Kosinski, M., & Bjorner, J. B. (2004). Item banking and the improvement of health status measures [Special issue]. *Quality of Life Newsletter, Fall*, 3–5.
- Ware, J. E., Jr., Kosinski, M., Bjorner, J. B., Bayliss, M. S., Batenhorst, A., Dahlöf, C. G., ... Dowson, A. (2003). Applications of computerized adaptive testing (CAT) to the assessment of headache impact. *Quality of Life Research, 12*, 935–952.
- Ware, J. E., Jr., Kosinski, M., Bjorner, J. B., Turner-Bowker, D. M., Gandek, B., & Maruish, M. E. (2007). *User's manual for the SF-36v2 Health Survey* (2nd ed.). Lincoln, RI: QualityMetric Incorporated.
- Ware, J. E., Jr., Kosinski, M., DeBrotta, D. J., Andrejasich, C. M., & Bradt, E. W. (1995). *Comparison of patient responses to SF-36 Health Surveys that are self-administered, interview administered by telephone, and computer-administered by telephone*. Paper presented at the Eastern Regional Meeting of the American Federation for Clinical Research, Washington, DC.
- Ware, J. E., Jr., Kosinski, M., & Dewey, J. E. (2000). *How to score version 2 of the SF-36 Health Survey (standard and acute forms)*. Lincoln, RI: QualityMetric Incorporated.
- Ware, J. E., Jr., Kosinski, M., Dewey, J. E., & Gandek, B. (2001). *How to score and interpret single-item health status measures: A manual for users of the SF-8 Health Survey*. Lincoln, RI: QualityMetric Incorporated.
- Ware, J. E., Jr., Kosinski, M., & Gandek, B. (2000). *SF-36 Health Survey: Manual & interpretation guide*. Lincoln, RI: QualityMetric Incorporated.
- Ware, J. E., Jr., Kosinski, M., Gandek, B., Aaronson, N. K., Apolone, G., Bech, P., ... Sullivan, M. (1998). The factor structure of the SF-36 Health Survey in 10 countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology, 51*, 1159–1165.
- Ware, J. E., Jr., Kosinski, M., Gandek, B., Sundaram, M., Bjorner, J. B., Turner-Bowker, D. M., & Maruish, M. E. (2010). *User's manual for the SF-12v2 Health Survey* (2nd ed.). Lincoln, RI: QualityMetric Incorporated.
- Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1994). *SF-36 physical and mental health summary scales: A user's manual*. Boston: The Health Institute.
- Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1995). *How to Score the SF-12 physical and mental health summary scales*. Boston: The Health Institute.
- Ware, J. E., Jr., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*, 220–233.
- Ware, J. E., Jr., Kosinski, M., Turner-Bowker, D. M., & Gandek, B. (2002). *How to score version 2 of the SF-12 Health Survey (with a supplement documenting version 1)*. Lincoln, RI: QualityMetric Incorporated.
- Ware, J. E., Jr., Manning, W. G., Duan, N., Wells, K. B., & Newhouse, J. P. (1984). Health status and the use of outpatient mental health services. *American Psychologist, 39*, 1090–1100.
- Ware, J. E., Jr., Miller, W. G., & Snyder, M. K. (1973). *Comparison of factor analytic methods in the development of health related indexes from questionnaire data* (Publication No. PB 239 517. 58). Rockville, MD: Health Services Research Methods Branch.

- Ware, J. E., Jr., Nelson, E. C., Sherbourne, C. D., & Stewart, A. L. (1992). Preliminary tests of a 6-item general health survey: A patient application. In A. L. Stewart & J. E. Ware, Jr. (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 291–308). Durham, NC: Duke University Press.
- Ware, J. E., Jr., & Sherbourne, C. D. (1992). The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care*, *30*, 473–483.
- Ware, J. E., Jr., Sherbourne, C. D., & Davies, A. R. (1992). Developing and testing the MOS 20-item Short-Form Health Survey: A general population application. In A. L. Stewart & J. E. Ware, Jr. (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 277–290). Durham, NC: Duke University Press.
- Ware, J. E., Jr., Snow, K. K., Kosinski, M., & Gandek, B. (1993). *SF-36 Health Survey manual and interpretation guide*. Boston: The Health Institute.
- Ware, J. E., Jr., & Snyder, M. K. (1975). Dimensions of patient attitudes regarding doctors and medical care services. *Medical Care*, *13*, 669–682.
- Ware, J. E., Jr., Snyder, M. K., McClure, R. E., & Jarrett, I. M. (1972). *The measurement of health concepts* (NTIS No. PB-293-508/AS). Springfield, VA: National Technical Information Service.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.
- Weinberger, M., Oddone, E. Z., Samsa, G. P., & Landsman, P. B. (1996). Are health-related quality-of-life measures affected by the mode of administration? *Journal of Clinical Epidemiology*, *49*, 135–140.
- Weiss, D. J. (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.
- Wells, K. B., Hays, R. D., Burnam, M. A., Rogers, W. H., Greenfield, S., & Ware, J. E., Jr. (1989). Detection of depressive disorder for patients receiving prepaid or fee-for-service care: Results from the Medical Outcomes Study. *Journal of the American Medical Association*, *262*, 3298–3302.
- Wenger, N. K., Mattson, M. E., Furberg, C. D., & Elinson, J. (1984). Assessment of quality of life in clinical trials of cardiovascular therapies. *American Journal of Cardiology*, *54*, 908–913.
- Wetzler, H. P., Lum, D. L., & Bush, D. M. (2000). Using the SF-36 Health Survey in primary care. In M. E. Maruish (Ed.), *Handbook of psychological assessment in primary care settings* (pp. 583–621). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wiklund, I., Lindvall, K., Swedberg, K., & Zupkis, R. V. (1987). Self assessment of quality of life in severe heart failure: An instrument for clinical use. *Scandinavian Journal of Psychology*, *28*, 220–225.
- Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., & Erikson, P. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health*, *8*, 94–104.
- Williams, A. H., Ware, J. E., Jr., & Donald, C. A. (1981). A model of mental health, life events, and social supports applicable to general population. *Journal of Health and Social Behavior*, *22*, 324–336.
- Williams, J. D., & Lindem, A. C. (1976). *A computer program for two-way analysis of variance with multiple covariates (ANCOVA2)*. Grand Forks, ND: University of North Dakota Computer Center.
- Wilson, I. B., & Cleary, P. D. (1995). Linking clinical variables with health-related quality of life: A conceptual model of patient outcomes. *Journal of the American Medical Association*, *273*, 59–65.
- Wilson, A. S., Kitas, G. D., Carruthers, D. M., Reay, C., Skan, J., Harris, S., ... Bacon, P. A. (2002). Computerized information-gathering in specialist rheumatology clinics: An initial evaluation of an electronic version of the Short Form 36. *Rheumatology*, *41*, 268–273.
- Wood, G. C., & McLauchlan, G. J. (2006). Outcome assessment in the elderly after total hip arthroplasty. *Journal of Arthroplasty*, *21*, 398–404.
- Wrennick, A. W., Schneider, K. M., & Monga, M. (2005). The effect of parenthood on perceived quality of life in teens. *American Journal of Obstetrics and Gynecology*, *192*, 1465–1468.
- Wu, A. W., Rubin, H. R., Mathews, W. C., Ware, J. E., Jr., Brysk, L. T., Hardy, W. D., ... Richman, D. D. (1991). A health status questionnaire using 30 items from the Medical Outcomes Study: Preliminary validation in persons with early HIV infection. *Medical Care*, *29*, 786–798.
- Wyrwich, K. W., Fihn, S. D., Tierney, W. M., Kroenke, K., Babu, A. N., & Wolinsky, F. D. (2003). Clinically important changes in health-related quality of life for patients with chronic obstructive pulmonary disease: An expert consensus panel report. *Journal of General Internal Medicine*, *18*, 196–202.

- Wyrwich, K. W., Metz, S. M., Kroenke, K., Tierney, W. M., Babu, A. N., & Wolinsky, F. D. (2006). Interpreting quality-of-life data: Methods for community consensus in asthma. *Annals of Allergy, Asthma and Immunology*, *96*, 826–833.
- Wyrwich, K. W., Nelson, H. S., Tierney, W. M., Babu, A. N., Kroenke, K., & Wolinsky, F. D. (2003). Clinically important differences in health-related quality of life for patients with asthma: An expert consensus panel report. *Annals of Allergy, Asthma and Immunology*, *91*, 148–153.
- Wyrwich, K. W., Nienaber, N. A., Tierney, W. M., & Wolinsky, F. D. (1999). Linking clinical relevance and statistical difference in evaluating intra-individual changes in health-related quality of life. *Medical Care*, *37*, 469–478.
- Wyrwich, K. W., Spertus, J. A., Kroenke, K., Tierney, W. M., Babu, A. N., & Wolinsky, F. D. (2004). Clinically important differences in health status for patients with heart disease: An expert consensus panel report. *American Heart Journal*, *147*, 615–622.
- Wyrwich, K. W., Tierney, W. M., & Wolinsky, F. D. (1999). Further evidence supporting an *SEM*-based criterion for identifying meaningful intra-individual changes in health-related quality of life. *Journal of Clinical Epidemiology*, *52*, 861–873.
- Yoshida, K., Sekiguchi, M., Otani, K., Mashiko, H., Shiota, H., Wakita, T., ... Konno, S. (2011). A validation study of the Brief Scale for Psychiatric Problems in Orthopaedic Patients (BS-POP) for patients with chronic low back pain (verification of reliability, validity, and reproducibility). *Journal of Orthopaedic Science*, *16*, 7–13.

Glossary

Acceptability: the general level of approval for an instrument in field use.†

Accuracy: the degree of conformity of a measure to a standard or a true value.†

Acute: describing a temporary state or condition.*

Affect: an emotional or feeling state.*

Algorithm: the rules that define the numerical coding of responses to survey items, and the formulas for combining item response scores to produce health domain scale and component summary measure scores.

Alpha (coefficient): Cronbach's alpha coefficient, an estimate of internal-consistency reliability based on the average inter-item correlation and number of items.*

Alternate-form reliability: an estimate of reliability based on the correlation between two forms constructed to be equivalent measures (i.e., equal mean, variance, and content) of the same concept.

Alternative forms: two versions of a test that have been shown to be equivalent in eliciting information about the same characteristic or variable.†

Anxiety/depression: feelings of anxiety, nervousness, tenseness, depression, moodiness, and downheartedness.†

Assessment: in the term *health assessment*, a standardized procedure used to quantify an individual's health.*

Attribute: a characteristic of an individual.*

Battery: a collection of measures.*

Behavioral functioning: the performance of normal or usual behaviors and activities, usually observable. Behavioral functioning is distinct from well-being, which pertains to subjective, internal states that cannot be directly observed.†

Bodily pain: the intensity, duration, and frequency of physical pain and limitations in usual activities due to pain, such as headaches or backaches.†

Calibration: the equating of scores with other measures of the same or similar health domains using a common metric (e.g., mean = 50, $SD = 10$), such that a known domain score from one instrument can be used to estimate the score for another instrument.

CAT: see *computerized adaptive testing*

Ceiling effect: a concentration of observed scores at the highest possible scale level (see also *floor effect*).

Chronic: describing a state or condition that is persistent or long lasting, usually more than 3 months.*

Classical psychometric methods: the traditional psychometric methods based on true score theory, which were dominant in the health care field prior to the introduction of item response theory (IRT).

Clinical trial: a study, usually a randomized groups experiment, typically designed to evaluate treatment; referred to as a *controlled trial* if a comparison with another treatment or placebo group is involved.

Closed-ended question: a question that contains specific response options (e.g., *yes* or *no*).*

Coarse measure: a measure that has relatively fewer possible scale levels.*

Cognitive functioning: orientation to time and place, memory, attention span, and alertness.†

Comorbid condition: a condition, in addition to the disease or condition under study, that may account for some or all of the measured health differences.

Complete data MSE: a rule applied to Missing Score Estimation (MSE) that allows for the calculation of a scale score only when the respondent has provided a response for every item representing the scale.

* Definition taken directly (some with minor modifications) from the glossary published in Stewart & Ware (1992).

† Definitions edited from definitions published in Bungay & Ware (1993).

Component: part of a larger concept or construct. For example, anxiety is a component of psychological distress.*

Component summary measure: a summary measure calculated using the scores from *all* eight of the Short Form survey health domains, using the methods of principal components factor analysis.

Computerized adaptive testing (CAT): an adaptive, computer-administered test or survey that selects items that are matched to the respondent's level of health or other construct being measured. The CAT administration proceeds in this manner, evaluating score estimates after each answer until specified stopping rules are met.

Concurrent validity: a form of criterion validity in which the measure being tested and the comparison measure are administered at the same point in time.*

Condition-specific measures: a category of health measures that describe problems, such as low-back pain, or particular interventions or treatments, such as knee replacement or coronary artery bypass graft surgery.†

Confidence interval: an estimate of how likely the observed result is; usually defined in terms of a range between an upper and lower limit and associated with a particular probability (e.g., the 95% confidence interval around a mean is the range of mean scores that would be expected 95% of the time).

Construct: something constructed especially by mental synthesis (e.g., to form a construct of a physical object by mentally assembling and integrating sense-data); a variable that is relatively abstract, as opposed to concrete, and is defined or operationalized in terms of observed indicators. Anxiety is an example of a mental health construct.*†

Construct validity: a process in which validity is evaluated as the extent to which a measure correlates with variables in a manner consistent with theory.*

Content validity: the extent to which a measure or battery represents the universe of measurement objects or domains (i.e., adequacy of coverage).*

Convergent validity: a form of construct validity indicating the strength of association between two methods of measuring the same construct.*

Convergent-discriminant validity: a form of construct validity in which reliability coefficients, convergent validity coefficients, and discriminant validity evidence are simultaneously interpreted (e.g., a multitrait-multimethod matrix of correlations with reliability coefficients in the diagonal).*

Correction for overlap: correction of a correlation coefficient for the inflation due to inclusion of the item in the scale score. A correlation corrected for overlap is the correlation of the item with the sum of other items in the same scale (multitrait scaling analysis). When a correlation coefficient is calculated between an item and the scale it is part of (to determine if the item has convergent validity), the scale is scored with the item omitted in order to remove the bias of correlating the item with itself. The item-scale correlation is said to be corrected for item overlap.*

Correlation: an index of association between two continuous variables.

Criterion validity: the extent to which a measure corresponds to an accurate or previously validated measure of the same concept.*

Cross-validation: testing the usefulness of an operational definition derived from one sample on a second sample.*

Data quality evaluation: a systematic evaluation of responses to items and scales to determine their usefulness in estimating scores.

Descriptive statistics: indicators that characterize score distributions for a particular sample, such as the mean, standard deviation, range, skewness, and percentage missing.*

Dimension: a distinct component of a multidimensional construct that can be theoretically or empirically specified; for example, physical and mental health are dimensions of health.*

Dimensionality: the number and nature of distinct components of a construct.*

Disability: a limitation in the performance of a usual social role.

Discriminant validity: an aspect of construct validity in which a measure is shown to correlate higher with concepts it is intended to measure than with concepts it is not intended to measure.*

Disease-specific measures: a category of health measures of severity, symptoms, or functional limitations that are specific to a particular disease state, condition, or diagnostic grouping (e.g., arthritis or diabetes).*†

Domain: any one of the twelve dimensions of health first defined by Campbell: community, education, family life, friendships, health, housing, marriage, nation, neighborhood, self, standard of living, or work.†

Dynamic Health Assessment (DYNHA): QualityMetric's proprietary CAT software, which uses an item bank that incorporates calibrated questions from the Short Form surveys and other widely used health surveys, standardized scoring algorithms, and modern measurement methods to make health status surveys very short, precise, and valid over a wide range of score levels; for use in risk screening and health outcomes monitoring.

DYNHA: see *Dynamic Health Assessment*

Dysfunction: a limitation or decrement in the performance of usual or normal activities.*

Empirical validity: evidence of validity based on the analysis of data.*

Empirically distinct: an instance in which analysis of data yields evidence that two measures do not have the same interpretation.*

External validity: representativeness or generalizability of results.*

Face validity: extent to which a measure "looks like" what it is intended to measure; whether respondents understand a measure's questions and find the answers appropriate.*†

Face-to-face administration: in-person administration of a questionnaire by an interviewer, as opposed to over the telephone (see also *telephone administration*).

Factor: a latent (unobserved) variable or theoretical construct operationalized in terms of the associations among the indicators in a factor analysis.*

Factor analysis: a multivariate analytic method for testing the extent to which underlying hypothetical constructs are defined by a set of measures. Factor analysis is also used to determine whether a set of measures can be reduced to a smaller set without loss of information.*

Factorial validity: a sophisticated form of construct validity; the extent to which the structural relationship among measures corresponds to their underlying theoretical framework.*

Fixed-form instrument: a test or survey containing pre-selected questions and response choices. The paper-and-pencil versions of all the Short Form instruments are considered fixed-form instruments.

Floor effect: a concentration of observed scores at the lowest possible scale level (see also *ceiling effect*).

Frequency distribution: the number of respondents who score at each level of a scale.*

Full MSE: a rule applied to Missing Score Estimation (MSE) that allows for the calculation of a scale score when the respondent has provided a response for at least one item representing the scale. Additionally, Full MSE allows for the calculation of component summary measure scores when at least seven of the eight health domain scale scores are available; however, the Physical Functioning scale cannot be missing when calculating the Physical Component Summary measure score and the Mental Health scale cannot be missing when calculating the Mental Component Summary measure score.

Functional status: the extent to which individuals currently perform their normal or usual behaviors and activities without limitations due to health problems; often used to refer to a variety of concepts of behavioral functioning and well-being.†

Functioning: the ability of individuals to perform their normal or usual behaviors and activities; usually observable; distinct from well-being, which pertains to subjective, internal states that cannot be directly observed.*

General health perceptions: the beliefs and evaluations of a person's overall health, including current and prior health, health outlook, and resistance to illness.†

General population: the population at large, including sick and well persons, rather than a patient population. General population samples are relatively healthier than patient samples.*

Generic measures: general, as opposed to disease-specific, health assessment measures; a category of health measures that are appropriate for all types of patients as well as general populations and that have reliability and validity to measure health in populations with diverse characteristics.†

Guttman scale: a cumulative scale in which each item consists of increasingly more severe or extreme items (e.g., Can you walk a block? Can you walk a mile? Can you walk several miles?). In a perfect Guttman scale, each person's response to items in the scale can be determined from their total scale score.*

Half-scale rule: rule applied to Missing Score Estimation (MSE) that allows for the calculation of a scale score when the respondent has provided responses for at least half of the items representing the scale.

Health: according to the World Health Organization, a state of complete physical, mental, and social well-being rather than merely the absence of disease or infirmity.

Health assessment: a standardized procedure used to quantify an individual's health.†

Health burden: the total impediment in physical, mental, and social functioning and well-being in the personal evaluation of health.

Health dimension: a theoretical component of health, such as physical or mental.*

Health domain: see *domain*

Health framework: a systematic and comprehensive way of organizing health constructs; a theoretical model that specifies distinct health concepts and how they relate to one another.*

Health indicator: an operational definition of health.*

Health Insurance Experiment (HIE): a randomized experiment conducted by the RAND Corporation between 1974 and 1981.*

Health outlook: expectation for health in the future; for example, as measured by the Health Outlook scale in the Health Perceptions Questionnaire (Ware, 1976a).

Health-related quality of life (HRQOL): personal health status; usually refers to those aspects of people's lives that are dominated or significantly influenced by their mental or physical well-being.†

HOS: see *Medicare Health Outcomes Survey*

Index: an aggregation of two or more distinct health measures into an overall summary measure.*

Internal consistency: the extent to which a set of items in a scale measures the same attribute; also called homogeneity. Score reliability increases with internal consistency.†

Internal consistency reliability: a method for estimating score reliability from the correlations among the items in the scale. Cronbach's alpha coefficient (or coefficient alpha) is an internal-consistency reliability coefficient.*

Internal validity: refers to research designs, not measures; confidence in conclusions drawn regarding relationships (adequacy of controls).*

International Quality of Life Assessment Project (IQOLA): a worldwide effort launched in 1991 to translate, norm, and validate the SF-36 for use in multinational clinical trials and general population studies.

Interpolation: the process of estimating a value (e.g., percentage) associated with a specific score within a given range of scores.

Interval scale: a scale in which the distances between all levels along the scale have known numerical values.*

IQOLA: see *International Quality of Life Assessment Project*

Item: a single question or statement and its standardized set of responses.*

Item analysis: the application of quantitative methods to determine the statistical properties of individual test or survey items; also, a qualitative approach used in conjunction with quantitative approaches to further enhance interpretation of individual patient scores.

Item homogeneity: average inter-item correlation.

Item response theory (IRT): a modern psychometric method comprising a set of statistical models that can be used to analyze several categorical variables measuring the same concept (e.g., survey items from a Short Form health domain scale). IRT provides the psychometric basis for CAT.

Item weights: for some scales, items are given differential emphasis in the scoring rules and are thus weighted unequally. When no weights are assigned, equal weights are assumed.*

Known-groups validity: the usefulness of a measure in distinguishing between (or among) groups of people with known characteristics (most often a kind of construct validity).*

Latent variable: an unobserved construct defined in terms of a weighted linear combination of observed or measured variables.*

Likert scale: a scale evaluated and scored according to the method of summated ratings in which items are summed or averaged to obtain an overall score. Items shown to be linearly related to the total scale score are included.*

Limitation: a problem such as having pain, difficulty, or fatigue upon performance of a particular activity.*

Loading: a correlation between a measure and a factor.*

Long form: a survey in its original, full-length form and content, as opposed to a short-form measure constructed to reproduce the survey with fewer items.

Mean: the average calculated by summing the items and dividing by the number of items.*

Measure: a single-item or multi-item scale or index; can be a nominal, ordinal, interval, or ratio scale; a set of questions and answers that elicit statistically useful and consistent information from individuals. Measure is synonymous with questionnaire, tool, survey, or instrument.*†

Measurement error: random error occurring in the measurement of an attribute; the portion of observed score that is *not* true score.*

Median: the midpoint of a particular score distribution marking the 50th percentile.*

Medical expenditure prediction: the predicted average monthly medical expenses for an individual or group of individuals using SF-36, SF-36v2, SF-12, or SF-12v2 data and patient demographic data.

- Medical Outcomes Study (MOS):** a study launched in 1983 to look at variations in styles of practice and outcomes for patients with chronic conditions treated in different systems of care and to advance the state-of-the-art, patient-based assessment methods for assessing health outcomes.
- Medicare Health Outcomes Survey (HOS):** an annual assessment using the SF-12v2 to measure the physical and mental health of Medicare beneficiaries enrolled in managed care plans, as required by the Centers for Medicare & Medicaid Services (CMS).
- Mental health:** a person's emotional, cognitive, and intellectual status.†
- MID:** see *minimally important difference*
- Minimally important difference (MID):** an important group-level score difference, in the context of patient-reported outcomes (PRO), that emphasizes both the perspective of the patient (rather than the clinician) and the importance of considering many types of evidence, including but not limited to clinical evidence.
- Missing Score Estimation (MSE):** scoring algorithms that make it possible to calculate scale and summary scores for survey respondents who did not answer every item and for whom scores would be missing if only the standard scoring algorithms were available.
- Modern psychometric methods:** psychometric approaches employing IRT or Rasch models, as opposed to classical methods, that result in a score estimate and a reliability coefficient that are specific to a particular score level.
- MOS:** see *Medical Outcomes Study*
- Multitrait scaling:** a method for evaluating scale items that considers both item convergence (whether each item correlates substantially with the scale it is part of) and item discrimination (whether each item correlates significantly higher with the scale it is part of than with other conceptually similar scales).*
- Multitrait-multimethod matrix:** a correlation matrix used to examine convergent and discriminant validity. The matrix contains correlations among two or more constructs (traits) measured in two or more ways (methods), as well as reliability coefficients.
- National Survey of Functional Health Status (NSFHS):** a 1998 health status survey, which included both the SF-36 and SF-36v2 forms, administered to a national sample drawn from the sampling frames maintained by National Family Opinion (NFO) Research. Data from this survey served as the bases for the 1998 norms for the SF-36v2, SF-36, and SF-12v2.
- Nominal scale:** a scale in which the numeric values assigned to scale levels are arbitrary and have no numeric meaning. Categories are classifications rather than ordered values (e.g., 1 = male, 2 = female).*
- Norm:** an empirical benchmark based on the scores obtained for a defined sample (e.g., the general population mean); used in interpreting the score for an individual or group.
- Normative data:** data obtained from unspecialized populations that allows for broad comparisons and interpretations of unlike populations.†
- Norm-based scoring (NBS):** see *T scores*
- NSFHS:** see *National Survey of Functional Health Status*
- Objective:** expressing or dealing with facts or conditions as perceived without distortion by personal feelings, prejudices, or interpretations; the opposite of subjective.†
- Ordinal scale:** a scale in which the numbers reflect levels ordered from most to least with respect to some attribute. The relative distance between each level differs throughout the scale, and the number assigned to each level does not reflect an exact quantity. For example, the rating of health as *excellent*, *good*, *fair*, or *poor* is an ordinal scale.*
- Outcome:** a measure of health used specifically as an endpoint or dependent variable. It can be used in evaluating treatment or health care interventions.†
- Out-of-range:** values that do not correspond to the response codes described in the manual for the survey.
- Pain:** see *bodily pain*
- PAQ:** See *Patient Assessment Questionnaire*.
- Patient Assessment Questionnaire:** one survey used in the Medical Outcomes Study.*
- PAQ baseline sample:** an MOS sample of 3,053 patients who completed the baseline Patient Assessment Questionnaire (PAQ). A subset of these patients was selected to become the MOS panel sample.*
- Patient-reported outcomes (PRO):** outcomes of an intervention as reported from the patient's perspective, usually in the form of scores from scales, summary measures, indexes, or other variables derived from the administration of a patient self-report test or survey.
- Personal evaluation:** a respondent's own rating of his or her health, such as based on the widely used rating of health in terms of *excellent* to *poor* (e.g., see the SF-36v2 General Health scale).
- Physical abilities:** ability to perform everyday activities.†
- Physical functioning:** performance of physical activities such as self-care, walking, climbing stairs, and vigorous activities.†

Physical limitations: limitations in performance of self-care, mobility, and physical activities.†

Pilot study: a small study, usually of a convenience sample, to test preliminary measurement decisions and identify unanticipated problems in fielding the instruments in a study.*

Power: see *statistical power*

Precision: extent to which a measure is capable of detecting small differences.*

Predictive validity: a form of criterion validity in which the hypothesis being tested is whether the measure can forecast the probability of another event (e.g., use of health care services) or a future score.*

PRO: see *patient-reported outcomes*

Product-moment correlation: a widely used index of association between two continuous variables, published by Pearson.

Profile: a graphical or tabular display of scores for multiple scales, as estimated from the administration of a test or survey.

Psychological distress: frequency and intensity of negative psychological states including anxiety, depression, and loneliness.†

Psychological well-being: frequency and intensity of general positive affect, behavioral-emotional control, and feelings of belonging.†

Psychometrics: the psychological theory or technique of mental measurement; the use of tests to measure an attribute of an individual object.*

Psychophysiologic symptoms: physical symptoms that can have either a physical health or mental health cause. For example, loss of appetite can be caused by illness or emotional distress.*

QALYs: see *quality adjusted life years*

Quality adjusted life years (QALYs): a method for combining length of life (mortality) and quality of life (health status) for the purposes of economic evaluations of total health benefit. The quality of a year of life is scored using a preference-based utility index ranging from 0.0 (death) to 1.0 (perfect health states across all domains).

Quality of life: an evaluation of all aspects of our lives, such as where we live, how we live, and how we play. It encompasses such life factors as family circumstances, finances, housing, and job satisfaction.†

Questionnaire: a set of questions for obtaining statistically useful or personal information from individuals; a survey made by the use of a questionnaire that includes standardized questions and response choices. Synonyms are measure, test, tool, survey, or instrument.†

Range: the difference between the highest and lowest observed scores for a given variable; also, the full gamut of levels for a given variable or domain (e.g., from well-being to deathly ill).†

Rating: data obtained from a respondent that are subjective and include an evaluative component. Ratings are based on the standards and preferences of the individual patient.†

Ratio scale: a scale with all the properties of an interval scale but, in addition, has an absolute zero (i.e., the point at which there is none of the property being measured), so that ratios between values are meaningful.*

Raw score: data or numbers in their original state, prior to being statistically manipulated.

Recall period: the interval of time the respondent is instructed to consider in reporting or rating a given health phenomenon (e.g., health during the past 4 weeks).

Recode: to assign new numeric values to response choices, following a predetermined set of rules.

Reliability: the accuracy and precision of a measurement procedure; the extent to which a measure reproduces results on repeated trials; the extent to which a measure is free of measurement error; the ratio of the true score to observed score variance.*†

Report: data obtained from a respondent that gives an objective account of an occurrence, not influenced by emotional or personal prejudice.†

Respondent: the person answering questions or completing a survey.*

Respondent burden: the amount of time and effort required of those completing questionnaires.*

Responder definition: important individual patient-level score differences, in the context patient-reported outcomes (PRO), that emphasize both the perspective of the patient (rather than the clinician) and the importance of considering many types of evidence, including but not limited to clinical evidence; previously referred to as *responder criteria*.

Response choices: categories offered to respondents for use in answering a question.

Response level: a particular choice or category defined by an item or combination of items.*

Response scale: the response choices (numbers and their definitions) presented to a respondent with which to answer a particular question (e.g., 1 = yes, 2 = no).*

Response set: a tendency of respondents to answer questions in patterned ways, irrespective of content (e.g., the tendency to present oneself in a favorable light, the tendency to agree with questions regardless of item content).*

- Role functioning:** the degree to which an individual performs or has the capacity to perform activities typical for a specified age and level of social responsibility, such as working at a job, housework, schoolwork, child care, community activities, or volunteer work. †
- Scale:** an item or aggregation of one or more items designed to elicit information concerning a variable or domain; may be used to refer to a graded series of tests. Items are combined in such a way as to satisfy the rules underlying a scale construction method. In health-related measures where data concerning multiple domains are solicited, groups of questions in a domain or in a portion of a domain are grouped together to create a scale. Scales may then be grouped together to provide an index or indices.* †
- Scale level:** a point on a scale that defines a particular rank order or quantity of the concept being measured (e.g., the 21 levels of the SF-36v2 Physical Functioning scale).
- Scale score:** the result of the aggregation and manipulation of the responses to the individual items in a scale.
- Score conversion:** conversion of observed scores from one metric (e.g., 0–100 scores) to another metric (e.g., *T* scores); also, algorithms that allow a user to convert SF-36v2 or SF-12v2 scores to SF-36 or SF-12 scores, respectively, and to convert SF-36 or SF-12 scores to the scoring metric used for the SF-36v2 or SF-12v2, respectively.
- Scoring rules:** numbers assigned to item responses and, if applicable, the formula for their aggregation into a multi-item scale or index.*
- Self-administration:** when respondents read and answer the questions by themselves, without assistance.*
- Self-report:** questions answered by respondents about themselves, either by self-administration or by responding to an interviewer's question.*
- Sensitivity:** the extent to which a measure detects true differences or changes in the construct being measured.*
- Short form:** a scale constructed, from a subset of items contained in a full-length measure, to be shorter in length (e.g., the 36-item SF-36, the 17-item Duke Health Profile [Parkerson, Broadhead, & Tse, 1990]).
- Skewness:** the extent of asymmetry in a frequency distribution.*
- Social functioning:** the ability to develop, maintain, and nurture mature social relationships, including family, friends, neighbors, marital functioning, and sexual functioning. Often separated into two areas: (a) whether and with what frequency social contacts are occurring and (b) the nature of the person's social network or community. †
- Somatic:** pertaining to the body.*
- Split-half correlation:** administering a test in halves. Each half should obtain the same information, and thus the results for each half should correlate. †
- Stability:** the consistency of the results of a questionnaire on repeated applications; often determined by repeated administrations of a test. †
- Standard:** something established by authority, custom, or general consent as a model or example; criterion; something set up and established for the measure of quantity, weight, extent, value, or quality. †
- Standard deviation (*SD*):** an indicator of dispersion or variation around the mean. The standard deviation is the square root of the variance, which is the average squared deviation around the mean.*
- Standard error of measurement (*SEM*):** a statistic used to determine the confidence interval around an individual score; equal to the standard deviation times the square root of one minus the score reliability.*
- Standardization:** consistency in the wording and content of items and the manner in which they are administered and scored so that the results can be meaningfully compared across all administrations of a test or survey.
- Standardize:** to convert raw scores so that the resulting mean and standard deviation have specific values.*
- Statistical power:** the probability of detecting an effect of a given size under the conditions of a particular study.*
- Subjective:** relating to or determined by the mind as the subject of experience; characteristic of or belonging to reality as perceived rather than as independent of mind; experience or knowledge as conditions by personal mental characteristics or states; peculiar to a particular individual; arising out of or identified by means of one's perception of one's own states and processes. The opposite of objective. †
- Subscale:** a scale within a scale; an analyzable smaller unit of a more inclusive scale or index.*
- Summary score:** an aggregate of scale scores that represent related constructs of health.
- Supplemental norms:** subsets of general norms for tests or surveys, representing normative data specific to a particular subsample of the total sample based on demographic (e.g., age, gender, race, education), clinical (e.g., healthy, diabetic), or other variables of interest (e.g., patients who benefited from treatment).

Telephone administration: interviewer administration of a questionnaire over the telephone, as opposed to in person (see also *face-to-face administration*).

Test-retest reliability: a method of estimating reliability by correlating scores from two different repeated administrations of a test, separated by a short time interval.*

Tracer condition: a medical condition defined in order to have a somewhat homogeneous sample by which to trace the effects of health care interventions. For example, in the MOS the following tracer conditions were defined: hypertension, diabetes, heart disease (myocardial infarction, congestive heart failure), and depression.*

T scores: a simple linear transformation of Short Form health domain scale and component summary measure scores that makes each easier to interpret in relation to population norms. Use of *T* scores transforms 0–100 scores to a metric with a mean of 50 and a standard deviation of 10 in the 2009 U.S. general population; previously referred to as *norm-based scores*.

Utility index: a quantitative summary measure of health status appropriate for use in economic evaluations, such as to determine quality adjusted life years (QALYs). The SF-6D is a utility index that can be scored from any version of the SF-36 or SF-12.

Validity: the extent to which an instrument measures what it is supposed to measure and does not measure what it is not supposed to measure (see also *content validity*, *criterion validity*, *predictive validity*, *concurrent validity*, *face validity*, and *construct validity*).*

Variability: the extent to which all possible scale levels are observed.*

Vitality: feelings of energy, pep, fatigue, and tiredness.†

Well-being: subjective bodily and emotional states; how an individual feels; a state of mind distinct from functioning that pertains to behaviors and activities.*

z score: a standardized score that indicates how far a score deviates from the mean in standard deviation units.

Index

A

- Acute (1-week recall) form 6, 30, 31–32, 42, 65, 97, 107–108, 118–119, 149, 151, 153–157, 159–164, 166, 172, 197, 215, 216, 224, 257, 260, 268, 269–272, 277, 281
 - comparability with standard form 217–218, 224
 - factor loadings 264–266
 - interscale correlations 266–267
 - normative data 232, 238, 239–243, 245
 - sample 232, 236, 237, 238
- Administration 41–53
 - addressing problems and questions 43–45
 - determining eligibility 41–42
 - dos and don'ts 46
 - effects of data collection method 48–50
 - guidelines 42–45
 - modes 23, 30, 45–48, 60
 - effect on results 48–50
 - fax 45, 46
 - interview, face-to-face or telephone 45, 46–47
 - mail-out/mail-back (MO/MB) 45, 46, 47
 - online 46, 47–48
 - paper-and-pencil 45, 47
 - smartphone 46
 - tablet/kiosk 46
- Aggregation *See* Scoring
- Algorithms 10, 35, 55, 57, 59, 67, 211–212, 214, 220, 224, 226, 227, 228, 236–237, 239, 245, 252, 281 *See also* *T scores and Scoring*
- Applications 20–27

B

- Bias 8, 44, 56, 69, 172, 224, 226
- Bodily Pain (BP) 15–16, 31, 200
 - abbreviated item content 17
 - interpretation 73–75, 77–78, 85, 96, 107, 118, 141–142, 161, 174
 - item order 42, 47

scoring 58–59

- Burden of disease 3, 19, 20, 21, 22, 27, 80, 173, 174, 175, 226

C

- Calibration on a common metric 9, 226, 227
- Case studies
 - group data 179–183
 - individual respondent data 186–191
- CAT *See* Computerized adaptive testing
- Ceiling 7, 9, 11, 17, 33, 34, 36, 49, 70, 78, 80, 176, 211, 215, 216, 217, 226, 239, 242, 243, 284
- Chronic diseases and conditions 5, 9, 12, 13, 19, 21–22, 24, 26, 47, 80, 132, 142, 143, 147, 152, 153, 161, 162–163, 164, 171, 173, 174, 175, 226, 241–243, 257, 268, 272, 273
- Coefficients, reliability 78, 216, 255–256, 257, 260
- Completeness of data 33, 66, 67
- Completion time 32, 42, 71
- Component summary measures, Mental (MCS) and Physical (PCS) 6, 16–17, 18, 19, 29, 131, 218–224, 239–240 *See also* Mental Component Summary (MCS) *and* Physical Component Summary (PCS)
- Computerized adaptive testing (CAT) 4, 10–11, 30, 37
- Conceptual framework 12–13, 196, 197–198
- Concurrent validity 264, 272–277
- Condition-specific measure 12, 20, 24, 26, 45, 80
- Confidence intervals (CIs) 17, 35, 75, 78–79, 175–176, 260–261
- Confirmation of the two-component structure 33, 66, 70–71, 220, 224
- Construct validity 211, 220, 263, 264–272
- Content validity 263, 264, 277–279
- Content-based interpretation 83–130 *See also* Interpretation
- Convergent validity 69, 266–267
- Criterion validity 263–264, 272–277

Criterion-based interpretation 131–167 *See also* Interpretation

D

Data entry 46, 56, 57, 66, 67, 68

Data quality evaluation (DQE) 33, 65–72

considerations 65–66

qualitative checks 71

patterned responses 71

results inconsistent with respondent presentation 71

unusually quick or long completion time 71

quantitative checks 66–71

completeness of data 33, 66, 67

confirmation of the two-component structure 33, 66, 70–71

consistent responses 33, 67–68

item discriminant validity 33, 70

item internal consistency 33, 69–70

percentage of estimable scale scores 33, 66, 68–69

responses within range 33, 67

scale reliability 33, 70

Development 196–202, 204–210

Difference *See* Statistical power

Discriminant validity 70, 266–267

Disease burden 3, 19, 20, 21, 22, 27, 80, 173, 174, 175, 226

Disease Impact Project 12–13

Disease-specific measures 12, 20, 24, 26, 45, 80

DYNHA Computerized Adaptive Health Assessments 11, 29, 30, 34, 35, 37, 226

E

Effect size 56, 172, 236, 272, 281

F

F ratio *See F* statistic

F statistic 268

Factor analyses 16, 202, 219, 220, 222, 264–266

Floor 7, 17, 33, 34, 35, 36, 49, 80, 199, 211, 215, 216, 217, 218, 226, 239, 240, 242, 243, 284

Forms, acute (1-week recall) and standard (4-week recall) 6, 30, 31–32, 42, 65, 172, 215, 216, 224, 257, 260, 268, 277, 281 *See also* Acute (1-week recall) form *and* Standard (4-week recall) form

Full Missing Score Estimation (Full MSE) 33, 57, 60, 66, 68–69, 227, 238–239 *See also* Appendix C

G

General Health (GH) 16, 31, 200–201

abbreviated item content 17

interpretation 73–75, 77–78, 85, 96, 107, 118, 143, 162–163, 174

scoring 58–59

General Health Rating Index (GHRI) 11, 200

Generic measures 3, 9, 12–13, 15, 20, 26, 27, 45, 278

H

Half-Scale Rule 33, 57, 226–228

Health domain scales 15–16, 58–59, 131, 198–202, 239–240

Bodily Pain (BP) 15–16, 200

General Health (GH) 16, 200–201

Mental Health (MH) 16, 30, 202

Physical Functioning (PF) 15, 198–199

Role-Emotional (RE) 16, 30, 201

Role-Physical (RP) 15, 30, 199–200

Social Functioning (SF) 16, 201

Vitality (VT) 16, 30, 201

Health Insurance Experiment (HIE) 4–5, 196, 277

Health Outcomes Scoring Software *See* QualityMetric Health Outcomes Scoring Software *and* QualityMetric Health Outcomes Scoring Software 5.0

Health Outcomes Survey (HOS) *See* Medicare Health Outcomes Survey (HOS)

Health status assessment 32, 169, 195

conceptual framework for 12–13, 196, 197–198

context 3–4

disease-specific 12, 20, 24, 26, 45, 80

generic 3, 9, 12–13, 15, 20, 26, 27, 45, 278

improvements 4, 7–10

short form 4–7, 29–31

Health Utility Index *See* SF-6D Health Utility Index

Healthcare Effectiveness Data and Information Set (HEDIS) 6

Health-related quality of life (HRQOL) 6, 7, 12, 13, 23, 24, 26, 29, 30, 132, 137, 138, 143, 146, 147, 151, 164, 203, 231, 232, 263

I

Improvements 4–10, 202–203, 211–212, 214–218, 225–229

Information from other instruments 80–81

Internal consistency reliability 33, 34, 69–70, 216–217, 239, 255, 256–259

International Quality of Life Assessment (IQOLA) Project 5, 8, 29, 34, 66, 197, 201, 203, 228–229 *See also* Translations and adaptations

International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 4, 229

- International Society for Quality of Life Research (ISOQOL) 4
- Interpolation 119, 129–130, 166
- Interpretation 19, 42, 45, 55, 56, 212, 214, 222, 223, 224, 236, 264, 266
- content-based 83–130
 - component summary measures, acute form 97, 107–108
 - component summary measures, standard form 84–85, 93
 - health domain scales, acute form 108, 118–119
 - health domain scales, standard form 93, 96–97
 - criterion-based 131–167
 - component summary measures, acute form 149, 151, 153–157, 159–160
 - component summary measures, standard form 132–133, 137
 - health domain scales, acute form, 160–164, 166
 - health domain scales, standard form 137–138, 141–143, 146–149
 - norm-based 73–81
 - component summary measures 75, 77
 - considerations 73–75, 78–81
 - health domain scales 77–78
 - of group data
 - case studies 179–183
 - considerations 75
 - of individual respondent data
 - case studies 186–191
 - considerations 74–75, 185–186
- Interview scripts 41, 42, 43, 45, 47
- IQOLA Project *See* International Quality of Life Assessment (IQOLA) Project
- IRT *See* Item response theory (IRT)
- Item banks 3, 9, 10–11, 30, 226 *See also* DYNHA Computerized Adaptive Health Assessments
- Item characteristic curves 295–296
- Item pools *See* Item banks
- Item recalibration 58, 200, 201 *See also* Scoring
- Item recoding 58, 200 *See also* Scoring
- Item response theory (IRT) 4, 11, 30, 57, 79, 211, 227–228, 239 *See also* Appendix C
- Items
 - content 15–16, 17, 75, 198–202
 - missing responses 33, 49, 57, 58, 60, 67, 186, 203, 226–228, 238–239
 - multiple responses 57
 - response scales 8, 9, 30, 172, 200, 201, 203, 211, 215
 - scoring 57–58
 - selection and origin 198
 - wording 5, 8, 9, 30, 44, 56, 58, 197, 202, 203, 211, 215, 217, 228–229
- Item-scale correlations 69–70, 215, 217, 239, 256, 257, 266
- Interscale correlations 266–267
- K**
- Known-groups comparisons 263, 267–272, 273, 274
- M**
- Measurement precision *See* Precision
- Medical outcome 3
- Medical Outcomes Study (MOS) 5–6, 11, 69, 197, 201, 231, 232, 233, 266
- Medical Outcomes Trust (MOT) 195, 196, 229
- Medicare Health Outcomes Survey (HOS) 5–6, 22, 24, 69, 171
- Mental Component Summary (MCS) 6, 16–17, 18, 19, 29, 131, 218–224, 239–240
 - development of 224
 - interpretation 75, 77, 84, 85, 93, 107–108, 133, 137, 154–157, 159–160, 173
 - scoring 59–60, 223–224
- Mental Health (MH) 16, 30, 31, 202
 - abbreviated item content 17
 - interpretation 73–75, 77–78, 93, 97, 108, 119, 147–149, 164, 166, 175
 - response scale 8, 30, 203, 211, 215
 - scoring 58–59
- Mental Health Inventory (MHI/MHI-5/MHI-38) 34, 197, 201, 202, 256
- Minimally important difference (MID) 169–177, 245
 - criteria 173–175
 - definition 169
 - general considerations 169–173
 - criterion-/anchor-based approaches 170–172
 - distribution-based approaches 172–173
 - minimally important change (MIC) 172
 - responder definition 175–177
 - values 177
- Missing data 33, 57, 58, 60, 66, 67, 68–69, 186 *See also* Appendix C
 - Complete Data 66, 68–69
 - Full Missing Score Estimation (Full MSE) 33, 57, 60, 66, 68–69, 227, 238–239 *See also* Appendix C
 - Half-Scale Rule 33, 57, 226–228
- Missing Score Estimation (MSE) 6, 33, 55, 57, 60, 66, 67, 68–69, 226–228, 239
- Missing Data Estimation (MDE) xxv, 6 *See also* Missing Score Estimation (MSE)

Missing Score Estimation (MSE) xxv, 6, 33, 55, 57, 60, 66, 67, 68–69, 226–228, 239
 MOS SF-20 22, 69, 197, 199, 200, 201, 202

N

National Committee for Quality Assurance (NCQA) 6, 50
 National Institutes of Health (NIH) 22, 26
 National Survey of Functional Health Status (NSFHS) 6, 214, 228
 Norm-based interpretation 33, 73–81, 231 *See also* Interpretation
 Norm-based scores xxv, 30, 33 *See also* T scores
 Norming study *See* QualityMetric 2009 Norming Study *and* Norms
 Norms, 1998 29, 32, 35, 56, 65, 203, 211–212, 214–215, 224, 232, 236–237, 243, 245, 252–253, 263, 264, 266, 267, 274
 Norms, 2009 9, 33, 35, 56, 75, 83, 84, 93, 97, 108, 131, 132, 137, 149, 160, 236–240, 264, 265, 266, 267, 277, 281
 comparability with 1998 norms 56, 243, 245, 252–253
 data collection 234–236
 demographics 236, 237, 238
 forms 232–234
 norming study 6–7, 9, 10, 79, 131, 231–236, 238–239, 260, 272, 274, 277
 sampling 232
 supplemental 240–243
 age 240–241
 disease-specific 241–243
 gender 240–241
 gender-by-age 240–241
 tables
 demographics 237, 238
 disease-specific 244, 246–247, 248–251
 total sample, acute (1-week) form 240, 242
 total sample, standard (4-week) form 240, 241

O

Online scoring service 60–61
 Order effects 50
 Orthogonal components 222–223, 224, 264
 Out-of-range response values 57, 58, 60, 67

P

Pain Impact Questionnaire (PIQ-6) 232, 233
 Partial credit model 295, 296 *See also* Item response theory (IRT)
 Patient-Reported Outcomes Measurement Information System (PROMIS) 26–27

Patient-reported outcomes (PRO) measures 9, 13, 20, 26–27, 46, 169, 196
 Patterned responses 71
 Percentage of estimable scale scores 66, 68–69
 Physical Component Summary (PCS) 6, 16–17, 18, 19, 29, 131, 218–224, 239–240
 development of 224
 interpretation 75, 77, 84–85, 97, 107, 132–133, 151, 153, 173
 scoring 59–60, 223–224
 Physical Functioning (PF) 15, 31, 198–199
 abbreviated item content 17
 interpretation 73–75, 77–78, 84, 93, 97, 107, 108, 118, 137–138, 160–161, 173–174
 scoring 58–59
 Precision 4, 8, 7, 9, 10, 11, 19, 30, 31, 32–33, 78, 169, 170, 175, 197, 199, 200, 211, 226, 228, 240, 255, 256, 281
 Predictive validity 175, 177, 202, 236, 264, 277, 278
 Principal components 16, 215, 219, 220, 222, 223, 224, 264, 265
 Procedure-specific measure 12
 Profile of scores 19, 20, 73–74, 77–78, 80, 212, 213, 214, 226
 Published literature 14, 34, 169, 171, 172, 198, 202, 214, 274, 277

Q

Quality adjusted life years (QALYs) 20, 225
 QualityMetric 2009 Norming Study 6–7, 9, 10, 131, 231–236, 260, 272, 277
 QualityMetric Health Outcomes Scoring Software 6
 QualityMetric Health Outcomes Scoring Software 5.0 10, 33, 57, 58, 60–61, 65, 226 *See also* Appendix A *and* Appendix B

R

Recalibration of items 58, 200, 201 *See also* Scoring
 Recall period 10, 31–32 *See also* Acute (1-week recall) form *and* Standard (4-week recall) form
 Recoding item response values 58, 200
 Relative validity (RV) coefficients 268, 269, 270, 271, 272
 Reliability 7, 9, 17, 33, 34, 35, 56, 70, 78, 176, 255–261
 coefficients 78, 216, 255–256, 257–259
 internal consistency 33, 34, 69–70, 216–217, 239, 255, 256
 new approaches 7, 261
 test-retest 50, 256, 260, 282, 283
 Reliability coefficients 78, 216, 255–260
 Reported Health Transition (HT) item xxv–xxvi, 16, 24
 See also Self-Evaluated Transition (SET) item

- Respondent burden 7, 21, 30, 196, 199, 211, 226, 232, 256
- Responder definition 175–177
- Response Consistency Index (RCI) 55, 60, 67–68
- Responses outside of range 57, 58, 60, 67
- Results inconsistent with respondent presentation 71
- Reverse scoring 58, 68, 70
- Role-Emotional (RE) 16, 30, 31, 201
 - abbreviated item content 17
 - interpretation 73–75, 77–78, 85, 93, 96–97, 108, 119, 146–147, 164, 175
 - response scale 8, 9, 30, 201, 203, 211, 215
 - scoring 58–59
- Role-Physical (RP) 15, 30, 31, 199–200
 - abbreviated item content 17
 - interpretation 73–75, 77–78, 85, 93, 107, 118, 138, 141, 161, 174
 - response scale 8, 9, 30, 200, 203, 211, 215
 - scoring 58–59
- RV coefficients *See* Relative validity (RV) coefficients
- S**
- Sample size 30, 31, 32, 33, 35, 78, 83, 169, 232, 235, 236, 245, 268, 281, 282, 283, 284 *See also* Statistical power
- Score profile 19, 20, 73–74, 77–78, 80, 212, 213, 214, 222, 226
- Scoring 55, 56–60
 - common problems 57
 - component summary measures 59–60
 - health domain scales 58–59
 - items 58
 - reports 47, 60–61 *See also* Appendix A and Appendix B
 - software and services 60–61
 - steps 57–60
- Scoring, 0–100 metric 55–56, 58–59, 171, 212, 214, 243, 245
- Scoring algorithms *See* Algorithms
- Scoring services 35, 58, 60–61, 78, 80, 243, 252
- Scoring software *See* QualityMetric Health Outcomes Scoring Software and QualityMetric Health Outcomes Scoring Software 5.0
- Scoring, *T*-score metric 9, 55–56, 59, 74, 171, 203, 211–212, 215
- Selecting a Short Form survey 31–37
- Self-Evaluated Transition (SET) item xxvi, 16, 24, 58, 172, 202, 220, 232
- SET item *See* Self-Evaluated Transition (SET) item
- SF-6 Health Survey 10
- SF-6D Health Utility Index 3, 6, 19–20, 79, 220, 225, 239–240
- SF-8 Health Survey 10, 30, 226
- SF-10 Health Survey for Children 29, 31
- SF-12 Health Survey 8, 30
- SF-12v2 Health Survey 9–10, 30, 36–37, 225, 226
- SF-20 *See* MOS SF-20
- SF-36 Health Survey 7–8, 29, 195, 197–203, 211–212, 214–218
- SF-36v2 Health Survey 8–9, 29–30, 195
 - background 196–198
 - calibration on a common metric 226, 227
 - comparability of standard and acute forms 217–218
 - comparability with other Short Form Health Surveys 36–37
 - comparability with the SF-12v2 36
 - comparability with the SF-36 56, 202–203, 211–212, 214–218, 224
 - conceptual framework 12–13, 19, 196, 197–198
 - development of 197, 198–202, 204–211
 - improvements 4–10, 202–203, 211–212, 214–218, 225–229
 - items 15–16, 17, 198–202
 - renorming 6–7, 9, 10, 79, 131, 231–236, 238–239, 260, 272, 277
- Short Form family of instruments 7–10
 - applications 20–27, 34–37
 - comparability 36–37
 - features 31–34
 - precision 4, 7, 9, 10, 11, 19, 30, 31, 32–33, 78, 169, 170, 175, 197, 199, 200, 211, 226, 228, 240, 255, 281
 - selecting a form 31–37
 - translations and adaptations 5, 23, 30, 34, 35–36, 42, 66, 201, 202, 203, 228–229
- Smart Measurement System 10, 46, 60
- Social Functioning (SF) 16, 31, 201
 - abbreviated item content 17
 - interpretation 73–75, 77–78, 85, 96, 107–108, 118–119, 146, 164, 174–175
 - scoring 58–59
- Standard (4-week recall) form 6, 30, 31–32, 42, 65, 84–85, 93, 96–97, 132–133, 137–138, 141–143, 146–149, 172, 197, 215, 216, 232, 224, 257, 260, 268, 269, 277, 281
 - comparability with acute form 217–218, 224
 - factor loadings 264–266
 - interscale correlations 266–267
 - normative data 232, 238, 239–243, 245
 - sample 232, 236, 237
- Standard deviation (*SD*) 6, 9, 48, 55–56, 59, 74, 171, 172, 211, 212, 214, 215, 216, 217, 218, 224, 226, 232, 239, 240, 260–261, 281

Standard error of measurement (*SEM*) 73, 78–79, 172, 175–176, 255, 260–261

Standardization 42, 45, 55–56, 78, 212, 268

Statistical power 35, 177, 281–284

and sample sizes 281, 282, 283, 284

and scale measurement properties 283–284

and *T* scores 281–282

experimental studies 281–282

nonexperimental studies 282–283

Step size 175

Supplemental norms *See* Norms

T

Test-retest reliability 50, 256, 260, 282, 283

Translations and adaptations 5, 23, 30, 34, 35–36, 42, 66, 201, 202, 203, 228–229 *See also* International

Quality of Life Assessment (IQOLA) Project

True score variance *See* Variance

T scores xxv, 9, 30, 31, 33, 55–56, 59, 74–75, 77, 79, 83, 131, 212, 214, 226, 243, 245

T-score interpretation *See* Interpretation *and T* scores

U

Unusually quick or long completion time 71

U.S. Food and Drug Administration guidelines 22, 169, 196

Use of information from other instruments 80–81

V

Validity 7, 9, 33, 35, 42, 45, 56, 69, 70, 71, 175, 177, 215, 220, 223, 224, 263–279, 283

construct 211, 220, 222, 263, 264–272

convergent 69, 266–267

discriminant 70, 223, 266–267

factor analyses 264–266

interscale correlations 266–267

item-scale correlations 69–70, 215, 217, 239, 266

known-groups comparisons 263, 267–272

content 263, 264, 277–279

criterion 263–264, 272–277

concurrent 264, 272–277

predictive 175, 177, 236, 264, 277

Variance 70, 200, 215, 219, 222, 223, 224, 264, 265, 266, 268, 281

Vitality (VT) 16, 30, 31, 201

abbreviated item content 17

interpretation 73–75, 77–78, 85, 96, 107, 118, 143, 146, 163, 174

response scale 8, 30, 203, 211, 215

scoring 58–59

W

Weighted maximum likelihood (WML) 296 *See also* Item response theory (IRT)

Weights 59

Wording 5, 8, 9, 30, 44, 56, 58, 197, 202, 203, 211, 215, 217, 228–229

Z

z scores 59 *See also* Scoring

User's Manual for the
**SF-36v2 Health Survey,
Third Edition**

ISBN: 1-891810-28-6



QualityMetric
24 Albion Road Bldg 400
Lincoln, RI 02865
800.572.9394
www.QualityMetric.com