SF-36 Health Survey

Manual & Interpretation Guide

SF-36 Health Survey

Manual and Interpretation Guide

John E. Ware, Jr., Ph.D.

with

Kristin K. Snow, M.S.
Mark Kosinski, M.A.
Barbara Gandek, M.S.

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

Cover and book design by Kate Fallon
Printed by Nimrod Press, Boston, Massachusetts

To Katie and Rick

CONTENTS

LIST OF TABLES

Appendices

ACKNOWLEDGMENTS

Ten years ago, Alvin R. Tarlov, M.D., then at the University of Chicago and now at The Health Institute, New England Medical Center (NEMC); Edward B. Perrin, Ph.D., at the University of Washington; Michael Zubkoff, Ph.D.; and Eugene C. Nelson, Sc.D., at the Dartmouth Medical School; and John E. Ware, Jr., Ph.D., then at The Rand Corporation, now at The Health Institute, began planning for the Medical Outcomes Study (MOS). We gratefully acknowledge their collaboration, faithful service on the MOS Steering Committee, and loyalty to the study for these many years.

Preparation of this manual was supported by The Health Institute, NEMC, from its own research funds. Development and validation of the SF-36 Health Survey was supported by grant # 89-6515 to NEMC from the Henry J. Kaiser Family Foundation, Menlo Park, CA. The MOS was sponsored by The Robert Wood Johnson Foundation, Princeton, NJ; the Functional Outcomes Program of the Henry J. Kaiser Foundation; The Health Institute; The Pew Charitable Trusts, Philadelphia, PA; the Agency for Health Care Policy and Research, Rockville, MD; The National Institute on Aging, Bethesda, MD; and the National Institute of Mental Health, Bethesda, MD.

We gratefully acknowledge the contribution of Colleen A. McHorney, Ph.D.; Martha Bayliss, M.Sc.; Anastasia Raczek, A.B.; Rachel Lu, Sc.M.; and Sue Keller, M.S., Ph.D. Cand., for their efforts in scaling, validation, and norming of the SF-36; William H. Rogers, Ph.D., for his advice on statistical issues; James E. Dewey, Ph.D., of Response Technologies, Inc. in East Greenwich, RI, and Allyson Ross Davies, Ph.D., of the Department of Quality Assessment, NEMC, for their work on the design team for the RT-2000 SF-36 computer software project; Kate Follen for the manual design; Denise R. Short for copyediting; and Kathleen A. Clark, Rebecca Voris, and John Gonsalves for their administrative support.

National norms for the SF-36 were estimated from surveys fielded by the National Opinion Research Center (NORC) under the direction of Richard Rubin and Lisa Thalji, with sponsorship from the Functional Outcomes Program of the Henry J. Kaiser Family Foundation; The Health Institute, NEMC, Boston, MA. The design team for that survey also included Colleen A. McHorney, Ph.D.; Eileen Connor, M.H.S.A.; Sol Levine, Ph.D.; and Edward Schor, M.D. MOS questionnaires that included items chosen or adapted for the SF-36 were fielded and processed by the Survey Research Group at The Rand Corporation under the direction of Sandra H. Berry, M.A. The MOS Functional Status and Well-Being Questionnaire, which

was the source for items selected or adapted for the SF-36, was developed and evaluated by a multi-institutional MOS team as documented in *Measuring Functioning and Well-Being: The Medical Outcomes Study Approach* (Stewart & Ware, 1992).

Several other documents about the SF-36, its scoring, and health status assessment issues in general proved useful in preparing this manual. These include: scoring manuals prepared by the staff at The Health Institute, NEMC, under sponsorship from the John A. Hartford Foundation of New York, entitled *How To Score The MOS 36-Item Short-Form Health Survey (SF-36)* (Ware, 1988; THI, 1991); and *Scoring Exercise for the MOS SF-36 Health Survey* (THI, 1992); and a 1993 publication, *Measuring and Monitoring Health-Related Quality of Life*, co-authored by Kathleen M. Bungay, Pharm. D., and John E. Ware, Jr., Ph.D. Cathy D. Sherbourne, Ph.D. assisted in the preparation of a summary of the selection of items for the SF-36, which was published in *Medical Care* (Ware & Sherbourne, 1992).

The script for the personal administration of the SF-36 reproduced here was adapted from that used by NORC in the 1990 national survey with improvements added by the Department of Quality Assessment, NEMC. The guidelines for administering the SF-36 were adapted from those developed by the "Measuring Health Concepts" research project at Southern Illinois University under the direction of W. Russell Wright, Ph.D., and others at the School of Medicine (Ware et al., 1972).

We acknowledge John E. Brazier, M.Sc.; Crispin Jenkinson, Ph.D.; Andrew M. Garratt, M.Sc.; Michael Marmot, M.D.; and their colleagues and The King's Fund of London for their contributions to the development of the U.K. version of the SF-36. We also thank Hernán Quiñones for his work on the development of the Mexican-American version of the SF-36, and Klemens B. Meyer, M.D., for allowing the reproduction of SF-36 findings from his study of patients receiving hemodialysis treatment.

Finally, we thank the numerous investigators who have used the SF-36 and whose published findings are reproduced here. Their contributions are cited throughout the manual, and their studies are included in the annotated bibliography.

PREFACE

We are in an era during which health care outcomes will increasingly be evaluated from the patient point-of-view. The benefits of specific treatments and of the health care delivery system in general will be judged in terms of the extent to which changes in a patient's functioning or well-being meet his or her needs and expectations (Ware, 1992).

With the new era in mind, my colleagues and I began planning for the Medical Outcomes Study (MOS) over 10 years ago. We had two principal objectives: (1) to determine whether variations in patients' outcomes are explained by differences in the system of health care delivery, provider training, and variations in styles of practice; and (2) to develop more practical tools for the routine monitoring of patient outcomes (Tarlov et al., 1989). We launched the fieldwork for the MOS in 1986.

To advance the state-of-the-art of health assessment methods, we emphasized the development and validation of *short-form* surveys, limited to the most important general health concepts and the most efficient measures of those concepts. Cathy Donald Sherbourne, Allyson Ross Davies, and I constructed an 18-item survey that was included in a 1984 Harris Poll (Louis Harris and Associates, 1984). Two items were added to create the MOS SF-20, which was administered to 11,336 MOS patients in 1986. These surveys demonstrated the potential of well-constructed short-form measures and the usefulness of general population norms in determining the health burden of specific chronic conditions (Stewart et al., 1988, 1989; Ware & Sherbourne, 1992; Wells et al., 1989). MOS studies and dozens of reports published by others identified the strengths and shortcomings of the SF-20. The latter included the omission of important health concepts, scales that were too coarse, and a generally poor fit with advances in an evolving conceptual framework of health.

Demand for an improved short-form survey increased notably following publication of Paul Ellwood's article on outcomes management in the *New England Journal of Medicine* in June 1988, other requests for accountability for outcomes in health care (Relman, 1988; Roper et al., 1988), and passage of a 1989 federal law calling for functional status and well-being assessment. In response, a new 36-item short form (SF-36) was constructed. My objectives in developing the SF-36 were to accomplish the following as quickly as possible: (1) standardization of the content and scoring of an improved form, and (2) accumulation and evaluation of information that would be useful in interpreting results. There were two difficult choices: which concepts to use and whether to construct new scales. I chose eight health concepts, from dozens under study in the MOS, for inclusion in the SF-36. New

short-form scales were constructed to measure five of the eight concepts. The scales were evaluated using data from the MOS.

A prepublication (Developmental) form of the survey and preliminary scoring algorithms were made available to dozens of organizations and to InterStudy, which distributed the form to other groups for testing and general use (Ware, 1988).

The information necessary for the first objective — standardization of content and scoring — was gathered and digested by the fall of 1990, after a 2-year developmental period. That period ended with administration of the standardized form to a representative sample of the U.S. population and its first administration in the MOS, in a fourth-year follow-up survey of participants, both in the fall of 1990. A summary of the development of the SF-36 and its relationship to MOS long-form measures and other instruments that preceded it was submitted for publication in January 1991, the review and publication process required 18 months (Ware & Sherbourne, 1992).

Meanwhile, a scoring manual for the Developmental version was made available (Ware, 1988); and a scoring manual for the Standard version of the SF-36, *How To Score The MOS 36-Item Short-Form Health Survey (SF-36)* (THI, 1991), was prepared. More recently, the *Scoring Exercise for the MOS SF-36 Health Survey*, which included an SF-36 test data set on diskette, was developed to help users evaluate the accuracy of their scoring algorithms in relation to those used in the MOS (THI, 1992).

With the publication of this manual, I seek to begin a partnership with users of the SF-36. The objectives of this manual are: (1) to better meet your needs and to learn from your successes and your failures; (2) to advance understanding of the strengths and weaknesses of the SF-36; (3) to improve the scoring and interpretation of the SF-36; and (4) to increase understanding of how to advance the state-of-the-art of health assessment in general. So that we can share updates with you in a timely manner, I encourage you to fill out and return the registration form (see Appendix D). I hope we become partners in working toward these objectives.

John E. Ware, Jr., Ph.D.

I. HOW TO USE THIS MANUAL

This chapter offers suggestions on how to use this manual and how to find specific information quickly.

Applications of the SF-36

Chapter 2 provides an introductory statement regarding the state-of-the-art of health care assessment, various applications of the SF-36 and health surveys in general, and the importance of maintaining standards of content and scoring as the foundation for reliable and valid interpretation of scale scores. Examples of results from specific applications are presented in *Chapter 11*.

Background and Development

Chapter 3 reviews the background and development of the SF-36. Although some of this information has been summarized in published articles, we attempt to clarify common misunderstandings. *Chapter 3* details the logic behind the selection of items for each SF-36 scale and answers the question "How does each scale differ from its predecessors?" Questions asked most often about each scale are addressed in *Chapter 3*.

Versions of the SF-36

Chapter 3 also compares the content of various versions of the SF-36, including the Standard version, which uses a 4-week recall; the Acute version, which uses a 1-week recall; and the U.K. version. *Chapter 3* documents the evolution of each SF-36 item from the version fielded in the Medical Outcomes Study (MOS), to the Developmental version, and to the Standard version.

Administration Guidelines

The SF-36 was designed to be self-administered, as well as administered by personal interview, or on the telephone. *Chapter 4* presents guidelines for the administration of the SF-36 and discusses personal interviews. See *Appendix C* for the SF-36 script for personal interviews.

Scaling Tests

Chapter 5 explains the application of psychometric theory and methods to the construction and testing of SF-36 scales. Tests of item internal consistency and discriminant validity from the MOS database are summarized in *Chapter 5* along with results published by others.

Scoring

Chapter 6 updates widely used SF-36 scoring guides (THI, 1991; Ware, 1988) with explanations of advances in item and scale scoring algorithms and transformations and guidelines for comparing results across studies.

Reliability, Precision, and Data Quality

Chapter 7 presents estimates of the reliability of SF-36 scores in the MOS and from other sources using internal consistency, test–retest, and alternate forms methods of estimation. *Chapter 7* explains the importance of reliability and other indicators of data quality. Confidence intervals for interpreting individual patient scores are presented in *Chapter 7*, along with estimates of sample sizes required to achieve statistical power using the SF-36 in various study designs.

Validity

Chapter 8 evaluates the content of the SF-36 in relation to other widely used health surveys, explains scale validation methods, summarizes available information based on various validation strategies, and summarizes guidelines for the interpretation of SF-36 scales.

Content- and Criterion-Based Interpretation

Chapter 9 presents guidelines for interpreting differences in SF-36 scales and self-reported health transitions derived from content- and criterion-based findings from the MOS and other published studies.

Norm-Based Interpretation

Chapter 10 presents 21 tables of SF-36 norms for a representative sample of the general U.S. population and descriptive data for various patient groups. The tables include norms and descriptive data for seven age groups, men

and women; clinical populations (including medical and psychiatric patients), and 13 chronic conditions.

Applications

Chapter 11 presents examples of the use and interpretation of the SF-36. Sample results are presented for each application: a general population survey, estimating the burden of different conditions, clinical trials of treatment effects, and monitoring outcomes for individual patients in clinical practice.

Future Directions

Work in progress is summarized in *Chapter 12*. It covers improvements in the content of SF-36 items and scales, advances in scoring, summary health indexes, and generic and disease-specific measures. Translations and cultural adaptations under way in 15 countries and advances in scanning, faxing, and other data processing systems are also discussed.

References and Bibliography

Complete citations for more than 150 referenced publications and an annotated bibliography summarizing the objectives, methods, and results from 30 publications related to the SF-36 are included at the end of this manual.

Appendices

Several forms (fax, computer scanning, and booklet) of the Standard SF-36; measures used during the development of the SF-36; British adaptations of the SF-36 for use in the United Kingdom; the Acute version of the SF-36; and sample items from the Mexican-American SF-36 (Version 1.0) are appended. The Appendices also contain frequency distributions for transformed scores in the general U.S. population and a script for personal and telephone administrations of the SF-36.

2. INTRODUCTION*

During the past decade, one of the more important developments in the health care field has been the recognition of the centrality of the patient 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 that patient's needs or expectations. This sentiment was well expressed in the medical literature much earlier in this century (Codman, 1914; Lembcke, 1952, cited in Silver, 1990). Nearly 40 years ago, Lembcke 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.

More recently, these objectives were echoed 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). Although the patient is the best source of information regarding the achievement of these goals, information from patients about their experiences of disease and treatment has not been routinely collected in clinical research or medical practice. Since this information is not a part of the medical record, it is unavailable for analysis in the current health care database.

In the new era that we are now entering, information about functional status, well-being, and other important health outcomes will be used by policy analysts, who compare the costs and benefits of competing ways of organizing and financing health care services, and by managers of health case organizations, who seek to produce the best value for each health care dollar. The information will also be utilized by clinical investigators who are evaluating new treatments and technologies and by practicing physicians and other providers who are trying to achieve the best possible patient outcomes. The primary source of new information on general health outcomes is rapidly

* An edited version of this introduction about the MOS survey methods in general appeared elsewhere (Ware, 1992); this version attempts to place the SF-36 into a broader context.

becoming standardized patient surveys that have been serving research effectively during the past decades.

Several advances in methods for assessing patient perspectives about functional status, well-being, and other important health care outcomes have been presented during the past decade. These advances have been the subject of numerous conferences (Katz, 1987; Lohr, 1989, 1992; Lohr & Ware, 1987; Wenger et al., 1984). Some significant advances are: (1) an improved understanding of the major dimensions of health and the validity of specific measurement scales in relation to those dimensions (Hays & Stewart, 1990; Liang, 1986; Ware et al., 1981); (2) demonstration of the usefulness of standardized health surveys in clinical trials (Bombardier et al., 1986; Croog et al., 1986; Fowler et al., 1988); (3) health policy evaluations (Brook et al., 1983; Ware et al., 1986); (4) general-population health surveys (Bergner et al., 1981; Stewart et al., 1988, 1989; Ware et al., 1986); and (5) medical practice (Nelson & Berwick, 1989). Until now, it has been impractical to use most of these assessment methods on a large-scale basis. Questions remain regarding the appropriateness of these methods for use among patients with chronic conditions and for patients in all age ranges.

The Improvement of Health Status Surveys

The use of standardized surveys to assess functional status and well-being can be traced back over 300 years. Methodologic interest, however, has been greatest during the last half of this century (Katz et al., 1963). Most health measures used prior to the 1970s were not based upon methods of scale construction. The psychometric techniques of scale construction, now more widely used in the health field, have been available for most of the past century (Guttman, 1944; Likert, 1932; Thurstone & Chave, 1929). In the past 50 years, psychometric techniques have been used successfully in constructing numerous health status scales (Berki & Ashcraft, 1979; DiCocco & Apple, 1958; Dupuy, 1984; Ware, 1976; Williams & Lindem, 1976).

Both the techniques of constructing measures and their content have changed. The earlier focus of measures was limited to the presence or absence of negative health status, functional limitations, symptoms of disease, and acute and chronic problems. Some health measures still focus exclusively on such negative content (Kaplan, 1989). During the last half of this century, the changing content of published measures of functioning and well-being has been

well documented (McDowell & Newell, 1987; Stewart et al., 1978; Ware et al., 1978, 1979; Wenger et al., 1984).

The Health Insurance Experiment (HIE)

One of the most 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; Valdez et al., 1989; Ware et al., 1986). The goal in the HIE was to construct the best possible scales for measuring a broad array of functional status and well-being concepts for non-aged adults and children. That work was summarized in an eight-volume set of Rand Corporation technical reports and in *Medical Care* (Brook et al., 1979; Eisen et al., 1980). The HIE clearly demonstrated the potential of scales constructed from self-administered surveys as reliable and valid tools for assessing changes in health status in the general population. The HIE left two basic questions unanswered: (1) Can methods of data collection and scale construction such as those used in the HIE work in sicker and older populations?, and (2) Can more efficient scales be constructed? The answer to these questions was the challenge for the Medical Outcomes Study (MOS). 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 the elderly.

The Medical Outcomes Study (MOS)

The MOS surveys, like the HIE surveys, were based on a multidimensional model of health. The MOS surveys were more comprehensive, assessing 40 physical and mental health concepts. The SF-36 Health Survey was constructed to represent eight of the most important health concepts included in the MOS and other widely used health surveys.

The SF-36 is referred to as a generic measure because it assesses health concepts that represent basic human values that are relevant to everyone's functional status and well-being (Ware, 1987, 1990a). Such measures are called generic not only because they are universally valued but also because they are not age, disease, or treatment specific. Generic health measures assess health-related quality of life outcomes, namely, those known to be most directly affected by disease and treatment.

The SF-36 survey of generic health concepts is a promising tool for monitoring the results of care. Prior to the SF-36, a comprehensive array of generic functional status and well-being measures had not received widespread adoption, nor had one been shown to be suitable for use across diverse populations and health care settings. As a result, the opportunity to describe differences in functioning and well-being for both the sick and the well was lost. Little was known about how patients suffering from one chronic medical or psychiatric condition differed from each other in terms of functional status and well-being. (One noteworthy exception is the Sickness Impact Profile [Bergner et al., 1981].) The SF-36 provides a common yardstick to compare those patients with chronic health problems to those sampled from the general population.

Generic health measures are not designed nor intended to serve as substitutes for traditional measures of clinical endpoints. To the contrary, the greatest advances in this field during the next decade are likely to come from studies that test generic health measures in parallel with clinical measures. The potential of such comparisons is illustrated in the profiles of functional status and well-being for patients with different medical and psychiatric conditions and in contrast to the general U.S. population (see Chapter 10). These comparisons serve at least two important purposes. The comparisons test the validity of SF-36 scales in describing groups of patients known to differ in functional status and well-being. These comparisons also facilitate understanding among clinicians of the meaning of differences in SF-36 scale scores because these diagnostic groups are familiar.

New Standards of Measurement Evaluation

For the SF-36, a new standard of evaluation was established. The MOS team has 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 predictive tests of validity in relation to subsequent utilization of health care resources. Others have also published such tests and have expanded them to include tests of sensitivity to change over time (Katz et al., 1992).

New standards of measurement evaluation were necessary because the old standards addressed the wrong questions for the MOS approach. Traditionally, longer measures are generally more reliable and more valid (Manning

et al., 1982). The best tests, however, are those most clearly approximating the intended use of the measure (Ware, 1990a). The new direction in the assessment of outcomes calls for new standards formulated to address two questions: Which concepts should be measured?, and How much measurement is enough for each concept and for a particular purpose?

Considerations of respondent burden and of the costs of data collection caused rethinking of measurement goals and, accordingly, the criteria used to construct and evaluate standardized health surveys. It is no longer adequate for a battery of health measures to excel in relation to traditional psychometric standards of reliability, validity, and precision. Today's new opportunities to measure health status routinely demand the best compromise between traditionally defined psychometric elegance and the new standard of feasibility and practicality. The SF-36 attempts to achieve reductions in respondent burden without sacrificing measurement precision below the critical level. This reduction was accomplished by constructing scales from more efficient items. In the HIE, for example, 25 items were necessary to define seven levels of physical functioning (Stewart et al., 1978). With the SF-36 Physical Functioning scale, only 10 items are necessary to define 21 levels of functioning (Stewart & Kamberg, 1992).

The SF-36 is practical because it is shorter. Lengthier research tools served as a point of departure in developing the SF-36. Short- and long-form measures that differed greatly in respondent time and in the cost of collecting and processing data were tested.

The SF-36 is also practical because, for the great majority of respondents, it can be self-administered. The reliance on self-administration as the primary mode of data collection, even for surveys with more than 250 questions, was based in part on the successful use of relatively lengthy self-administered questionnaires in the HIE (Ware, Brook, et al., 1980). However, the MOS population studied in developing the SF-36 was much older and sicker than the HIE population. Half of those in the MOS longitudinal panel were sixty years or older, about 40% were eligible for Medicare, and all had one or more chronic conditions. Self-administered surveys were adopted for use in the MOS on the strength of pilot studies in which self-administration worked well using standard survey methods.

In summary, factors limiting the rate of progress in monitoring health

outcomes from the patient point of view have included the absence of measurement tools with good psychometric properties that are easily administered and well documented. The SF-36 offers one approach for achieving these objectives. Standardization of SF-36 content and scoring will make meaningful interpretation and comparison of results across studies possible.

3. DEVELOPMENT OF THE SF-36

Background and history of major concepts

Interest in short-form health surveys became necessity during the Health Insurance Experiment (HIE) when some study participants refused to complete a lengthy health survey (Ware, Brook et al., 1980). So that they would not be lost to follow-up, we developed a very short survey that could be administered in about 5 minutes by telephone. This strategy worked well in gaining cooperation and yielded preliminary data supporting the use of short-form scales.

Subsequently, several of those short-form scales were used successfully (Brook et al., 1987; Davies & Ware, 1981; Fowler et al., 1988; Lurie et al., 1984; Nelson et al., 1983, 1987; Read et al., 1987; Spiegel et al., 1988). Other analyses of HIE data demonstrated that a well constructed multi-item scale, even a scale 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 a more comprehensive questionnaire achieved a higher level of validity in predicting subsequent medical expenditures than a relatively short multi-item scale (Manning et al., 1982). These analyses underscore the trade-offs involved in the choice between short- and long-form scales.

Our first attempt to construct a *comprehensive* short-form health survey was an 18-item form measuring physical functioning, role limitations due to poor health, general mental health, and current health perceptions. It was constructed for a 1984 national survey fielded by Louis Harris and Associates (Montgomery & Paranjpe, 1985). That survey, which was developed from items that had been used successfully in previous studies, is described elsewhere (Ware, Sherbourne, & Davies, 1992).

In 1986, we added two items measuring social functioning and bodily pain to create a 20-item short form, SF-20 (Ware, Sherbourne, & Davies, 1992). SF-20 was administered to 11,336 Medical Outcomes Study (MOS) participants sampled from 523 practices in Boston, Chicago, and Los Angeles. The resulting data sets have been 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 among patients with chronic medical and psychiatric conditions (Stewart et al., 1988, 1989; Ware, Sherbourne, & Davies, 1992; Wells et al., 1989).

Since the 18-item and 20-item short-forms were first used, we have accumulated considerable experience with the trade-offs between breadth of concepts represented and the depth of measurement for each concept in constructing a short-form health status survey. We have also identified strategies for improving the precision of short-form scales in measuring those concepts.

Conceptual Framework

A survey can be shortened by excluding some health concepts. 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 concepts and multiple manifestations of functioning and well-being for each concept (Ware, 1987, 1990a). From these standards and empirical work to date, multiple categories of operational definitions were chosen to measure each health concept: (a) behavioral functioning, (b) perceived well-being, (c) social and role disability, and (d) personal evaluations (perceptions) of health in general. Table 3.1 shows the physical and mental health phenomena assumed to be represented by each SF-36 scale.

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 psychological state that cannot be completely inferred from observable behavior (Ware, 1987, 1990a). We chose to define perceived well-being 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

TABLE 3.1 SUMMARY OF HEALTH PHENOMENA CAPTURED BY SF-36 SCALES

		Physical				Mental			
Scale	Label	Function	Well-being	Disability	Pers. Eval.	Function	Well-being	Disability	Pers. Eval.
Physical Functioning	PF	●							
Role–Physical	RP			●					
Bodily Pain	BP		●	●					
General Health	GH				●				●
Vitality	VT		●				●		
Social Functioning	SF			●				●	
Role–Emotional	RE							●	
Mental Health	MH					●	●		

preferences of the individual. Who else is more qualified to evaluate current health status or expectations for health in the future? For this reason, perceptions of health in general (i.e., personal evaluation of current health status, susceptibility to illness, and health outlook), were also included. It is well documented that such evaluations provide a good summary of health status and reflect the impact of specific symptoms and other health states experienced but not captured explicitly by measures in the other three categories. (Davies & Ware, 1981).

Selection and Origin of Items

The content of items in the SF-36 will appear very familiar to those who follow the literature on health assessment. Many of the selected items have their roots in instruments that have been in use for more than 20 years. With our colleagues, we have reviewed the content of various source instruments for measuring limitations in physical, social, and role functioning (Donald & Ware, 1984; Stewart et al., 1978, 1981); general mental health (Veit & Ware, 1983; Ware et al., 1979); and general health perceptions (Davies & Ware, 1981; Ware, 1976; Ware & Karnos, 1976). In fact, it is the accumulation of experience with these full-length scales that makes it feasible to construct useful short-form health scales.

The most difficult task in developing the SF-36 was the selection of a subset of 8 health concepts from the more than 40 concepts and scales studied in the MOS. Among those seriously considered, but not chosen, were measures of health distress, sexual functioning, family functioning, and sleep adequacy.

Summary of concepts and measures

As summarized in Table 3.2, the SF-36 includes one multi-item scale measuring each of eight health concepts: (1) physical functioning, (2) role limitations due to physical health problems, (3) bodily pain, (4) general health, (5) vitality (energy/fatigue), (6) social functioning, (7) role limitations due to emotional problems, and (8) mental health (psychological distress and psychological well-being).

A major problem in the field has been the absence of agreed-upon criteria for constructing and validating health scales. In selecting items for each SF-36 scale, we used the corresponding full-length MOS scale as the criterion. Items in each SF-36 scale were selected to reproduce that "parent" scale as much as possible. Other psychometric standards were also considered. Considerably more data were available for applying these strategies to construct the SF-36 than the SF-20. These forms are compared in Table 3.3 in terms of concepts represented and the numbers of items and scale levels for each concept. Specific strategies for selecting SF-36 items, which varied across concepts, are summarized below.

Physical Functioning

Because of the importance of distinct aspects of physical functioning and the necessity of sampling a 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 the SF-20. First, items were added to better represent levels and kinds of limitations between the extremes, including lifting and carrying groceries; climbing stairs; bending, kneeling, or stooping; and walking moderate distances. As with SF-20 and HIE versions of this scale, only one self-care item was included to represent limitations in self-care activities. Although limitations in self-care activities are very important and can be measured in

TABLE 3.2 INFORMATION ABOUT SF-36 HEALTH STATUS SCALES

Concepts	No. of Items	No. of Levels	Meaning of Scores: Low	Meaning of Scores: High
Physical Functioning	10	21	Limited a lot in performing all physical activities including bathing or dressing due to health	Performs all types of physical activities including the most vigorous without limitations due to health
Role–Physical	4	5	Problems with work or other daily activities as a result of physical health	No problems with work or other daily activities as a result of physical health
Bodily Pain	2	11	Very severe and extremely limiting pain	No pain or limitations due to pain
General Health	5	21	Evaluates personal health as poor and believes it is likely to get worse	Evaluates personal health as excellent
Vitality	4	21	Feels tired and worn out all of the time	Feels full of pep and energy all of the time
Social Functioning	2	9	Extreme and frequent interference with normal social activities due to physical or emotional problems	Performs normal social activities without interference due to physical or emotional problems
Role–Emotional	3	4	Problems with work or other daily activities as a result of emotional problems	No problems with work or other daily activities as a result of emotional problems
Mental Health	5	26	Feelings of nervousness and depression all of the time	Feels peaceful, happy, and calm all of the time
Reported Health Transition	1	5	Believes general health is much better now than one year ago	Believes general health is much worse now than one year ago

Note. Adapted from "The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection" by J.E. Ware & C.D. Sherbourne, 1992, *Medical Care, 30,* 473–483.

considerable detail (Katz et al., 1963, 1970), they are relatively rare in both general and patient populations (Stewart et al., 1978, 1981, 1982a, 1982b, 1988; Ware et al., 1992). Only 7.4% of 11,336 patients screened in doctors offices for the MOS reported limitations in self-care activities. Thus, it is not efficient to routinely administer a lengthy battery of these items in a general health survey.

Second, standardized response choices were revised to estimate the severity of each limitation and thereby increase the precision of scores. Consistent

TABLE 3.3 COMPARISON OF NUMBER OF ITEMS AND SCALE LEVELS IN SF-20 AND SF-36 FORMS

	SF-20		SF-36	
Concepts	No. of Items	No. of Levels	No. of Items	No. of Levels
Physical Functioning	6	7	10	21
Role Functioning	2	3	—	—
Role–Physical	—	—	4	5
Role–Emotional	—	—	3	4
Bodily Pain	1	6	2	11
Current Health Perceptions	5	21	—	—
General Health Perceptions	—	—	5	21
Vitality	—	—	4	21
Social Functioning	1	6	2	9
Mental Health	5	26	5	26
Reported Health Transition	—	—	1	5

Note. Adapted from "The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection" by J.E. Ware & C.D. Sherbourne, 1992, *Medical Care, 30*, 473–483.

with HIE methods, each item in SF-20 measured the duration (more or less than 3 months) of any limitation reported. Because the great majority of physical limitations are chronic, measures of duration proved to be of little value in data analysis and have been ignored in scoring items for over 10 years (Stewart et al., 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 & Ware, 1992). Some performance-based measures ignore this distinction (Kaplan & Anderson, 1988), while others do not (Jette et al., 1986). SF-36 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 10 questionnaire items was doubled relative to the number achieved with dichotomous items, and the precision of hypothesis testing was increased without adding to respondent burden. This substantial departure from the SF-20 form was based on empirical evidence of gains in precision with these changes.

Role Functioning

The SF-20 Role Functioning scale was constructed from two widely used questions about health-related limitations in the kind or amount of work (Ware, Sherbourne, & Davies, 1992). The result was a rather coarse, three-level scale. The SF-36 includes a subset of the 11 role functioning items from MOS long-forms. They differ from the SF-20 and other widely used surveys in two important respects (Stewart & Ware, 1992). First, the SF-36 items cover a richer array of role limitations, including: (1) limitations in kind of work or other usual activities, (2) reducing the amount of time spent in work or other usual activities, and (3) difficulty performing work or other usual activities. Thus, in addition to defining more levels of role limitations due to health problems, the two SF-36 Role Functioning scales (Role–Physical and Role–Emotional) are more applicable to retired individuals and those with more than one usual role.

Second, SF-36 items define two scales that distinguish between role limitations due to physical health and mental problems. The latter are sometimes missed by the SF-20 scale and other similar surveys that do not ask explicitly about limitations due to emotional problems (McHorney et al., 1992; Stewart & Ware, 1992). Hypothesized gains in validity have been demonstrated. The two SF-36 Role Functioning scales achieved improved precision in discriminating among groups known to differ in medical and psychiatric conditions (Hays & Stewart, 1990; Sherbourne et al., 1992).

Bodily Pain

The SF-36 retained the SF-20 question about the intensity of bodily pain or discomfort and adds a second item measuring the extent of interference with normal activities due to pain. The latter item was chosen because it is the best predictor ($r = 0.84$) 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 compared to a 6-level scale), relative to the SF-20 version (McHorney et al., 1992).

General Health

SF-20 combined the widely used single-item rating of health (in terms of excellent–poor) and four items from the Current Health scale constructed from the Health Perceptions Questionnaire (HPQ) (Davies & Ware, 1981; Ware, 1976). Although this five-item scale had performed well in studies to

data, a number of potential improvements were achieved with the SF-36 five-item version.

For the SF-36, we chose to reproduce the General Health Rating Index (GHRI) summary score to represent general health, rather than just the Current Health subscale used in SF-20. The SF-36 GH scale: (a) achieves a more comprehensive sample of the content of the HPQ (current health, resistance to illness, and health outlook); (b) correlates highly ($r = 0.96$) with the 22-item GHRI constructed from the HPQ; and (c) is more acceptable to respondents because its items appear less redundant. The GH scale also 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 summary score and other health measures is quite consistent with hypotheses (Davies & Ware, 1981); and GHRI differentiates the impact of serious and minor acute symptoms (Shapiro et al., 1986). It is a good predictor of medical care expenditures (Manning et al., 1982), return to work after a heart attack (Smith et al., 1986), and it proved useful in detecting health outcomes in the HIE (Ware et al., 1986).

Vitality

A four-item measure of vitality (energy level and fatigue), not included in SF-20, was added to better capture differences in subjective well-being. The selected items have an impressive track record in terms of empirical validity and strike a balance between favorably and unfavorably worded items to control for response set effects. They were adapted from the Mental Health Inventory (MHI) fielded in the HIE; the latter was derived from the 1976 HANES survey by the National Center for Health Statistics (Dupuy, 1984). These studies all yielded thorough evaluations of the scale's psychometric properties and documented item-discriminant validity and scale reliability. The scale's sensitivity to the impact of disease and treatment has been demonstrated in recent clinical trials involving patients with hypertension (Croog et al., 1986), prostate disease (Fowler et al., 1988), and those differing in severity of AIDS (Wachtel et al., 1992; Wu et al., 1991).

Social Functioning

In contrast to physical and mental health concepts that tend to "end at the skin" (Ware, 1986), the Social Functioning scale extends measurement beyond the individual to capture both the quantity and quality of social activities with others. The SF-20 included only one social functioning item. The SF-36 retains an improved form of that item and adds 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 activity ask respondents to report the number of contacts and activities or frequency of participation in different activities (Donald & Ware, 1984). They do not usually 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 et al., 1988). To measure health outcomes, SF-36 items ask specifically about the impact of either physical health or emotional problems on social activities. The resulting two-item SF-36 scale defines more levels of social functioning and achieves a higher level of precision (McHorney et al., 1992).

Mental Health

The five-item Mental Health scale (MHI-5) used in SF-20 was retained with modifications only in format. It has been in use for several years (Berwick et al., 1991; Croog et al., 1986; Fowler et al., 1988; Read et al., 1987; Stewart et al., 1988; Stewart & Ware, 1992; Wächtel et al., 1992; Wu et al., 1991). MHI-5 was constructed from the 5 items that best predicted the summary score for the 38-item MHI. It 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). The simple sum of the 5 short-form items (without weights) correlated 0.95 with the full-length 38-item MHI. This correlation was 0.93 on cross-validation using data from the HIE.

The MHI-38, which served as the "gold standard" in constructing MHI-5, is well documented elsewhere (Davies et al., 1988; Veit & Ware, 1983; Ware et al., 1979). Evidence of its 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 et al.,

1984), utilization of general medical services (Manning et al., 1982), and mental health measured after 3 years (Williams et al., 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).

Reported Health Transition

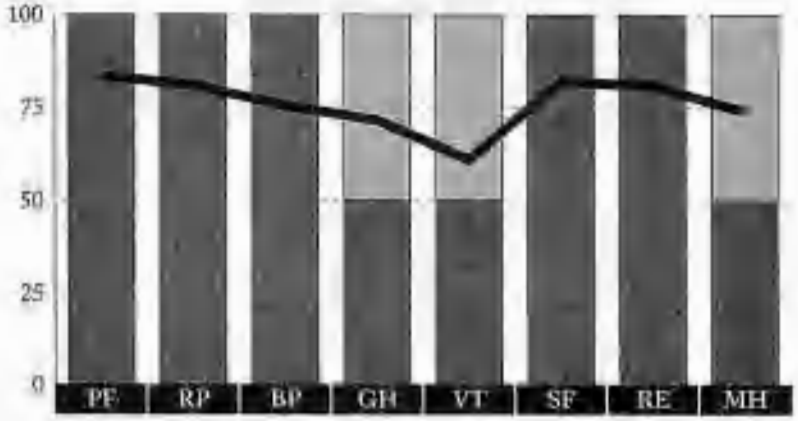
The SF-36 includes a sixth general health rating item, which asks respondents the amount of change in their health in general over a 1-year period. This item is not used to score any of the eight multi-item scales. The transition item can be analyzed as a categorical variable or as an ordinal-level or interval-level scale. As documented in Chapter 9, this item provides useful information about actual changes in health status during the year prior to the administration of the SF-36.

SF-36 health profiles

The SF-36 Health Survey was constructed to achieve 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 disease and other medical conditions, and the effect of treatments on general health status. Figure 3.1 illustrates the SF-36 Profile and calls attention to important features of the eight scales in this regard. As explained in later sections and shown in Figure 3.1, the SF-36 Profile orders scales from left to right from the best physical health measure (Physical Functioning) to the best mental health measure (Mental Health). The validity of these scales and the empirical evidence for this ordering of the scales is presented in Chapters 8 through 10. This ordering facilitates interpretation of the SF-36 profile. Differences on the left side of profiles reflect physical health status and differences on the right side reflect mental health status.

Figure 3.1 also illustrates an important feature about the range of measurement for each of the eight scales. Five scales (PF, RP, BP, SF, and RE) define

FIGURE 3.1
SF-36 HEALTH PROFILE FOR U.S. ADULTS



PF = Physical Functioning
RP = Role-Physical
BP = Bodily Pain
GH = General Health
VT = Vitality
SF = Social Functioning
RE = Role Emotional
MH = Mental Health

Norm

health status as the absence of limitation or disability. For these scales, the highest possible score of 100 is achieved when no limitations or disabilities are observed. These scales are indicated with dark shading in Figure 3.1.

Three of the scales (GH, VT, and MH) are "bipolar" in nature and measure a much wider range of negative and positive health states. For these scales, a score in the mid-range is earned when respondents report no limitations or disability, as illustrated in Figure 3.1 by the dark shading and the light shading. A score of 100 on these bipolar scales is only earned when respondents report positive states and evaluate their health favorably.

The implication of these features of the scales for the "peaks" and "valleys" in the profile of SF-36 scale scores is illustrated by the profile drawn in Figure 3.1, which is the SF-36 profile for the general U.S. population. As would be expected, five of the highest average scores are observed for those scales that require only the absence of limitations to achieve the highest possible score (scales with dark shading from 0 to 100 in Figure 3.1). Further, as would also be expected, the three lowest average scores are observed for those scales that tap the well-being range. These scales require the presence of positive states of health to achieve the highest possible score (see scales with light shading from 50 to 100 in Figure 3.1).

Comparison of MOS short-forms/alternate versions

Five authorized English language versions of the SF-36 have been studied at various stages in its development: (1) the MOS PAQ version, (2) the Developmental (Pre-Publication) version, (3) the Standard (Published) version, (4) the Acute version, and (5) the U.K. version. Persons using one of the above forms, other than the Standard version, can use the SF-36 label but should document that their form was the Developmental, Acute, or U.K. version. Table 3.4 compares instructions, questions, and response choices in the three U.S. versions that use a standard recall period of 4 weeks, so that users of the SF-36 can better understand its development and can determine which form they are using.

For all new studies, we recommend use of the Standard (Published) version of the SF-36 as documented here and as published elsewhere (Ware & Sherbourne, 1992). This form is in the public domain and is the most widely used version in general population studies, clinical trials, and methodological studies. Longitudinal studies already under way with the Developmental version may be best served by continued use of that version. Use of the scoring algorithms in Chapter 6 will maximize comparability of scores based on the Developmental version to those from the Standard version.

MOS PAQ Items Selected for the SF-36

Prior to final selection of concepts and specific items for the SF-36 at THI in fall and winter 1988, versions of candidate items were scattered throughout the 149-item Patient Assessment Questionnaire (PAQ) being used in the MOS (Stewart, Sherbourne, et al., 1992). These items, response choices, and instructions are listed in column three ("original MOS PAQ items") of Table 3.4.

The Standard SF-36 differs from items in the original MOS PAQ in a number of important respects. First, versions of the 36 items in the original MOS PAQ were much longer than the Standard version (844 versus 677 words, about 25% longer). The SF-36 was shortened through adoption of a more efficient response grid format and by not repeating instructions to respondents as often as in the MOS. For example, field tests confirmed that questions 9a through 9i could be substantially shortened without loss of data quality (see Table 3.4).

TABLE 3.4 COMPARISON OF CONTENT IN THE STANDARD SF-36, DEVELOPMENTAL, AND ORIGINAL MOS PATIENT ASSESSMENT QUESTIONNAIRE (PAQ) VERSIONS

Item Number and Concept	SF-36 Standard Version	Developmental (Pre-pub.) Version	Original MOS PAQ Items
Overall 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. Answer every question by marking the answer as indicated.[3] If you are unsure about how to answer a question, please give the best answer you can.	This survey asks for your views about your health. This information will be summarized in your medical record and will help your doctor keep track of how you feel and how well you are able to do your usual activities. 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 **left margin**.	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 will be different. 2. Answer the questions by circling the appropriate number 1, 2 or filling in the answer as requested. 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.
Question 1 GH	In general, would you say your health is:	Same	Same *Page 374, Q1*[4]
Q1 Responses	Excellent Very Good Good Fair Poor	Same	Same
Question 2 RT	**Compared to one year ago**, how would you rate your health in general **now**?[5]	Same	Compared to one year ago, how would you rate your health in general **now**?[5] *PAQ12 Page 3, Q7*
Q2 Responses	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	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	Much better now than one year ago A little better now than one year ago About the same A little worse now than one year ago Much worse now than one year ago
Question 3 PF	The following items are about activities you might do during a typical day. Does **your health now limit you** in these activities? If so, how much?	The following questions are about activities you might do during a typical day. Does **your health** limit you in these activities? If so, how much?	The following items are activities you might do during a typical day. **Does your health limit you** in these activities? *Page 375, Q1*
3a PF	**Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	Same	Same *Page 375, Q1a*
3b PF	**Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	Same	Same *Page 375, Q1b*
3c PF	Lifting or carrying groceries	Same	Same *Page 375, Q1c*

TABLE 3.4 CONTINUED

Item Number and Concept	SF-36 Standard Version	Developmental (Pre-pub.) Version	Original MOS PAQ Items
3d PF	Climbing several flights of stairs	Same	Same *Page 375, Q1d*
3e PF	Climbing one flight of stairs	Same	Same *Page 375, Q1e*
3f PF	Bending, kneeling, or stooping	Same	Same *Page 375, Q1f*
3g PF	Walking more than a mile	Same	Same *Page 375, Q1g*
3h PF	Walking several blocks	Same	Same *Page 375, Q1h*
3i PF	Walking one block	Same	Same *Page 375, Q1i*
3j PF	Bathing or dressing yourself	Bathing and dressing yourself	Same (as Standard SF-36) *Page 375, Q1j*
Q3 Responses	Yes, Limited a Lot Yes, Limited a Little No, Not Limited at All	Same	Same
Question 4 RP	During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?	Same	Same *Page 380, Q1*
4a RP	Cut down the amount of time you spent on work or other activities	Same	Same *Page 380, Q1a*
4b RP	Accomplished less than you would like	Same	Same *Page 380, Q1c*
4c RP	Were limited in the kind of work or other activities	Same	Same *Page 380, Q1e*
4d RP	Had difficulty performing the work or other activities (for example, it took extra effort)	Same	Same *Page 380, Q1f*
Q4 Responses	Yes No	Same	Same
Question 5 RE	During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?	Same	Same *Page 380, Q2*

TABLE 3-4 CONTINUED

Item Number and Concept	SF-36 Standard Version	Developmental (Pre-pub.) Version	Original MOS PAQ Items
5a RE	Cut down the **amount of time** you spent on work or other activities	Same	Same *Page 380, Q8a*
5b RE	**Accomplished less** than you would like	Same	Same *Page 380, Q8b*
5c RE	Didn't do work or other activities as **carefully** as usual	Same	Same *Page 380, Q8c*
Q5 Responses	Yes No	Same	Same
Question 6 SF	During the **past 4 weeks**, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?	Same	Same *Page 379, Q7*
Q6 Responses	Not at all Slightly Moderately Quite a bit Extremely	Same	Same
Question 7 BP	How much **bodily** pain have you had during the **past 4 weeks**?	Same	How much **bodily** pain have you generally had during the **past 4 weeks**? *Page 374, Q9*
Q7 Responses	None Very mild Mild Moderate Severe Very Severe	Same	Same
Question 8 BP	During the **past 4 weeks**, how much did **pain** interfere with your normal work (including both work outside the home and housework)?	Same	Did you experience **any** bodily pain in the **past 4 weeks**? (Yes/No)? *Page 378, Q1* During the **past 4 weeks**, how much did pain interfere with the following things? *Page 379, Q4* Your normal work (including both work outside the home and housework) *Page 379, Q4d*
Q8 Responses	Not at all A little bit Moderately Quite a bit Extremely	Same	Same

TABLE 3.4 CONTINUED

Item Number and Concept	SF-36 Standard Version	Developmental (Pre-pub.) Version	Original MOS PAQ Items
Question 9 VT, MH	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...	These questions are about how you feel and how things have been with you during the past month. 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 past month...	How often during the past 4 weeks...* Page 378, Q13
9a VT	Did you feel full of pep?	Same	Same Page 377, Q1e
9b MH	Have you been a very nervous person?	Same.	How much of the time, during the past month, have you been a very nervous person? Page 396, Q9
9c MH	Have you felt so down in the dumps that nothing could cheer you up?	Same	How much of the time, during the past month, have you felt so down in the dumps that nothing could cheer you up? Page 390, Q16
9d MH	Have you felt calm and peaceful?	Same	How much of the time, during the past month, have you felt calm and peaceful? Page 396, Q16
9e VT	Did you have a lot of energy?	Same	Same Page 377, Q1a
9f MH	Have you felt down-hearted and blue?	Same	How much of the time, during the past month, have you felt down-hearted and blue? Page 390, Q10
9g VT	Did you feel worn out?	Same	Same Page 377, Q1e
9h MH	Have you been a happy person?	Same	How much of the time, during the past month, have you been a happy person? Page 392, Q17
9i VT	Did you feel tired?	Same	Same Page 378, Q11
Q2 Responses	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	Same

TABLE 3.4 CONTINUED

Item Number and Concept	SF-36 Standard Version	Developmental (Pre-pub.) Version	Original MOS PAQ Items
Question 10 SF	During the **past 4 weeks**, how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting with friends, relatives, etc.)?	Has your **health limited your social activities** (like visiting with friends or close relatives)?[f]	During the **past 4 weeks**, how much of the time has your **physical health** or **emotional problems** interfered with your social activities (like visiting with friends, relatives, etc.)?[c] *Page 194, Q4*
Q10 Responses	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 Standard SF-36)
Question 11 GH	How **TRUE** or **FALSE** is **each** of the following statements for you?	Please choose the answer that best describes how **true** or **false** each of the following statements is for you.	Same (as Standard SF-36) *Page 195, Section 5*
11a GH	I seem to get sick a little easier than other people	Same	Same *Page 198, Q21*
11b GH	I am as healthy as anybody I know	Same	Same *Page 198, Q27*
11c GH	I expect my health to get worse	Same	Same *Page 198, Q24*
11d GH	My health is excellent	Same	Same *Page 198, Q23*
Q11 Responses	Definitely True Mostly True Don't Know Mostly False Definitely False	Definitely True Mostly True Not Sure[f] Mostly False Definitely False	Same (as Standard SF-36)

[a] Page numbers and item number refer to the placement of the original item in the MOS Patient Assessment Questionnaire (PAQ), as shown in the Appendix of Stewart & Ware, 1992.

[b] The RT-2000 form reads "Answer every question by marking the appropriate oval."

[c] Words in boldface are emphasized in the individual questionnaires. The emphasis may either be bold-faced, underlined, or both. These are considered equivalent for the purposes of this comparison.

[d] This item was first fielded in the PAQ12 rather than the PAQ00.

[e] A positive response to this question was necessary for administration of the next two questions (Q8 and Q8d on page 179 of the PAQ).

[f] Vitality questions 9a, 9e, 9g, and 9i were administered separately in one grid, with the mental health questions 9b, 9c, 9d, 9f, and 9h administered as separate items without a grid. The other two forms include both the vitality and mental health items in one grid.

[g] This item was included as part of a grid in Question 9 of the Developmental version along with Mental Health and Vitality items. It is a single item in the SF-36 Standard version.

[h] Empirical studies have shown that the numerical scale values for "Don't Know" and "Not Sure" do not differ significantly. Therefore, we consider the categories interchangeable.

SF-36 Developmental (Pre-Publication) Version

Following selection of concepts and scales and the editing of original MOS PAQ versions of the items, a Developmental (Pre-Publication) version of the SF-36 was distributed by THI beginning in early 1989. This version was made available to InterStudy for testing under its Outcomes Management Systems (OMS) program and to numerous other investigators and projects (Ware, 1988). In the fall of 1990, MOS staff at THI finalized the content and format of the SF-36 as documented in the first column of Table 3.4 (SF-36 Standard version).

Changes include reference to "the past four weeks" rather than "the past month" in response to ambiguity reported by field test participants regarding whether the recall period was the number of days that had elapsed in the current month or a period of time of 4-weeks' duration. Numerous other changes based on the comments of respondents during the 2-year test period were also adopted, including the underlining or boldfacing of key words in the instructions, questions, and response choices. Improvements contributed to the success of the administration and processing of the Standard version of the SF-36. Many individuals and organizations contributed useful suggestions (see Acknowledgments).

As Table 3.4 shows, 35 of the 36 items are virtually identical across the Standard and Developmental versions of the SF-36. Response choices are identical for all but three of the items and instructions to respondents are identical or nearly identical for all of the items. Although numerous changes in the original MOS PAQ versions of SF-36 items and instructions were adopted, very few changes were necessary in making the transition from the Developmental version to the Standard version. Given the wide-spread adoption of the Developmental version, a very high priority was placed on maintaining comparability.

SF-36 Acute Version

Studies of treatments for acute conditions may be better served by using the Acute version of the SF-36. This version uses a "1-week" recall period rather than the "4-week" recall period used in the Standard version. When SF-36 forms are administered weekly or biweekly, it is usually best to use the Acute version so that the recall period will not overlap across administrations.

Many studies using the Acute version are under way and much information

from them is expected in the near future. The effects of the Acute versus the Standard version on responses to SF-36 questions and the sensitivity of the two forms to acute changes in health are the subject of several randomized methodological studies currently under way.

SF-36 U.K. Versions

A number of teams have been working on a British adaptation of the SF-36 for use in Great Britain (Brazier et al., 1992; Garratt et al., 1993; Jenkinson et al., 1993). Their objective has been to determine the extent and nature of revisions necessary to adapt the SF-36 form for use in the United Kingdom. The King's Fund in London established a network of investigators that pooled their experience with the goal of producing a standardized U.K. version. John Brazier at Sheffield University represents the U.K. network as a member of the team of scientists working on the International Quality of Life Assessment (IQOLA) Project. The IQOLA Project is translating and adapting the SF-36 for use in 15 countries (Aaronson et al., 1992).

It is encouraging that after extensive evaluation of the SF-36, the British adaptation required only minor changes involving 5 of the 36 items. As shown in Table 3.5, the U.K. version does not use the word "block" as a distance measure, instead it uses "half a mile" and "100 yards." The other adaptations involve only minor word substitutions: "life" instead of "pep," "low" instead of "blue," and "ill more easily" instead of "sick a little easier."

Because the first adaptations in the United Kingdom were initiated soon after introduction of the Developmental version of the SF-36, early study results were based on the U.K. Developmental version of the SF-36 (reproduced in Appendix B). These studies have yielded a rich database including population norms and profiles for specific conditions (Brazier et al., 1992; Garratt et al., 1993; Jenkinson et al., 1993).

The Whitehall epidemiologic studies, being directed by Dr. Michael Marmot at University College in London, use a British adaptation of the Standard SF-36 version (M. Marmot, personal communication, December 14, 1990). These studies will also yield a vast array of cross-sectional and longitudinal data on the U.K. Standard version (reproduced in Appendix B). Their results will be useful in interpreting scores in terms of the burden of specific diseases and the amount of change in scores over time due to different treatments.

TABLE 3.5 COMPARISON OF SF-36 ITEMS THAT ARE NOT IDENTICAL IN U.S. AND U.K. VERSIONS

Items	U.S. Version	U.K. Version
3h.	Does your health now limit you in these activities?...	
	Walking several blocks	*Walking half a mile*
3i	Does your health now limit you in these activities?...	
	Walking one block	*Walking 100 yards*
9a	How much of the time during the past 4 weeks...	
	Did you feel full of pep?	*Did you feel full of life?*
9f	How much of the time during the past 4 weeks...	
	Have you felt downhearted and blue?	*Have you felt downhearted and low?*
11a	How true or false is each of the following statements for you?...	
	I seem to get sick a little easier than other people	*I seem to get ill more easily than other people*

The U.K. version of the SF-36 is available royalty free from The Medical Outcomes Trust. The British adaptations of SF-36 items described in Table 3.5 have been incorporated in both the Developmental and Standard U.K. versions. The only difference remaining between the two versions is the number of response choices for the second Social Functioning item. This item has six response choices in the Developmental U.K. version and five in the Standard U.K. version.

The choice between one of the two U.K. versions is made easier and the consequences are minimized by the scoring methods discussed in Chapter 6 and by the norms presented in Chapter 10, which include estimates of the

differences between scoring methods. Included are scoring algorithms that equate the two forms and estimates of the effect of differences in scoring on the most affected scale scores.

As shown in figures in Chapter 11, U.S. and U.K. norms for males and females are quite comparable for younger age groups across the U.S. and U.K. populations. Adjustments for differences in the scoring of the Bodily Pain and Social Functioning scales improve the comparability of average scores across the U.S. and U.K. populations (see Chapter 10).

Mexican-American SF-36 (Version 1.0)

Following more than a year of translation work and evaluation, Version 1.0 of the Mexican-American SF-36 has been made available to selected clinical trials and other studies including the National Institutes of Health (NIH) sponsored Breast Cancer Prevention Trial. It appears that Version 1.0 is a satisfactory reproduction of the source SF-36. Evaluation of Version 1.0 continues, and results will be submitted for publication within a year.

The Mexican-American SF-36 (Version 1.0) was translated using the methods of the IQOLA Project (Aaronson et al., 1992). Briefly, the first step involved "forward" translations from English to Spanish by independent professional translators familiar with health-related questionnaires, but not familiar with the SF-36. The target language was Spanish, as spoken and familiar to Spanish-speaking Americans across the United States. Variations across the two forward translations were reconciled by a team that included the translators and IQOLA project scientists. They considered conceptual, cultural, and psychometric issues in agreeing upon the content of the Mexican-American version. Pre-existing Spanish-language translations of the SF-36 and "parent" forms were also evaluated and considered. While the translation was targeted to all Spanish-speaking populations, pretesting of the form was conducted among Mexican-Americans.

The second step, referred to as "backward" translation, involved independent translations from Spanish to English. This step was performed by professional translators unfamiliar with the SF-36. Comparisons between the backward translations and the source SF-36 revealed that some items were not equivalent. They were recycled through the translation process until they were judged to be equivalent. Translation difficulty, quality, and equivalence

to the source were formally evaluated using standardized rating forms. Selected items from the Developmental version 1.0 of the Mexican-American SF-36 are reproduced in Appendix B.

4. ADMINISTRATION OF THE SF-36

Methodological issues

This chapter presents guidelines for administering the SF-36. The SF-36 has been successfully administered to persons age 14 and older using self-administration and interviewer administration by telephone and in-person.

Often, the SF-36 is completed by patients at the time of a doctor or clinic visit. However, the SF-36 can also be administered at home and in many other settings, including telephone interviews, mail-out/mail-back questionnaires, and face-to-face interviews. (The script for telephone and face-to-face administration of the SF-36 is included in Appendix C.) The SF-36 can also be included as one part of a longer interview, questionnaire, or other data collection effort.

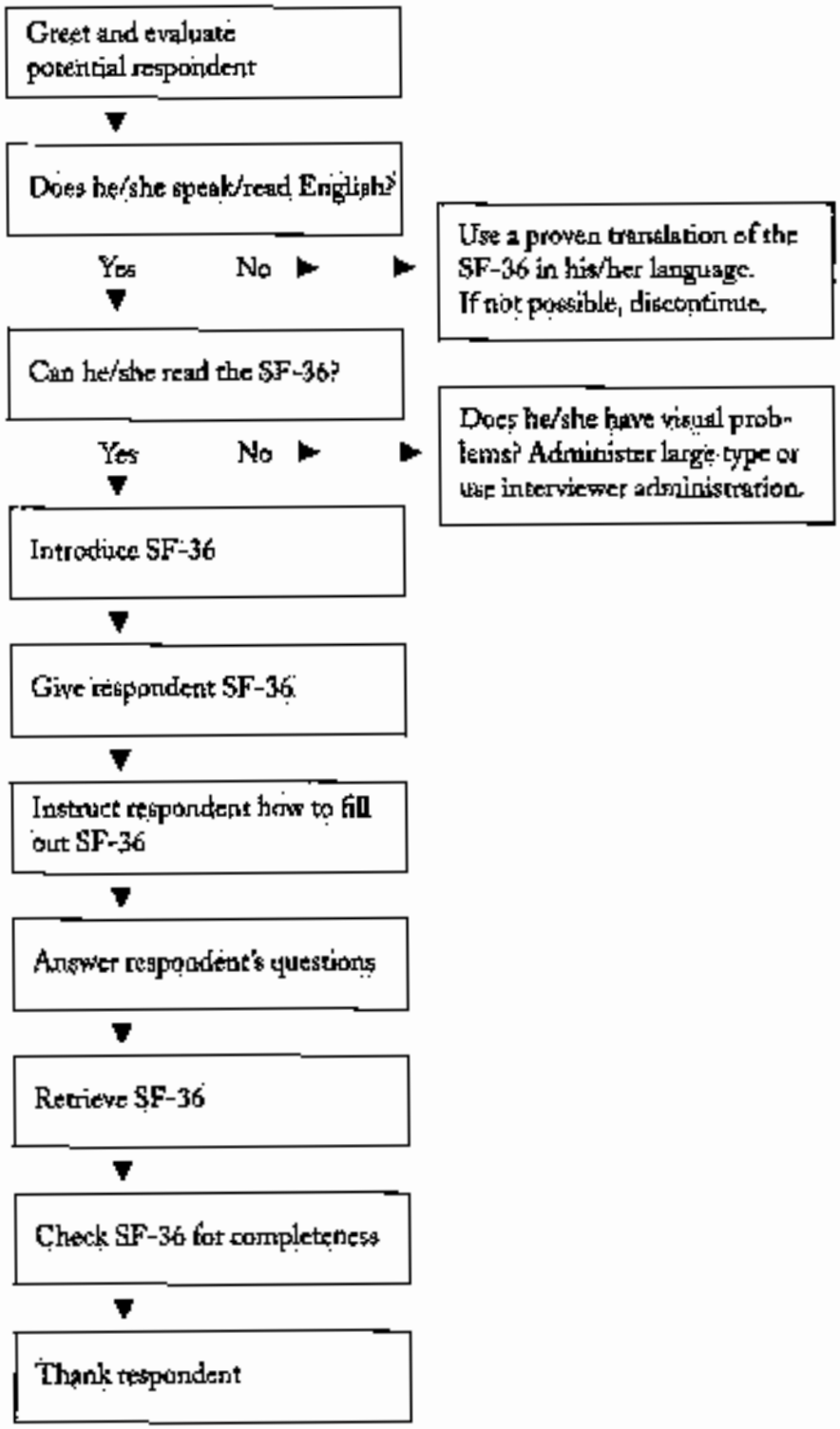
Since most questions on the proper administration of the SF-36 arise in the clinical setting, this chapter specifically addresses these issues. Figure 4.1 presents a flow chart detailing proper administration methods for the SF-36 questionnaire.

These guidelines assume that an appropriate person or persons has been designated to facilitate self-administration of the SF-36, and that a procedure has been established for identifying respondents. The questionnaire administrator is important for establishing rapport with the respondent and encouraging completion of the SF-36. The administrator can emphasize to respondents the importance of their answers to the completion of the study or to the addition to their medical records. The administrator can also answer questions and address concerns about the SF-36, and ensure the questionnaire is filled out correctly and completely. Respondents are more likely to fill out a questionnaire honestly and completely if they have a positive impression or relationship with the administrator.

Identifying the Sample

Because participation requires completion of a self-administered questionnaire, respondents are excluded if they are unable to read the questionnaire

FIGURE 4.1
SF-36 ADMINISTRATION FLOW CHART



due to limited reading ability. Before giving a respondent a questionnaire, determine whether they fall into one of the following two categories:

Foreign/Non-English Speaking Respondents
If a respondent does not speak English, check to see if any information is available regarding the respondent's ability to read English. If it is believed that the respondent is able to read English, offer them the questionnaire. If they are unable to read English, offer a proven translation or record that it was not completed due to a language barrier.

Reading Ability
Check to see if any information is available regarding the respondent's ability to read. If a respondent is unable to read, do not offer them the questionnaire. Again, record that it was not self-administered due to reading ability.

If the SF-36 is to be given to a large number of respondents who are unable to read, then the interviewer-administered accompanying script should be used (see Appendix C).

If a study is expected to have a large number of respondents who have visual problems, a large-type version of the questionnaire should be prepared. It should be noted that the printing of special forms adds to the cost and complexity of data collection and administration. However, when necessary, this is a good investment.

Timing of Data Collection

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

Guidelines for administrators

Specific "DOs" and "DON'Ts" for questioonaire administration are listed in Table 4.1.

Introducing the SF-36 Health Survey

Give target respondents a questionnaire when they check in for their appointment. The following script (or a variation appropriately reworded to sound more like your style of speech) is suggested for introducing the SF-36:

> *We would like to better understand how you and other persons in this study 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 and other persons, 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.*
>
> *You should answer these questions by yourself. Spouses, or other family members, or visitors, should not assist you in completing the questionnaire.*
>
> *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 you have completed it.*

Administering and Completing the SF-36

Provide a firm writing surface such as a clipboard or table top. Provide a number 2 pencil if the respondent is completing the SF-36 for computerized scoring.

Closing

When the respondent returns the SF-36, check the questionnaire for completeness. Note whether the questionnaire is complete. If it is not complete, ask the respondent whether he/she had any difficulty completing it and record the reasons for non-completion.

TABLE 4.1 QUESTIONNAIRE ADMINISTRATION DOS AND DON'TS

DOs	DON'Ts
Do have the respondents fill out the questionnaire before they fill out any other health data forms and before they see their physicians	Do not discuss respondents' health, health data, or emotions with them before they fill out the questionnaire
Do be warm, friendly, and helpful	Do not force or command respondents to fill out the questionnaire
Do request and encourage respondents to fill out the questionnaire	Do not accept an incomplete questionnaire without first encouraging the respondent to fill out unanswered questions
Do read and repeat a question verbatim for the respondent	Do not interpret or explain a question
Do tell respondents to answer a question based on what they think the question means	Do not force or command respondents to fill out a particular question
Do have respondents fill out the questionnaire by themselves	Do not allow spouses or family members to help the respondent fill out the questionnaire
Do encourage respondents to fill out all questions	Do not minimize the importance of the questionnaire
Do thank respondents for filling out the questionnaire	
Do inform respondents if they will be asked to fill out the same questionnaire again at other clinic visits	

Finally, thank the respondent using the following exit script (or a variation appropriately reworded to sound more like your 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 SF-36 may not be appropriate.

Be sure to put the completed questionnaire in a safe and secure place to ensure confidentiality.

Addressing Problems and Questions

What should I do if the respondent refuses to fill out the SF-36?

Respondents are not required to fill out the questionnaire. If the respondent is able to self-administer the SF-36 but *refuses* to participate, tell the respondent that completion of the questionnaire is voluntary, but that it will provide helpful health-related information. In clinical settings, this will help their physician better understand their health problems.

Emphasize that this data is as important as any of the other medical information. The questionnaire responses are essential in order to get a complete picture of the respondent's health. Emphasize that the questionnaire is simple to fill out. Suggest that it is possible that this questionnaire is different from others they have filled out in the past, and that they may even enjoy filling out this questionnaire. If respondent still refuses, take back the questionnaire, record the reason for refusal, and thank the respondent.

What if a respondent does not complete the SF-36?

If noncompletion is a result of the respondent having trouble understanding particular items, ask the respondent to explain why they had difficulty responding to those items. Reread the question for them verbatim, but do not rephrase the question. 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 survey due to difficulty understanding questions. If the respondent is unable to self-administer the questionnaire, 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 questionnaire, some respondents might ask for clarification of *specific items* so that they can better understand and respond to a question. If this happens, the staff member can assist the respondent by rereading the question for them verbatim. If the respondent asks what something means, do not try to explain what the question means, but suggest that the respondent use his or her own interpretation of the question. All respondents should answer the questions based on what they think the questions mean.

Sometimes respondents may have trouble with the response choices. They may say "I don't know" or something different than what is stated on the questionnaire. In these circumstances it is important to gently guide the respondent to respond in one of the pre-set categories 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 questions ask about their health in general.

If the respondent does not like a question, or thinks it is unnecessary or inappropriate, emphasize that all questions are in the survey for a reason that is very important to the study. They should try to answer all of the questions.

Differences in answers due to different wordings of the questionnaire can bias results. It is important to minimize these differences. If the respondent has repeated difficulties filling out the questionnaire which you cannot address with the above direction, thank the respondent, take back the questionnaire, and record the difficulty.

What should I do if a respondent wants to know what his/her answers mean?

If a respondent asks for interpretation of their responses or asks for their score on the questionnaire, tell respondents that you are not trained in how to score or interpret the questionnaire. Emphasize that their answers are to be kept confidential.

What should I do if a respondent is concerned someone will see their answers?
Emphasize that all respondents' responses to the SF-36 are to be kept confidential. You are not allowed to read the responses other than to check that all responses are answered. If an ID number is used to identify a respondent, point out that their names do not appear anywhere on the questionnaire, so that their results will be linked with an ID number and not their name. If this is for a clinical study, tell respondents that their answers to the questionnaire will be pooled with other respondents' answers and that they will be analyzed as a group rather than as individuals for the study.

What should I do if a respondent asks why the SF-36 must be filled out more than once?
Explain that respondents must fill out the same questionnaire at additional visits in order to see if their answers change over time. This will give a more complete picture of the respondent's health over the course of time.

5. TESTS OF SCALING ASSUMPTIONS

Psychometrics is the science of using standardized tests or scales to measure attributes of an individual or an object, in this case health status attributes. Psychometric theory and methods are used to translate a person's ratings and reports into scales to *measure* health attributes.

The SF-36 was constructed to measure eight health attributes using eight multi-item scales containing 2 to 10 items each (see Table 5.1). These scales are scored using Likert's (1932) method of summated ratings. This method is based on certain scaling assumptions that can be tested. For example, the method assumes that the distribution of responses to items within the same scale and item variances are roughly equal. The method also assumes that each item has a substantial *linear* relationship with the score for its scale (referred to as item internal consistency).

The use of each item to score only one scale assumes substantial item discriminant validity (i.e., that each item clearly measures one health concept more than other health concepts). Empirical methods are used to test these scaling assumptions. When the assumptions are well satisfied, items in the same scale can be scored without standardization and can be simply summed with good results.

Defining Scaling Successes

Item internal consistency is evaluated by inspecting the correlation between each item and its hypothesized scale. When scales have a relatively small number of items, as the SF-36 scales do, this test is very important. Therefore, a high standard of internal consistency (correlations above 0.40) has been adopted. With much longer scales this criterion can be relaxed, and minimum standards of score reliability can still be achieved.

With very short scales, it is also necessary to "correct" estimates of internal consistency for the "overlap" between each item and its scale score. Overlap refers to the fact that the item score is also in the scale score. This correction or adjustment amounts to estimating the correlation between the item and the sum of all other items in the same scale. The statistical method used

TABLE 5.1 ABBREVIATED CONTENT FOR ITEMS IN EACH SF-36 SCALE

Scale	Item[a]	Abbreviated Item Content
Physical Functioning (PF)	3a	Vigorous activities, such as running, lifting heavy objects, strenuous sports
	3b	Moderate activities, such as moving a table, vacuuming, bowling
	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 blocks
	3i	Walking one block
	3j	Bathing or dressing
Role–Physical (RP)	4a	Limited in the kind of work or other activities
	4b	Cut down the amount of time spent on work or other activities
	4c	Accomplished less than would like
	4d	Difficulty performing the work or other activities
Bodily Pain (BP)	7	Intensity of bodily pain
	8	Extent pain interfered with normal work
General Health (GH)	1	In your health: excellent, very good, good, fair, poor
	11a	My health is excellent
	11b	I am as healthy as anybody I know
	11c	I seem to get sick a little easier than other people
	11d	I expect my health to get worse
Vitality (VT)	9a	Feel full of pep
	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 would like
	5c	Didn't do work or other activities as carefully as usual
Mental Health (MH)	9b	Been a very nervous person
	9c	Felt so down in the dumps nothing could cheer you up
	9d	Felt calm and peaceful
	9f	Felt downhearted and blue
	9h	Been a happy person
Reported Health Transition (HT)	2	Rating of health now compared to one year ago

[a] Item numbers correspond to Standard form in Appendix B.

in making these "corrections for overlap" is that recommended by Howard and Forehand (1962).

Tests of item discriminant validity focus on the integrity of hypothesized item groupings relative to the health concepts hypothesized. When the correlation between an SF-36 item and its hypothesized scale (concept) is significantly higher than correlations with other SF-36 scales, its inclusion in that hypothesized item grouping is supported.

To evaluate item internal consistency and item discriminant validity, multitrait scaling techniques were employed. Success rates for each of these two scaling tests were evaluated for each scale. For the test of internal consistency, a scaling success was counted whenever the correlation between an item and its hypothesized scale (corrected for overlap) equalled or exceeded 0.40. The overall success rate for a given scale is equal to the number of scaling successes divided by the total number of scaling tests. For example, for the PF scale, 10 tests were performed (one for each item).

For item discriminant validity tests, a success was counted whenever an item correlated significantly higher (two standard errors or more) with its hypothesized scale compared to another SF-36 scale. The item discriminant validity success rate 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 (10 items and eight tests each). The software used in these corrections and other analyses reported below is available from The Health Institute at New England Medical Center.

Results of Scaling Tests

Table 5.2 presents item means, standard deviations, and correlations between items and SF-36 scales in the Medical Outcomes Study (MOS). These correlations are used to evaluate item internal consistency and item discriminant validity. Correlations that are hypothesized to be larger and that have been corrected for overlap have an asterisk (*) by them in Table 5.2.

The first important finding from Table 5.2 is that, within each scale, correlations between items and their hypothesized scale were roughly equal and with only one exception exceeded the 0.40 standard for internal consistency (see GH scale). Thus, as summarized in Table 5.3, a near perfect success rate

TABLE 5.2 ITEM MEANS, STANDARD DEVIATIONS, AND CORRELATIONS[a] WITH SF-36 SCALES; RESULTS FROM THE MOS (N=3,445)

Item[b]	MEAN	SD	PF	RP	BP	GH	VT	SF	RE	MH
3a	1.82	.79	.62*	.51	.42	.46	.39	.24	.16	.10
3b	2.48	.71	.78*	.54	.51	.45	.45	.40	.23	.16
3c	2.56	.66	.76*	.52	.50	.41	.40	.40	.22	.14
3d	2.25	.77	.78*	.50	.44	.46	.43	.33	.21	.13
3e	2.66	.60	.78*	.45	.41	.39	.38	.35	.20	.12
3f	2.44	.70	.71*	.50	.50	.40	.38	.32	.19	.12
3g	2.28	.81	.78*	.51	.46	.46	.43	.34	.20	.13
3h	2.52	.74	.80*	.48	.45	.43	.42	.37	.20	.14
3i	2.76	.55	.72*	.42	.42	.36	.35	.37	.18	.14
3j	2.88	.39	.49*	.31	.34	.27	.27	.35	.18	.14
4a	1.63	.48	.58	.67*	.55	.42	.43	.41	.28	.20
4b	1.62	.49	.51	.70*	.51	.42	.47	.46	.40	.27
4c	1.46	.50	.44	.65*	.46	.39	.51	.39	.41	.28
4d	1.57	.50	.49	.68*	.56	.41	.50	.45	.39	.32
7	4.31	1.30	.53	.54	.70*	.44	.46	.43	.27	.29
8	4.38	1.40	.53	.61	.70*	.44	.48	.51	.33	.33
1	3.15	.92	.58	.49	.48	.63*	.52	.40	.28	.28
11a	3.03	1.26	.49	.45	.44	.72*	.52	.40	.28	.32
11b	3.12	1.24	.42	.37	.35	.65*	.44	.34	.24	.27
11c	4.01	1.14	.22	.29	.30	.44*	.38	.36	.28	.35
11d	3.43	1.14	.26	.24	.20	.38*	.25	.16	.15	.19
9a	3.11	1.43	.45	.51	.42	.51	.75*	.46	.40	.44
9e	3.43	1.35	.43	.49	.40	.48	.74*	.43	.37	.42
9g	4.16	1.25	.42	.50	.48	.49	.69*	.54	.45	.55
9i	3.99	1.22	.40	.46	.44	.47	.70*	.50	.43	.50
6	4.18	1.05	.43	.51	.51	.42	.54	.74*	.52	.58
10	4.27	1.03	.38	.46	.45	.42	.52	.74*	.51	.62
5a	1.73	.45	.20	.39	.27	.29	.41	.50	.70*	.54
5b	1.60	.49	.22	.41	.28	.30	.45	.48	.73*	.56
5c	1.74	.44	.23	.36	.28	.29	.40	.45	.63*	.50
9b	4.67	1.31	.18	.26	.29	.33	.43	.47	.47	.65*
9c	5.34	1.06	.13	.27	.27	.28	.42	.58	.53	.73*
9d	3.96	1.39	.11	.28	.30	.32	.50	.53	.53	.79*
9f	4.77	1.20	.12	.27	.27	.32	.49	.59	.57	.81*
9h	4.14	1.30	.15	.30	.30	.37	.53	.58	.54	.81*
2	3.37	.92	.18	.17	.22	.22	.23	.24	.14	.19

continued on next page

TABLE 5.2 CONTINUED

a. Item-total correlation corrected for overlap. Standard error = .02.

b. Item numbers correspond to Standard form in Appendix B.

* Indicates highest hypothesized correlation and corrected for overlap.

Notes. A column that presents correlations between each item and the Reported Health Transition (HT) item (Item 2) has been excluded from this table. This column consistently showed low correlations (.09–.25). These correlations were included in the counts of successes shown in Table 5.3.

From "The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups" by C.A. McHorney et al., in press, *Medical Care*.

was observed in tests of the item internal consistency criterion in the MOS sample of patients with one or more chronic conditions. (See McHorney et al., in press, for a description of the sample.)

Tests of item discriminant validity involve a comparison between the correlation for each item and its hypothesized scale versus other correlations in the same row in Table 5.2. For example, in testing discriminant validity for item 3a in the PF scale, the item-scale correlation of 0.62 (corrected for overlap) was compared with other correlations in the same row, such as 0.51 with RP, and 0.42 with BP (see Table 5.2). As summarized in Table 5.3, tests of item discriminant validity were passed by all items in all SF-36 scales in tests performed for the MOS sample.

Conclusions

Results from SF-36 tests of scaling assumptions illustrated here and elsewhere strongly support the use of the method of summated ratings in computing scores for SF-36 scales. Examples of favorable results from similar tests have been reported by other researchers (Brazier et al., 1992; Garratt et al., 1993; Jenkinson et al., 1993). Forthcoming results from a study of 24 MOS subgroups also suggest that tests of scaling assumptions are generalizable (McHorney et al., in press).

Given that item variances tend to be comparable with few exceptions, it is not surprising that further transformation of item variances does not improve the reliability or validity of scale scores. Reliability coefficients for scale scores computed from items standardized to have equal variances did not differ at the second decimal place in the MOS. Further, across all scales and subgroups,

TABLE 5.3 SUMMARY RESULTS OF TESTS OF ITEM INTERNAL CONSISTENCY AND DISCRIMINANT VALIDITY: RESULTS FROM THE MOS (N=3,445)

		Range of Correlations		Internal Consistency Tests[d]		Discriminant Validity Tests[e]	
Scale	k[a]	Item-Internal Consistency[b]	Item-Discriminant Validity[c]	# Success/ Total	Success Rate (%)	# Success /Total	Success Rate (%)
PF	10	.49 - .80	.19 - .54	10/10	100	80/80	100
RP	4	.65 - .70	.12 - .58	4/4	100	32/32	100
BP	2	.70	.19 - .61	2/2	100	16/16	100
GH	5	.38 - .72	.09 - .58	4/5	80	40/40	100
VT	4	.69 - .75	.17 - .55	4/4	100	32/32	100
SF	2	.74	.20 - .62	2/2	100	16/16	100
RE	3	.63 - .73	.11 - .56	3/3	100	24/24	100
MH	5	.65 - .81	.11 - .59	5/5	100	40/40	100

[a] Number of items and number of item-internal consistency tests per scale.

[b] Correlations between items and hypothesized scale corrected for overlap.

[c] Correlations between items and other scales.

[d] Number ≥ 0.40.

[e] Number of correlations significantly higher/total number of correlations.

Note. From "The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups" by C.A. McHorney et al., in press. *Medical Care.*

correlations between scores for scales constructed from items with and without transformation to equate variances have been shown to be nearly perfect in the MOS (McHorney et al., in press). Thus, the requirement of equivalence of item means and variances appears to be well satisfied in studies to date using the scoring methods in Chapter 6. There is no evidence of gains in reliability or validity from further transformation to equate item variances.

Many of the analyses reported in Chapter 9 have implications for the assumptions underlining the construction of SF-36 scales and the methods used in their scoring (see "Criterion-based interpretation").

6. SCORING THE SF-36

This chapter provides scoring instructions for the eight multi-item scales and for the reported health transition item included in the SF-36 Health Survey. Chapter 3 describes the SF-36 scales and items. General scoring information and steps for data entry and scoring that are common to all items are discussed first (see Figure 6.1). Next, formulas for item aggregation and transformation of scale scores are presented. Finally, formal checks for errors in scoring are explained.

Importance of standardization

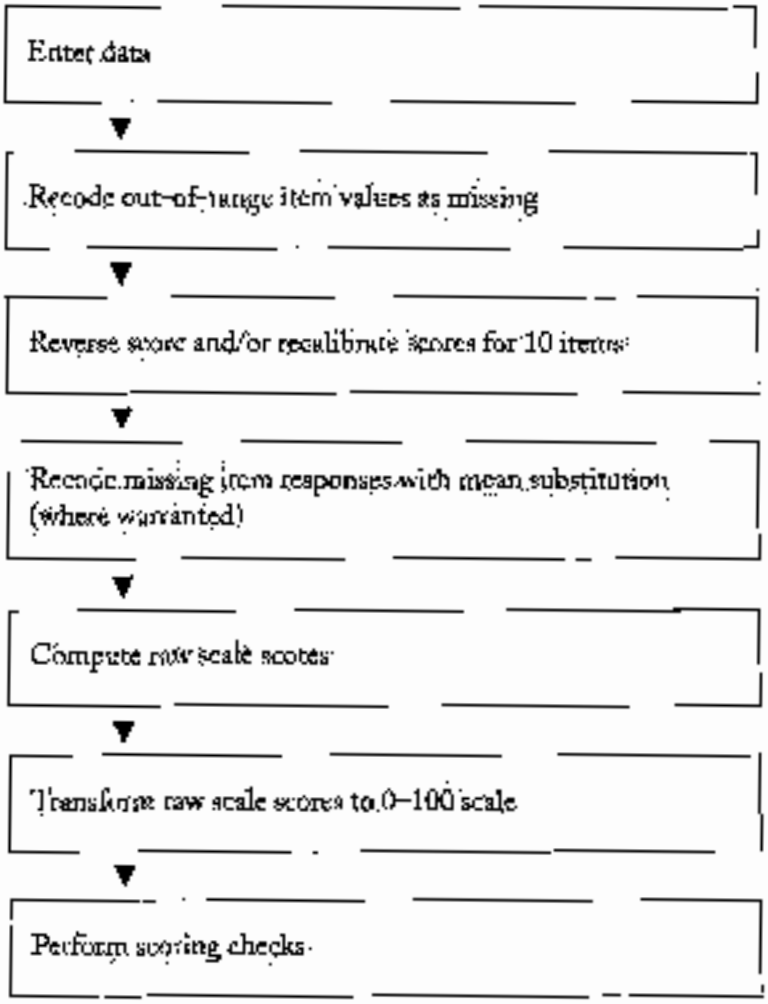
As with all standardized tests, standardization of content and scoring is what makes interpretation of the SF-36 scales possible. The content of the SF-36 form and the scoring algorithms were selected and standardized following careful study of many options. The algorithms described in this chapter were chosen to be as simple as possible while still satisfying the assumptions of the methods used to construct SF-36 scales.

Changes in the content of the survey or in scoring algorithms may compromise the reliability and validity of scores. Changes are also likely to bias scores sufficiently to invalidate normative comparisons and to prevent comparisons of results across studies.

There are at least two good reasons to adhere to the standards of content and scoring described in this manual. First, they are most likely to produce scores with the same reliability and validity as those reported here and in other Medical Outcomes Study (MOS) publications. Second, comparisons of results across studies are made possible to the benefit of all who use these content and scoring standards.

Prior to using the SF-36 scoring rules, it is essential to verify that the questionnaires being scored, including the questions asked (item stems), response choices, and numbers assigned to response choices at the time of data entry, have been reproduced exactly. The scoring rules described in this chapter are

FIGURE 6.1
FLOW CHART FOR SCORING THE SF-36



appropriate for the standard SF-36 survey questions, response choices, and numbers assigned to response choices as reproduced in Appendix B. The chapter ends with algorithms that help to equate scores for the Developmental version and the Standard version of the SF-36.

General scoring information

SF-36 items and scales are scored so that a higher score indicates a better health state. For example, functioning scales are scored so that a high score indicates better functioning and the pain scale is scored so that a high score indicates freedom from pain. After data entry, items and scales are scored in three steps:

(1) item recoding, for the 10 items that require recoding;

(2) computing scale scores by summing across items in the same scale (raw scale scores); and

(3) transforming raw scale scores to a 0 - 100 scale (transformed scale scores).

We recommend that both item recoding and scale scoring be performed by computer, using the scoring algorithms documented here or computer software available elsewhere (THI, 1992).

Data Entry

The SF-36 item responses should be keypunched as coded in the questionnaire. It is important to note that, although the numbers printed along with the response choices should be keypunched, they may not be the numbers ultimately assigned to those responses when SF-36 scales are scored.

In most cases, this means that the precoded number that is circled or marked by the respondent should be entered. However, sometimes it is not clear what number should be entered. Suggested rules for handling some of the more common coding problems are:

- If a respondent marks two responses which are adjacent to each other, randomly pick one and enter that number.

- If a respondent marks two responses for an item and they are not adjacent to each other, code that item "missing."
- If a respondent marks three or more responses for an item, code that item "missing."
- If a respondent answers the "yes/no" items by writing in "yes" or "no," code the answer as though "yes" or "no" had been marked.

Response Technologies Inc. and other companies have developed scanning forms for use with the SF-36, in both standard and acute formats. Sample forms appear in Appendix B. Optical scanning generally reduces the time required to process questionnaires, but may involve greater initial investment in form design. Some scanning forms may require special processing equipment; however, this method may be cost-effective, especially if the SF-36 is being administered frequently or to a large sample (see Chapter 12).

Tables 6.1 through 6.9 present scoring information for the items used in each of the eight SF-36 health scales and the reported health transition item. Each table presents the verbatim content of each question, response choices, and both the precoded values printed in the questionnaire and final values for scoring each item. Item numbers in Tables 6.1 through 6.9 correspond to those on the Standard SF-36 form (reproduced in Appendix B).

Item Recoding

The next stage after data entry is the recoding of response choices as shown in Tables 6.1 through 6.9. Item recoding is the process of deriving the item values that will be used to calculate the scale scores. Several steps are included in this process: (1) change out-of-range values to missing, (2) recode values for 10 items, and (3) substitute person-specific estimates for missing items.

Out-of-Range Values

All 36 items should be checked for out-of-range 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 (see Tables 6.1 through 6.9). Out-of-range values are usually caused by data-entry errors and, if possible, should be changed to the correct response through verification with the original questionnaire. If the questionnaire is not available, all out-of-range values should be recoded as missing data.

TABLE 6.1 PHYSICAL FUNCTIONING: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

3a. Vigorous activities, such as running, lifting heavy objects, 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 blocks

3i. Walking one block

3j. Bathing or dressing yourself

Precoded and Final Values for Items 3a - 3j

Response Choices	Precoded Item Value	Final Item Value
Yes, limited a lot	1	1
Yes, limited a little	2	2
No, not limited at all	3	3

Scale Scoring

Compute the simple algebraic sum of the final item scores as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better physical functioning.

Note. Precoded values are as shown on the appended form. This scale does not require recoding of items prior to computation of the scale score.

TABLE 6.2 ROLE-PHYSICAL: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

4a. Cut down the amount of time you spent on work or other activities

4b. Accomplished less than you would like

4c. Were limited in the kind of work or other activities

4d. Had difficulty performing the work or other activities (for example, it took extra effort)

Precoded and Final Values for Items 4a – 4d

Response Choices	Precoded Item Value	Final Item Value
Yes	1	1
No	2	2

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better Role-Physical functioning.

Note. Precoded values are as shown on the expanded form. This scale does not require recoding of items prior to computation of the scale score.

TABLE 6.3 BODILY PAIN: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

7 How much **bodily** pain have you had during the **past 4 weeks**?

8. During the **past 4 weeks**, how much did pain interfere with your normal work (including both work outside the home and housework)?

Precoded and Final Values for Item 7

Response Choices	Precoded Item Value	Final Item Value
None	1	6.0
Very mild	2	5.4
Mild	3	4.2
Moderate	4	3.1
Severe	5	2.2
Very severe	6	1.0

Scoring for Item 8 — if both Items 7 and 8 are answered

Response Choices	If Item 8 Precoded Item Value	and	Item 7 Precoded Item Value	then	Item 8 Final Item Value
Not at all	1		1		6
Not at all	1		2 through 6		5
A little bit	2		1 through 6		4
Moderately	3		1 through 6		3
Quite a bit	4		1 through 6		2
Extremely	5		1 through 6		1

Scoring for Item 8 — if Item 7 is not answered

Response Choices	Precoded Item Value	Final Item Value
Not at all	1	6.0
A little bit	2	4.75
Moderately	3	3.5
Quite a bit	4	2.25
Extremely	5	1.0

Scale Scoring

Compute the simple algebraic sum of final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored positively so that a high score indicates lack of bodily pain.

Note. Precoded values are as shown on the appended form. This scale requires recoding of both items prior to computation of the scale score.

TABLE 6.4 GENERAL HEALTH: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

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

11a. I seem to get sick a little easier than other people

11b. I am as healthy as anybody I know

11c. I expect my health to get worse

11d. My health is excellent

Precoded and Final Values for Items 1 & 11a–11d

Item 1	Response Choices	Precoded Item Value	Final Item Value
	Excellent	1	5.0
	Very good	2	4.4
	Good	3	3.4
	Fair	4	2.0
	Poor	5	1.0

Items 11a & 11c	Response Choices	Precoded Item Value	Final Item Value
	Definitely True	1	1
	Mostly True	2	2
	Don't Know	3	3
	Mostly False	4	4
	Definitely False	5	5

Items 11b & 11d	Response Choices	Precoded Item Value	Final Item Value
	Definitely True	1	5
	Mostly True	2	4
	Don't Know	3	3
	Mostly False	4	2
	Definitely False	5	1

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better general health perceptions.

Note. Precoded values are as shown on the appended form. This scale requires recoding of three items prior to computation of the scale score.

TABLE 6.5 VITALITY: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

9a. Did you feel full of pep?

9e. Did you have a lot of energy?

9g. Did you feel worn out?

9i. Did you feel tired?

Precoded and Final Values for Items 9a, 9e, 9g, & 9i

Items 9a & 9e	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	6
	Most of the time	2	5
	A good bit of the time	3	4
	Some of the time	4	3
	A little of the time	5	2
	None of the time	6	1

Items 9g & 9i	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	1
	Most of the time	2	2
	A good bit of the time	3	3
	Some of the time	4	4
	A little of the time	5	5
	None of the time	6	6

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates more vitality.

Note. Precoded values are as shown on the appended form. This scale requires recoding of two items prior to computation of the scale score.

TABLE 6.6 SOCIAL FUNCTIONING: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

Precoded and Final Values for Items 6 & 10

Item 6	Response Choices	Precoded Item Value	Final Item Value
	Not at all	1	5
	Slightly	2	4
	Moderately	3	3
	Quite a bit	4	2
	Extremely	5	1

Item 10	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	1
	Most of the time	2	2
	Some of the time	3	3
	A little of the time	4	4
	None of the time	5	5

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better social functioning.

Note Precoded values are as shown on the appended form. This scale requires recoding of one item prior to computation of the scale score.

TABLE 6.7 ROLE–EMOTIONAL: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

5a. Cut down the amount of time you spent on work or other activities

5b. Accomplished less than you would like

5c. Didn't do work or other activities as carefully as usual

Precoded and Final Values for Items 5a – 5c

Response Choices	Precoded Item Value	Final Item Value
Yes	1	1
No	2	2

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. See text for handling of missing item responses. This scale is scored so that a high score indicates better Role–Emotional functioning.

Note. Precoded values are as shown on the appended form. This scale does not require recoding of items prior to computation of the scale score.

TABLE 6.3 MENTAL HEALTH: VERBATIM ITEMS AND SCORING INFORMATION

Verbatim Items

9b. Have you been a very nervous person?

9c. Have you felt so down in the dumps that nothing could cheer you up?

9d. Have you felt calm and peaceful?

9f. Have you felt downhearted and blue?

9h. Have you been a happy person?

Recoded and Final Values for Items 9b, 9c, 9d, 9f, & 9h

Items 9b, 9c, & 9f	Response Choices	Recoded Item Value	Final Item Value
	All of the time	1	1
	Most of the time	2	2
	A good bit of the time	3	3
	Some of the time	4	4
	A little of the time	5	5
	None of the time	6	6

Items 9d & 9h	Response Choices	Recoded Item Value	Final Item Value
	All of the time	1	6
	Most of the time	2	5
	A good bit of the time	3	4
	Some of the time	4	3
	A little of the time	5	2
	None of the time	6	1

Scale Scoring

Compute the simple algebraic sum of the final item values as shown in Table 6.11. Set the raw for handling of missing item responses. This scale is scored so that a high score indicates better mental health.

Note. Recoded values are as shown on the recoded form. This scale requires recoding of two items prior to computation of the scale score.

TABLE 6.9 REPORTED HEALTH TRANSITION: VERBATIM ITEM AND SCORING INFORMATION

Verbatim Item

2. **Compared to one year ago**, how would you rate your health in general now?

Precoded and Final Values for Item 2

Response Choices	Precoded Item Value
Much better now than one year ago	1
Somewhat better now than one year ago	2
About the same as one year ago	3
Somewhat worse now than one year ago	4
Much worse now than one year ago	5

Note. Precoded item values are as shown on the appended form. The average measured change in health for respondents selecting each response choice is presented in Chapter 9.

Recode Values for 10 Items

Seven items are reverse scored. Reverse scoring of items is done to ensure that a higher item value indicates better health on all SF-36 items and scales. SF-36 items that need to be reverse scored are worded so that a higher precoded item value indicates a poorer health state.

Item Recalibration

For 34 of the SF-36 items, research to date offers good support for the assumption of a linear relationship between item scores and the underlying health concept defined by their scales. However, empirical work has shown that two items require recalibration to satisfy this important scaling assumption. These items are in two different SF-36 scales: the General Health (GH) scale and the Bodily Pain (BP) scale.

General Health Rating Item. The "Very Good" and "Good" responses to Item 1 are recalibrated to achieve a better linear fit with the general health evaluation concept measured by the GH scale. Empirical studies during the Health Insurance Experiment (HIE) were among the first to document that the intervals between response choices for this item are not equal (Davies & Ware, 1981). Subsequent studies of Item 1, using both the Thurstone Method of Equal-Appearing Intervals (Thurstone & Chave, 1929) and other empirical methods, have also consistently shown that the interval between "Excellent" and "Very Good" is about half the size of the interval between "Fair" and "Good" (Ware, Nelson et al., 1992). These results have been confirmed in studies of SF-36 translations from 10 countries participating in the International Quality of Life Assessment (IQOLA) Project. Finally, in all studies we are aware of to date, mean values for a criterion general health scale for respondents who choose each of the five levels defined by Item 1 depart significantly from linearity.

Results from two MOS studies that served as the basis for the recommended recalibration of Item 1 are summarized in Table 6.10. As shown in Table 6.10 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. Intervals between adjacent response categories were unequal, as observed in the HIE (Davies & Ware, 1981). For these reasons, item scale values are transformed as shown in Table 6.10 using specific results from the

TABLE 6.10 MEAN CURRENT HEALTH SCORES FOR RESPONDENTS CHOOSING EACH LEVEL OF SF-36 ITEM 1

	Mean Current Health		Recommended Scoring	
Response to Item 1	Screening Sample (N=18,573)	Baseline Sample (N=3,054)	1–5 Scale	0–100 Scale
Excellent	87.9	86.9	5.0	100
Very good	75.5	75.4	4.4	84
Good	57.6	55.9	3.4	61
Fair	30.0	30.6	2.0	25
Poor	10.8	10.8	1.0	0

Note. Adapted from "Preliminary results of a 6-item general health survey: A patient application" by J.E. Ware, E.C. Nelson et al., 1992, in A.L. Stewart & J.E. Ware (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (p. 299). Durham, NC: Duke University Press.

screening sample. The result is a very high 0.70 correlation with the sum of the other four items in the GH scale.

BodBy Pain Items. The scoring rules recommended for the Bodily Pain (BP) scale were based on three considerations: (1) the items offer both different numbers and different content of response choices, (2) administration of Item 8 depended on the response to an item like Item 7 in the MOS, and (3) empirical studies indicate that recalibration of Item 7 is necessary to achieve a linear fit with the scale score and with other measures of bodily pain.

As shown in Table 6.3, the two bodily pain items offer an unequal number 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 has been dropped to make the SF-36 easier to administer, the dependence between responses must be taken into account to compare results from new studies with published studies.

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 done by splitting those free of role interference due to pain into two different groups: (1) free of interference and free of pain (the best level), and (2) free of interference but with at least some pain (the next best level). Second, it approximates the dependence between the two items in MOS studies of reliability and validity to date (McHorney et al., 1992, 1993, in press).

Davies and Ware (1981) reported that recalibration of the bodily pain severity rating was necessary to satisfy the equal interval assumption in studies during the HIE. MOS studies have confirmed that the relationship between Item 7 and criterion measures of pain departs significantly 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 Item 7, using methods much like those illustrated in Table 6:10 for Item 1.

How to Treat Missing Data

Sometimes respondents leave one or more questionnaire items in a scale blank, although this happens infrequently (1 to 2% or less) in most surveys. One important advantage of multi-item scales is that a scale score can be estimated even though responses to some items are missing. Using a scoring algorithm that estimates missing values, it is usually possible to derive scale scores for nearly all respondents across the eight SF-36 scales.

We recommend that a scale score be calculated if a respondent answered at least half of the items in a multi-item scale (or half plus one in the case of scales with an odd number of items).

The recommended algorithm substitutes a person-specific estimate for any missing item when the respondent answered at least 50 percent of the items in a scale. A psychometrically sound estimate is the average score, across completed items in the same scale, for that respondent (Ware, Davies-Avery, & Brook, 1980). For example, if a respondent leaves one item in the 5-item Mental Health scale blank, substitute the respondent's average score (across the four completed mental health items) for that one item. When estimating the respondent's average score, use the respondent's final item values, as

defined in Tables 6.1 through 6.9. This step is easy to program using standard software packages (e.g., SPSS, SAS). Examples of program code and scoring software are available elsewhere (TH1, 1992).

Computing Raw Scale Scores

After item recoding, including handling of missing data, a raw score is computed for each scale. This score is the simple algebraic sum of responses for all items in that scale, as shown in Table 6.11. For example, the raw scale score for the Role–Physical scale is the sum of the scores for Items 4a, 4b, 4c, and 4d. Use recoded items values and imputed values where applicable. Generally, we recommend that if the respondent answers at least 50% of the items in a multi-items scale, the score should be calculated. If the respondent did not answer at least 50% of the items, the score for that scale should be set to missing. Some prefer a more conservative approach for the scales with only two items and set those scales to missing unless both items are completed.

This simple scoring method is possible because items in the same scale have roughly equivalent relationships to the underlying health concept being measured, and no item is used in more than one scale. Thus, it is not necessary to standardize or weight items. These assumptions have been extensively tested and verified across 24 patient groups (McHorney et al., in press).

Transformation of Scale Scores

The next step involves transforming each raw scale score to a 0 to 100 scale using the formula shown below. Table 6.11 provides the information necessary to apply this formula to each scale.

$$\text{Transformed Scale} = \left[\frac{(\text{Actual raw score} - \text{lowest possible raw score})}{\text{Possible raw score range}} \right] \times 100$$

This transformation converts the lowest and highest possible scores to zero and 100, respectively. Scores between these values represent the percentage of the total possible score achieved. While this final step is optional, it is strongly recommended because transformed scale scores can be compared with norms derived from the MOS (McHorney et al., 1992, 1993, in press).

TABLE 6.11 FORMULAS FOR SCORING AND TRANSFORMING SCALES

Scale	Sum Final Item Values *(after recoding items as in Tables 6.1–6.8)*	Lowest and highest possible raw scores	Possible raw score range
Physical Functioning	3a+3b+3c+3d+3e+3f+3g+3h+3i+3j	10, 30	20
Role–Physical	4a+4b+4c+4d	4, 8	4
Bodily Pain	7+8	2, 12	10
General Health	1+11a+11b+11c+11d	5, 25	20
Vitality	9a+9e+9g+9i	4, 24	20
Social Functioning	6+10	2, 10	8
Role–Emotional	5a+5b+5c	3, 6	3
Mental Health	9b+9c+9d+9f+9h	5, 30	25

Formula and example for transformation of raw scale scores

$$\text{Transformed Scale} = \left[\frac{(\text{Actual raw score} - \text{lowest possible raw score})}{\text{Possible raw score range}} \right] \times 100$$

Example: A Physical Functioning raw score of 21 would be transformed as follows:

$$\left[\frac{(21 - 10)}{20} \right] \times 100 = 55$$

Where lowest possible score = 10 and possible raw score range = 20

1990 National Survey of Functional Health Status, and other published and forthcoming results based on these scoring rules.

Raw and transformed scale scores are not calculated for the Reported Health Transition item. We recommend treating responses to this item as ordinal level data and analyzing the percentage of respondents who select each response choice or using the estimates of measured change reported for each response category in Chapter 9.

Scoring Checks

Because errors can occur while reproducing a form, entering data, programming or processing, which could lead to inaccurate scale scores, we strongly recommend formal scoring checks prior to using the scales. Any discrepancies observed during the following checks should be investigated for scoring errors:

(1) Calculate SF-36 scale scores by hand for several respondents and compare the results to those produced by your scale-scoring computer software.

(2) After items have been coded into their final item values, inspect the frequency distributions for the items to verify that only the final item values shown in Tables 6.1 through 6.9 are observed. Discrepancies should be limited to respondents with values estimated for missing data.

(3) After items have been recoded and scale scores have been computed, inspect the correlation between each scale and its component items to verify that all correlations are positive in direction and substantial in magnitude (0.30 or higher).

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

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

Scoring of the SF-36 Developmental version

Some studies have been based on the Developmental version of the SF-36 made available in December 1988 (Ware, 1988). This section explains how to score the Developmental version to be more comparable with the Standard version. Thirty-five items across seven scales in the Developmental version can be scored identically to items in the Standard version.

One Social Functioning (SF) item differs in both item content and response choice format in the two versions (Item 9j in the Developmental version and Item 10 in the Standard version). The item content in the Standard version asks specifically if "physical health or emotional problems interfered with your social activities," while the Developmental version asks if "health limited your social activities." Further, only five response choices are provided in the Standard version to equate the variance of the two SF items without recalibration; thus, its scoring is simpler. The Developmental version had six response choices for this item.

To make scores for the SF scale in the Developmental version (Items 6 and 9j) more comparable to the Standard version: (1) reverse the scoring of the first item (Item 6); (2) recalibrate the second item (Item 9j) so it ranges from "1" to "5" rather than from "1" to "6"; and (3) compute the scale by summing the two items. Table 6.12 details these scoring steps.

Scoring alternatives

Scoring algorithms made available to users of the SF-36 Developmental version in 1988 are identical to those for the SF-36 Standard version for six of the eight scales (Ware, 1988). Both Developmental and Standard scoring algorithms include the recalibration of Item 1 of the GH scale, as documented and explained earlier in this chapter. We are not aware of any published studies in the United States, United Kingdom, or elsewhere that do not use the SF-36 scoring algorithm for the GH scale.

Those using the Developmental version of the SF-36 have a choice between Developmental (old) and Standard (new) scoring algorithms for the SF scale. The Developmental version of the second SF item (Item 9j) offered six

TABLE 6.12 SCORING THE SF-36 DEVELOPMENTAL VERSION SOCIAL FUNCTIONING SCALE

Verbatim Items

6. During the **past 4 weeks**, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

9j. How much of the time during the past month has your health limited your social activities like visiting with friends or close relatives?

Precoded and Final Values for Items 6 & 9j

Item 6	Response Choices	Precoded Item Value	Final Item Value
	Not at all	1	5
	Slightly	2	4
	Moderately	3	3
	Quite a bit	4	2
	Extremely	5	1

Item 9j	Response Choices	Precoded Item Value	Final Item Value
	All of the time	1	1.0
	Most of the time	2	1.8
	A good bit of the time	3	2.6
	Some of the time	4	3.4
	A little of the time	5	4.2
	None of the time	6	5.0

Scale Scoring

Compute the simple algebraic sum of the final item values by summing Items 6 and 9j as described in the text. See text for handling of missing item responses. This scale is scored so that a high score indicates better social functioning.

Note. Precoded values are as shown on the appended form. This scale requires recoding of two items prior to computation of the scale score.

response choices, while the Standard version of this item (Item 10) has five response choices. The recalibration of this item, as recommended above (see "Scoring of the SF-36 Developmental version"), has no effect on the interpretation of SF scale scores. Instead, it adjusts the scale mean to be comparable to those in the normative data presented in Chapter 10.

Finally, we have published two options for scoring the Bodily Pain (BP) scale, which has identical item content across the Developmental and Standard versions. Advances in scoring for the Standard SF-36 are listed and explained above (see "Item Recalibration"). Early users of the Developmental version used the older (Developmental) scoring method (Ware, 1988). Thus, BP scales scored the old way will have means that are two to four points higher, on average, than BP scales scored the new (Standard) way. Scores for more than a third of respondents in the general population are shifted upward (i.e., towards better health) by five points or more when using the old scoring relative to the Standard scoring. The extent of this shift and implications for the precision and interpretation of BP scale scores, scored the old way, may vary depending on the proportion of respondents with chronic conditions and on their specific diagnoses. Therefore, we recommend routine use of the Standard scoring algorithms (as presented in this chapter). We encourage thorough documentation of any departures from this scoring system so that readers will know when they can and cannot compare results with other published studies.

Scoring Advances

We are presently evaluating several potential improvements in the scoring of the SF-36 including: (1) improvements in the enumeration of scale levels, (2) construction of aggregate (summary) indexes, and (3) norm-based scoring of scales and summary indexes. These and other SF-36 scoring issues that are likely to influence progress in the health assessment field are discussed in Chapter 12.

7. RELIABILITY, PRECISION, AND DATA QUALITY

Reliability

This chapter summarizes methods that are used to estimate scale score reliability and explains how reliability coefficients are interpreted. It also summarizes reliability estimates for SF-36 scales from the Medical Outcomes Study (MOS), for the general U.S. population, and for U.S. and U.K. studies published by others. Related issues, including the sample sizes required to detect differences in mean scale scores between groups and the sizes of confidence intervals around scores for individual patients, are also discussed. Finally, the quality of SF-36 data is discussed in terms of data completeness and the Response Consistency Index (RCI), which can be used to judge data quality for an individual patient. Norms for the RCI and for missing response rates are presented.

Background

To be useful, the data collection techniques and the rules for converting responses to numbers must produce information that is not only relevant but correct. Two crucial aspects of correctness are reliability, the extent to which measures give consistent or accurate results, and validity, the extent to which the results pertain directly to the desired attribute or characteristic being measured. A measurement procedure is reliable to the extent that a series of measurements give the same results. A ruler is normally a highly reliable measurement tool; repeated measurements of the same object with a well calibrated ruler would yield more reliable results than with a poorly calibrated ruler. However, consistent results may not be valid results. For example, at one time it was believed that the circumference of the head was directly related to the degree of intelligence. A tape measure proved to be a reliable tool for measuring the circumference of the head. However, the resulting score proved to be not valid in relation to intelligence.

Health measures also should be sensitive. Instruments differ in characteristics that determine their sensitivity: a crude measure may offer only two alternative answers, while a better graduated measure would offer a number of choices indicating amounts of the attribute. An example would be a question concerning feelings of depression. One could offer alternative answers of "Yes" or "No" as an example of a coarse measure. A question offering six

choices of answers, each defining a different level of severity, would provide more information about depression. The second graduated question is more sensitive to variations in feelings of depression among those who would have chosen "Yes" in response to the "Yes" or "No" item.

Interpreting Reliability Coefficients

The evaluation of the reliability of any measurement procedure consists of estimating how much of the variation in a score is real or truth as opposed to chance or random errors (Selltiz et al., 1976). A reliability of 0.70 indicates that 70% of the measured variance is reliable; reliability coefficients are therefore proportions. Reliability examines the consistency of results from different measures designed to evaluate the same variable. The exact questionnaires may be repeated, or alternative forms of the tool may be administered. Acceptable reliability differs depending on what is being analyzed: comparisons among individuals or across administrations to the same individual require high reliability (values > 0.90); group comparisons, needed to compare average health status scores between diagnostic or treatment groups, do not require as high a reliability (values of 0.50 or 0.70 or higher are acceptable) (Helmstadter, 1964; Nunnally, 1978).

Trends toward higher reliability coefficients for many of the newer health status surveys reflect both conceptual and methodologic advances. First, the amount of information gained from each questionnaire item has increased; 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, items in the same scale in the newer instruments define more homogeneous constructs and therefore yield more reliable scores. Homogeneity coefficients indicate the internal consistency of a measure, regardless of the number of items (Tyler & Fiske, 1968). Failure to consider homogeneity can be problematic. 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); they were quite heterogeneous (low in internal consistency) and yielded a relatively low reliability coefficient (Ware et al., 1979). This early health measure also had a complicated interpretation.

Reliability can be estimated by examining (1) the stability of scores from one

administration time to another or (2) the equivalence of individual answers across alternative forms of an instrument.

Stability is often determined by repeated administrations of a test, referred to as the test-retest procedure. This method is useful for estimating the reliability of single-item measures; however, recall of responses across administrations and changes in the attribute between tests introduce subject bias. Additionally, the dynamic nature of health often means that an individual's responses will vary as a result of genuine change. Therefore, the time interval between tests must be considered in studies of reliability that use the test-retest method. Researchers using this method are more likely to underestimate reliability rather than overestimate it.

Another approach is to administer alternate forms to the same individuals at the same time. In this procedure, alternate forms proven to be equivalent are given to the same individuals in a single testing session. The correlation between scores for alternate forms indicates the reliability of the measures. Use of this method has been rare even for the most widely used health measures because of the difficulty of constructing a truly alternate form (McHorney & Ware, forthcoming).

One of the first approaches to estimating reliability was the split-half correlation. At its simplest, that approach compares the two halves of a test, for example, the scores from the odd-numbered and even-numbered items. If developed to measure a single characteristic, both halves of the test should provide fairly equivalent results. A number of analogous techniques have been developed to improve this approach, most notably coefficient alpha, which is the average of all possible split-half reliabilities adjusted to the original number of items (Cronbach & Warrington, 1951).

Reliability of a scale score depends on the number of items in the scale and on the homogeneity of the items. As the demand for health status measurement tools has grown, different approaches to the trade-off between the length of a measure and its reliability and validity have been explored. Medical practitioners and clinical research scientists seem to find lengthy scales less practical for widespread use.

Single-item measures have been documented to correlate with long-form (parent) measures (Coates et al., 1987; Meyerboom-DeJong & Smith, 1990;

Nelson et al., 1987, 1990a, 1990b; Stewart & Ware, 1992). Single-item measures are satisfactory for use in detecting moderate to large differences between groups of 150 or more patients. From a practical point of view, the single-item measure may be attractive but probably will not detect meaningful differences at the individual patient level; 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 measures and long-form measures. For example, the five-item short-form Mental Health scale used in the SF-36 has 84% fewer items than the full-length scale, with only a 7% drop in precision (McHorney et al., 1992). Table 7.1 provides a comparison of the number of items and corresponding reliability coefficients for three different versions of the Mental Health Inventory (MHI). The implications for precision in comparing groups are shown in the last column. A well constructed short-form measure can provide a reasonable balance between the requirements of reliability and the needs of everyday use in clinical research and practice.

TABLE 7.1 NUMBER OF ITEMS AND RELIABILITY COEFFICIENTS FOR THREE VERSIONS OF THE MHI

Forms	Items	Administration Time (min)	Reliability	Relative precision[a]
MHI-32	32	5–8	.98	1.00
MHI-18	18	3–5	.96	.99
MHI-5	5	1 or less	.90	.93

[a] Reflects discrimination between groups of patients with and without psychiatric disorders.

Note: Data from "The validity and relative precision of MOS Short- and Long-Form Health Status Scales and Dartmouth COOP Charts: Results from the Medical Outcomes Study" C.A. McHorney et al., 1992, *Medical Care*, 30(Suppl.), MS253–265.

Summary of Findings

Estimates of score reliability for the SF-36 scales have been reported in 14 studies, as shown in Table 7.2. All estimates exceeded accepted standards for measures used in group comparisons. For each scale, the median of the reliability coefficients across studies equals or exceeds .80, with the exception of the Social Functioning scale (the median for this two-item scale is

TABLE 7.2 RELIABILITY ESTIMATES FOR SF-36 SCALES

Author	Year	Method	Reliability Estimates									Sample/Comments
			PF	RP	BP	GH	VT	SF	RE	MH	Range	
Stewart *et al.*	1989	IC								.88		Patients Seeing Doctor (MOS), N=9,335
Gelberg & Linn	1989	IC								.81		Homeless Adults, N=529
Cleary *et al.*	1991	IC				.90				.84	.84-.90	PTCA Patients, N=496
Wu *et al.*	1991	IC					.78			.86	.78-.86	AIDS Patients, N=137
Kantz *et al.*	1992	IC	.88	.90	.80	.83	.88	.77	.80	.82	.77-.90	Total Knee Replacement Patients, N=66
Kurtin *et al.*	1992	IC	.90	.76	.79	.82	.62	.76	.90	.67	.62-.90	Hemodialysis Patients, N=19
Nerenz *et al.*	1992	TRT	.90	.60	.43	.83	.68	.60	.60	.80	.43-.90	Diabetic Patients: 6 month Interval, N=235
Wachtel *et al.*	1992	IC								.82		HIV+ Patients, N=520
Brazier *et al.*[a]	1992	IC	.93	.96	.85	.95	.96	.73	.96	.95	.73-.96	General Practice Patients (U.K.), N=1,582
Brazier *et al.*[a]	1992	TRT	.81	.69	.78	.80	.80	.60	.63	.75	.60-.81	General Practice Patients: 2 week interval, N=187
Jenkinson *et al.*[a]	1993	IC	.90	.88	.82	.80	.85	.76	.80	.83	.76-.90	Random Sample of U.K. Population, N=9,332
Garratt *et al.*[a]	1993	IC	.92	.89	.86	.83	.86	.80	.66	.86	.80-.92	Patients with 1 of 4 Chronic Conditions, N=1310; and General Population Sample, N=542
McHorney *et al.*	*in press*	IC	.93	.84	.82	.78	.87	.85	.83	.90	.78-.93	Patients with 1 or More Chronic Conditions (MOS), N=3,445
McHorney *et al.*	*in review*	IC	.94	.89	.88	.83	.87	.63	.81	.82	.63-.94	Random Sample of U.S. Population, N=1,692[b]
McHorney & Ware	*forthcoming*	AF								.92		Patients with 1 or More Chronic Conditions (MOS), MHI Alternate Forms, N=1,437

Key: IC = Internal Consistency Reliability (Cronbach's α)
TRT = Test-Retest Reliability (Correlation)
AF = Alternate-Form Reliability (Correlation)

[a] Study used U.K. Developmental version of the SF-36.
[b] Estimates are based on the mail (self-administered) sample.

0.76). These results support the use of the SF-36 scales in studies of health status that are based on group-level analyses. Only the Physical Functioning scale consistently exceeded the 0.90 standard of reliability, which some consider a *minimum* standard for comparisons of scores for individual patients.

Most studies used the internal consistency method and Cronbach's coefficient alpha, although test-retest and alternate-forms estimates have also been reported. A range of patient populations and situations are represented in these studies, including patients at the time of doctor visits, patients with particular diagnoses (AIDS, orthopedics, renal disease, diabetes), general population studies in the United Kingdom and the United States, and patients with one or more conditions from the MOS. As in MOS studies reported to date, all of the published coefficients exceed the minimum standard of 0.50 suggested for group comparisons by Helmstadter (1964), and all but 11 exceed the 0.70 standard for individual comparisons suggested by Nunnally (1978).

Test-retest coefficients tend to be slightly lower than estimates based on the internal consistency method. Because both methods were used in the general population study reported by Brazier and his colleagues in the United Kingdom (1992), a direct comparison across methods is possible. The internal consistency estimates for the eight scales ranged from 0.73 to 0.96 with a median of 0.95, compared with a range of 0.60 to 0.81 and a median of 0.76 for the test-retest estimates in the U.K. population. These trends are important in understanding the internal consistency of the scales versus the stability of the health states they measure. If one believes that the items in each scale contain unique reliable variance (that is treated as error in the internal consistency estimates), one would expect the test-retest estimates to be higher than the internal consistency estimates. However, if one thinks that the scales are sensitive to relatively short-term changes such as those that would be observed in a 2-week interval between administrations, the test-retest estimates would be expected to be lower. The latter appears to be the case.

The test-retest correlations reported for diabetic patients by Nerenz and his colleagues (1992), which were based on an interval of 6 months between administrations, ranged from 0.43 to 0.90 with a median of only 0.64. Analyses from research in progress in the MOS confirm that substantial change occurs during a 6-month interval. Thus, it is reasonable to suspect

that these "test-retest" estimates underestimate the reliability of scores. They better represent the *stability* of the scores over a 6-month interval.

Not surprisingly, the largest number of reliability coefficients have been published for the five-item SF-36 Mental Health scale, also known as MHI-5. Results have been reported in 15 studies, as would be expected given that it was widely used before it was selected for the SF-36. It was used in the MOS SF-20 and in the Functional Status Questionnaire (Jette et al., 1986).

The level of reliability achieved by the SF-36 scales is lower than that achieved by the full-length versions of the MOS scales they were constructed to reproduce (e.g., mental health, general health perceptions) (McHorney et al., 1992, in press). Further, it has been well demonstrated that the SF-36 scales achieve substantially higher reliability than the single-item measures they replaced, such as those used in the SF-20 (e.g., social functioning, bodily pain) (McHorney et al., 1992).

Group Differences in Reliability Coefficients

The reliability of SF-36 scale scores has also been estimated using internal consistency methods for 24 subgroups of patients participating in the MOS (McHorney et al., in press). These patients (N=3,445) differed in sociodemographic characteristics, diagnosis, and disease severity. Table 7.3 summarizes these results. These coefficients varied across groups, with a range of coefficients from 0.65 to 0.94 across SF-36 scales and these subgroups. However, as shown in Table 7.3, minimum standards of reliability for purposes of group comparisons were satisfied in all 24 patient subgroups for all SF-36 scales. Minimum reliability standards required for comparisons of individual patients were most consistently met with the Physical Functioning scale across patient subgroups.

TABLE 7.3 RELIABILITY ESTIMATES FOR SF-36 SCALES IN MOS SUBGROUPS (N=3,445)

		PF	RP	BP	GH	VT	SF	RE	MH
Age	< 65	.92	.83	.81	.79	.87	.85	.83	.90
	65–74	.92	.86	.85	.78	.86	.84	.82	.88
	≥ 75	.92	.85	.81	.77	.82	.83	.82	.86
Gender	Female	.93	.84	.82	.79	.87	.84	.82	.90
	Male	.92	.84	.82	.77	.85	.87	.84	.89
Race	White	.93	.84	.83	.79	.88	.85	.82	.90
	Black	.93	.82	.80	.76	.83	.82	.82	.89
	Other	.92	.86	.80	.75	.82	.84	.82	.89
Education	< 8	.94	.88	.85	.79	.76	.85	.87	.89
	9-11	.94	.88	.87	.80	.86	.86	.84	.90
	12	.92	.82	.81	.78	.86	.85	.81	.90
	> 12	.92	.83	.80	.77	.88	.85	.82	.90
Poverty Status	Poverty	.94	.86	.85	.79	.85	.82	.86	.90
	Non Poverty	.92	.84	.81	.76	.88	.85	.82	.90
Diagnosis	Hypertension	.93	.84	.83	.77	.86	.84	.81	.87
	Diabetes	.93	.85	.86	.76	.86	.86	.81	.88
	CHF	.92	.82	.83	.78	.87	.90	.87	.84
	MI	.92	.83	.74	.78	.84	.84	.80	.86
	Clinical depression	.93	.84	.82	.79	.86	.82	.77	.86
	Symptomatic depression	.94	.83	.80	.77	.84	.79	.80	.86
Disease Severity	Uncomplicated Medical	.90	.81	.79	.75	.84	.76	.79	.84
	Complicated Medical	.91	.83	.82	.78	.87	.87	.83	.82
	Psychiatric and Uncomplicated Medical	.91	.80	.79	.72	.82	.84	.79	.85
	Psychiatric and Complicated Medical	.93	.70	.75	.65	.80	.72	.51	.88

Note. From "The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups" C.A. McHorney et al., in press, *Medical Care*.

Precision

Statistical Power

Better measures usually increase statistical power, which is the probability that a difference will be found when there is one. Measurement precision is important because "noisy" measures may add to the total variance in scores, which reduces power. Statistical power also improves with other features of study design, including the size of the difference (effect size) under study, the size of the sample, the number of groups compared, and how comparison groups were formed. For example, larger differences will be easier to detect, and a difference of any particular size is easier to detect with a larger sample. Comparisons of repeated measures between groups are generally more powerful when groups are formed randomly. Sample size is often a more important factor than measurement error in determining statistical power. That is why comparisons between groups can be performed successfully with less reliable measures (e.g., in the 0.50 to 0.70 range) than required for comparisons involving individual scores (Nunnally, 1978). We conclude this section with confidence intervals for individual scores.

Tables 7.4 through 7.8 present estimates of sample sizes necessary to detect small to large group differences in average SF-36 scale scores. We relied upon formulas published by Cohen (1988) and variance estimates from general U.S. population studies in estimating these sample sizes. We estimated sample sizes for five different designs beginning with the most powerful, an experimental comparison between two randomly formed groups with comparisons between repeated assessments over time. The least powerful design is a comparison between two group means. For all estimates, we assume a non-directional hypothesis (two-tailed test) with a false rejection rate of 5% and with a statistical power of 80%. In other words, a real difference in either direction is of interest and would be detected 80% of the time.

The differences in scores (effect sizes) modeled in Tables 7.4 to 7.8 were selected to represent two extremes: — very small and very large differences. Both are relevant to the design of outcome studies. Very small differences (2 points on a 100-point scale) are relevant to a study designed to rule out either a good or a bad health outcome on a population basis. Large sample sizes are required for confidence, although the size varies considerably across the SF-36 scales and across various study designs. The three bipolar scales

TABLE 7.4 SAMPLE SIZE NEEDED PER GROUP TO DETECT 2–20 POINT DIFFERENCES IN CHANGES OVER TIME BETWEEN TWO EXPERIMENTAL GROUPS, REPEATED MEASURES DESIGN

Scale		Number of Points Difference			
		2	5	10	20
Physical Functioning	PF	1364	219	55	15
Role–Physical	RP	2921	468	118	30
Bodily Pain	BP	1412	227	57	15
General Health	GH	1046	168	43	11
Vitality	VT	1108	178	45	12
Social Functioning	SF	1295	208	53	14
Role–Emotional	RE	2752	441	111	28
Mental Health	MH	824	132	34	9

Note. Estimates assume alpha = 0.05, two-tailed t-test, power = 80% and an intertemporal correlation between scores of 0.60 (Cohen, 1988).

TABLE 7.5 SAMPLE SIZE NEEDED PER GROUP TO DETECT 2–20 POINT DIFFERENCES BETWEEN TWO EXPERIMENTAL GROUPS, POST-INTERVENTION MEASURES ONLY

Scale		Number of Points Difference			
		2	5	10	20
Physical Functioning	PF	2131	342	86	22
Role–Physical	RP	4564	731	183	47
Bodily Pain	BP	2205	354	89	23
General Health	GH	1634	262	66	17
Vitality	VT	1732	278	70	18
Social Functioning	SF	2023	324	82	21
Role–Emotional	RE	4301	689	173	44
Mental Health	MH	1287	207	52	14

Note. Estimates assume alpha = 0.05, two-tailed t-test, power = 80% (Cohen, 1988). These estimates also apply to comparisons between self-selected groups, which are more difficult to interpret.

TABLE 7.6 SAMPLE SIZE NEEDED PER GROUP TO DETECT 2–20 POINT DIFFERENCES BETWEEN TWO SELF-SELECTED GROUPS, REPEATED MEASURES DESIGN

		Number of Points Difference			
Scale		2	5	10	20
Physical Functioning	PF	1705	274	69	18
Role–Physical	RP	3652	585	147	37
Bodily Pain	BP	1765	283	71	19
General Health	GH	1308	210	53	14
Vitality	VT	1385	222	56	15
Social Functioning	SF	1618	260	66	17
Role–Emotional	RE	3442	551	138	35
Mental Health	MH	1030	165	42	11

Note. Estimates assume alpha = 0.05, two-tailed t-test, power = 80% and an intertemporal correlation between scores of 0.60 (Cohen, 1988).

TABLE 7.7 SAMPLE SIZE NEEDED TO DETECT 2–20 POINT DIFFERENCES OVER TIME WITHIN ONE GROUP

		Number of Points Difference			
Scale		2	5	10	20
Physical Functioning	PF	852	137	35	9
Role–Physical	RP	1826	293	74	19
Bodily Pain	BP	883	142	36	10
General Health	GH	654	105	27	7
Vitality	VT	693	111	28	8
Social Functioning	SF	809	130	33	9
Role–Emotional	RE	1721	276	69	18
Mental Health	MH	515	83	21	6

Note. Estimates assume alpha = 0.05, two-tailed t-test, power = 80% and an intertemporal correlation between scores of 0.60 (Cohen, 1988).

TABLE 7.8 SAMPLE SIZE NEEDED TO DETECT 2–20 POINT DIFFERENCES BETWEEN A GROUP MEAN AND A FIXED NORM

		Number of Points Difference			
Scale		2	5	10	20
Physical Functioning	PF	1067	171	44	12
Role–Physical	RP	2282	366	92	24
Bodily Pain	BP	1103	177	45	12
General Health	GH	818	132	34	9
Vitality	VT	866	139	36	10
Social Functioning	SF	1012	163	41	11
Role–Emotional	RE	2152	345	87	22
Mental Health	MH	644	104	27	8

Note. Estimates assume alpha = 0.05, two-tailed t-test, power = 80% (Cohen, 1988).

(GH, VT, and MH) do the best in this regard. The 5-point column defines differences that are clinically and socially relevant. The 10-point column shows sample sizes necessary for moderate differences (0.50 SD units for most scales and roughly 0.30 SD units for the RP and RE scales).

Experimental Studies

Tables 7.4 and 7.5 present sample size estimates for two experimental study designs: two randomly formed groups with repeated SF-36 administrations and two randomly formed groups with post-intervention SF-36 assessments only. The repeated measures experimental design, as do others discussed in the following section, assumes a correlation of 0.60 between administrations of the SF-36. This is a reasonable assumption. Correlations ranging from 0.60 to 0.81 have been reported in test-retest studies (Brazier et al., 1992) and from 0.43 to 0.90 for repeated administrations 6 months apart (Nelsenz et al., 1992). The MOS has observed correlations ranging from 0.52 to 0.83 between repeated administrations 6 months apart.

Table 7.4 presents sample size estimates for a two-group randomized groups experiment with repeated SF-36 measures. This table reveals several trends

that are true of all study designs: (1) it takes many more subjects to detect small differences (2 points) than to detect large differences (20 points), and (2) approximately twice as many subjects are required to detect a given difference using the relatively coarse RP and RE disability scales.

Table 7.5 presents sample size estimates for comparisons between two experimental groups with post-intervention SF-36 measures only. When compared with those in Table 7.4, the estimates in Table 7.5 reveal the gains in power from a repeated measures experimental design. The number of subjects required to detect a difference is approximately 36% fewer for the repeated measures design than for the design with post-intervention measures only. For example, 2,131 subjects are required to detect the smallest difference in PF scores in Table 7.5 compared with only 1,364 in Table 7.4. The 0.60 correlation between repeated measures leads to a 36% reduction in the variance in scores.

Non-Experimental Studies

Tables 7.6 to 7.8 present sample size estimates for three non-experimental comparisons involving SF-36 scales: (1) comparisons between two self-selected groups with administrations before and after intervention(s) (Table 7.6); (2) repeated measures over time for a single group (Table 7.7); and (3) a comparison between a group mean score and a fixed score, such as the general population norm (Table 7.8).

The sample size estimates in Table 7.6 for a non-experimental, two-group study with repeated SF-36 measures assumes that difference scores will be analyzed to maximize the internal validity of the design. Comparisons between the entries in Table 7.6 and those in Table 7.4 illustrate the power gained from an experimental versus non-experimental two-group comparison. This gain is roughly 20%, for example, 1,705 versus 1,364 for the smallest difference in PF scale scores for non-experimental and experimental designs, respectively.

As shown in Table 7.7, sample sizes required to detect a change in SF-36 scale scores over time for a single group are much smaller than those discussed above. However, the results are likely to be much more difficult to interpret than those for comparisons between groups each receiving different treatment interventions (Cook & Campbell, 1979).

Finally, Table 7.8 presents estimates of the sample sizes needed to compare average SF-36 scale scores with a fixed norm, such as a general population norm. For example, a difference of 10 points on the PF scale between a group mean and a norm can be detected with only 44 subjects in the group, compared with 171 subjects for a difference of five points.

Confidence Intervals for Individual Scores

Measurement reliability is most important in determining the amount of fluctuation in a single score, such as the score for a particular patient. For example, as illustrated elsewhere (McHorney et al., 1992) the 95% confidence interval around the mental health score for an individual patient would be about ±13 points on a 100 point scale with a reliability of 0.90, which was the reliability observed in the MOS. In contrast, the 95% confidence interval around an individual score from a single-item measure with a reliability of 0.65 would be ±32 points, which is an increase of about 150%. Thus, whereas score reliability, at the levels observed to date for the SF-36 scales, is not an issue in group-level analyses, it may be more of an issue in studies of individual patients. Further research is necessary on this important issue.

The size of the confidence interval (CI) around an individual score is a function of the SD of the score distribution and the standard error of measurement (SEM). The size of the SEM, which is also referred to as the standard error of a score, is determined by the reliability of that score (Nunnally, 1978).

Table 7.9 presents estimates of CIs for SF-36 scale scores for an individual respondent's score. These estimates are based on the reliability of the scales in the general U.S. population. SDs used in estimating CIs are those for the general U.S. population (see Chapter 10). We have chosen to use SD estimates from the general population in order to standardize this important determinant of the CI, as suggested elsewhere (Mosteller et al., 1989).

The entries in Table 7.9 can be used to take into account fluctuations due to measurement error when interpreting scores for a patient or other individual. CIs for three levels of confidence are presented: 68% (1 SEM), 90% (1.64 SEM), and 95% (2 SEM). The use of these CI estimates is illustrated below.

A patient's score on the Physical Functioning (PF) scale would be expected

TABLE 7.9 SF-36 CONFIDENCE INTERVALS FOR INDIVIDUAL RESPONDENTS, GENERAL U.S. POPULATION

		Confidence Interval (CI)		
Scale	Label	68%[a]	90%[b]	95%[c]
Physical Functioning	PF	6.2	10.2	12.3
Role–Physical	RP	11.3	18.7	22.6
Bodily Pain	BP	7.5	12.4	15.0
General Health	GH	9.8	14.7	17.6
Vitality	VT	7.8	13.0	15.6
Social Functioning	SF	12.8	21.3	25.7
Role–Emotional	RE	14.0	23.2	28.0
Mental Health	MH	7.2	12.0	14.0

[a] 68% confidence interval equals 1 standard error of measurement (SEM).
[b] 90% confidence interval equals 1.64 SEMs.
[c] 95% confidence interval equals 2 SEMs.

to fall within 1 SEM (±6.2 points) about two-thirds (68%) of the time, according to Table 7.6. To be more certain about an individual PF score, use the 95% CI, which is ±12.3 for the PF scale.

Suppose that a clinician wanted to determine whether Mr. Smith, a 50 year old male who scored 70 on the PF scale, scored below the norm for his age and gender. The U.S. norm for a 50-year old male is about 86.5 (see Chapter 10). A score of 70 is more than two SEMs below the norm for a 50 year old male (86.5 - 70 = 16.5). Because 16.5 is greater than 12.3 (from Table 7.9), the clinician would be correct 95% of the time or more in concluding that Mr. Smith scored below the norm.

Data quality

The quality of SF-36 data can be evaluated in terms of the results of tests of scaling assumptions (Chapter 5), reliability of scores (discussed in this chapter), and the results of empirical tests of validity (Chapter 9). Two additional tests of data quality, the completeness of the data and response consistency, are discussed in the following section. Norms for these indicators are presented to facilitate evaluations of the quality of data gathered using the SF-36.

Data Completeness

Most surveys reproduce the SF-36 in a battery with other questions. The length of the total battery should be considered in evaluating data completeness. In the MOS, the SF-36 was embedded in a 245-item, self-administered questionnaire. The completeness of responses to SF-36 items has been extensively evaluated for 24 subgroups of MOS patients differing in sociodemographic characteristics, diagnosis, and disease severity (McHorney et al., in press). Rates of completed items ranged from 88% to 95% across scales in the total sample (N=3,445). Completion rates tended to be lower (75% to 91%) for older patients and for those reporting less education and tended to be higher for younger and more educated respondents (92% to 98%).

For most respondents with one or more items missing, an SF-36 scale score can be computed using the proration methods described in Chapter 6. For example, scale scores could be computed 90 to 98% of the time for groups with the most missing data in the MOS. These rates were 96 to 99% for the total sample and 98 to 100% for the groups with the least missing data.

Response Consistency

The quality of SF-36 data can also be evaluated by analyzing individual responses. Fifteen internal consistency checks based on pairs of SF-36 items have been identified and have proven useful in evaluating individual survey forms. These 15 checks are used to score the SF-36 Response Consistency Index (RCI). For example, a report of being able to walk "one mile" but not "one block" is considered an inconsistency in scoring the RCI.

Table 7.10 presents the frequency distributions of RCI scores for the general U.S. population (N=2,474) and for MOS patients with one or more chronic

conditions (N=3,434). For 90.3% of the general population and for 94.5% of MOS patients, inconsistent responses were not observed. Only 3.6% of the general population and 2.1% of MOS patients failed two or more checks, indicating problems with understanding how to complete the form or lack of motivation to respond carefully. For this reason, it is desirable to evaluate RCI scores for each patient before SF-36 profiles are interpreted. RCI scores are computed and printed for each survey form routinely by the SF-36 software included in the RT-2000 processing system. Information about RCI scores is available from The Health Institute.

TABLE 7.10 FREQUENCY DISTRIBUTION FOR THE RESPONSE CONSISTENCY INDEX (RCI)

Number of Inconsistent Responses	General Population (N=2,474)			MOS (N=3,434)		
	f	Percent	Cumulative Percent	f	Percent	Cumulative Percent
0	2234	90.3	90.3	3245	94.5	94.5
1	152	6.1	96.4	118	3.4	97.9
2	32	1.3	97.7	38	1.1	99.0
3	19	0.8	98.5	10	0.3	99.3
4	16	0.6	99.2	16	0.5	99.8
5	6	0.2	99.4	2	0.1	99.9
6	11	0.4	99.8	4	0.1	100.0
7	2	0.1	99.9	0	0.0	100.0
8	2	0.1	100.0	1	0.0	100.0
9-15	0	0.0	100.0	0	0.0	100.0

8. VALIDATION STRATEGIES AND INTERPRETATION GUIDELINES

Background

With the widespread use of the SF-36 to measure health status and to monitor outcomes, there is a greater need for useful interpretation guidelines. Validity is the extent to which a score means what it is supposed to mean — whether it has the intended interpretation. Guidelines for validating the SF-36 have been derived from those used to validate psychological and educational measures by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education (APA, 1985). This chapter explains SF-36 validation, compares the content of the SF-36 with that of other surveys, and summarizes interpretation guidelines.

The validity of questionnaires in the health field has most often been evaluated by means of content, construct, or criterion validation. Content validity (whether the test offers an adequate sample of the construct) is a challenge in the health field because of the breadth of health variables. Content validation requires the existence of a defining standard against which one can compare the content of a measure. Standards can be based on well-accepted theoretical definitions, on published standards, or on interviews with those who are experiencing the types of health problems under study. When construction of the SF-36 began 5 years ago, Ware published a set of standards for evaluating the content validity of general health measures intended to be comprehensive (1987). These standards were applied in constructing the SF-36.

When construct validation is used, both the test and the underlying theory must be evaluated. There are three steps to accumulate evidence of validity related to theoretical constructs: (1) specify the domain of variables, that is, prepare a blueprint for the constructs; (2) establish the internal structure of the observed variables; and (3) verify theoretical relationships between scale scores and external criteria. One method of testing the underlying theory is to test the differences between two patient groups known to differ in some way. For example, patients with a relatively minor and uncomplicated medical condition should score better in mental health (theorized construct) than patients with a psychiatric illness, and the average mental health scores of these patient groups should differ significantly. The mean

difference between such groups in the Medical Outcomes Study (MOS) was very large: 30.78 points on a 100 point scale (McHorney et al., 1993). The comparison demonstrates validity for the Mental Health (MH) scale because the mental health scores were much lower for patients with psychiatric disease (known to have poor mental health by definition of their disease).

Convergent and discriminant validity are at the foundation of construct validation. Convergent validity is supported when different methods of measuring the same construct provide similar results. Discriminant validity examines whether a measure of one underlying construct can be differentiated from another construct. For example, in the MOS, measures of physical functioning, mobility, and satisfaction with physical abilities were expected to yield results that "converge" at least moderately with one another because they are all hypothesized to assess physical health. In tests of discriminant validity, different measures are expected to yield different results. For example, one would not expect a measure of physical functioning to be highly related to a measure of depression or of loneliness. When more than one method of data collection or scale construction has been used to measure the same construct, they can be compared to test convergent validity; when both different methods and different constructs have been measured, both convergent and discriminant validity can be tested using the multitrait-multimethod procedure (Campbell & Fiske, 1959).

Criterion validity demonstrates that test scores are systematically related to one or more outcome criteria. This technique can be used when external evidence is available for use as a criterion against which the results of the test can be compared. In order to judge a measure in terms of external (known and independent) evidence, the evaluator must know what the anticipated result should be. Examples of correlations with external evidence occur when (1) health status and resource use are negatively correlated, (2) age and physical health are negatively correlated (according to the theory that physical function declines with increasing age), or (3) physical and mental health each have a positive correlation with general health.

Validity studies increase understanding of what a difference or a change in a score means. When enough evidence has been accumulated to show that a scale measures the intended health concept and does not measure other concepts, the scale is said to be *validated*. However, the process of validation continues as long as new information is produced about the interpretation

and meaning of scores. In the absence of agreed-upon criteria, or "gold standards," for validating health measures, norms can be very useful in interpreting scores. Guidelines for the norming of standardized surveys (APA, 1985) were followed in norming the SF-36 (see Chapter 10).

SF-36 Validation Strategy

Two kinds of strategies were used to evaluate the validity of the SF-36 and to accumulate information for interpreting scale scores. First, we judged content validity by comparing it with other widely used survey forms. Second, we used empirical approaches including factor analytic tests of construct validity, "criterion-based" approaches, and numerous correlational studies.

Comparison with Other Forms

Table 8.1 compares the content of the SF-36 with that of longer and shorter MOS measures that preceded it and with the content of seven other widely used measures. This table reveals that the SF-36 includes eight of the most frequently represented health concepts. One of the category/concept areas included in three or more surveys, but not included in the SF-36, is the Symptoms/Problems category. Table 8.1 also identifies those concepts that are not measured by the SF-36 but are included in other longer measures (e.g., the Sickness Impact Profile [SIP], MOS Long Form, and Health Insurance Experiment [HIE] battery). These omissions include sleep, cognitive functioning, and health distress.

Symptoms and problems that are specific to a particular condition are not included in the SF-36, because the SF-36 is a generic measure (see Chapter 3). As summarized in Chapter 9, SF-36 scales correlate substantially (r=0.40 or greater) with most of the omitted general health concepts and with the frequency and severity of many specific symptoms and problems. A noteworthy exception is sexual functioning, which is a good candidate for inclusion in questionnaires that supplement the eight SF-36 scales.

TABLE 8.1 COMPARISON OF CONTENT OF MOS AND OTHER WIDELY USED HEALTH SURVEYS

	MOS									
Concepts	Long-Form	SF-36	SF-20	HIE	NHP	SIP	DUKE	COOP	MHIQ	QWB
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						●				
Communication						●				
Quality of Life	●			●				●		

Long Form MOS 149-Item Functional Status and Well-Being Survey (Stewart & Ware, 1992)
SF-36 MOS 36-Item Short-Form Health Survey (Ware & Sherbourne, 1992)
SF-20 MOS 20-Item Short-Form Health Survey (Stewart et al., 1988; Ware, Sherbourne, & Davies, 1992)
HIE Health Insurance Experiment (Brook et al., 1979; Ware, Brook et al., 1980)
NHP Nottingham Health Profile (Hunt et al., 1981)
SIP Sickness Impact Profile (Bergner et al., 1981)
DUKE Duke Health Profile (Parkerson et al., 1990)
COOP Dartmouth COOP Function Charts (Nelson et al., 1990b)
MHIQ McMaster Health Index Questionnaire (Chambers, 1988)
QWB Quality of Well-Being Scale (Patrick et al., 1973)

How to interpret the SF-36

Table 8.2 summarizes information presented in Chapters 9 and 10 that may be most useful for interpreting scores for SF-36 scales. Table 8.2 orders scales according to their validity, from the scale known to be the most valid measure of the physical component of health status, Physical Functioning (PF), to the last scale in the table, Mental Health (MH), which is the most valid measure of the mental component of health status. Interestingly, MH is the poorest measure of the physical component, and PF is the poorest measure of the mental component. Scales in between PF and MH are ordered according to their validity in measuring physical and mental components. Scales in the middle have substantial or moderate validity for both components of health status and should be interpreted accordingly.

The number of items and the number of levels defined by each scale is presented. The most precise (least coarse) scales are those with 20 or more levels (PF, GH, VT, and MH). The relatively coarse role disability scales (RP and RE) each measure only four or five levels.

Means and standard deviations for each of the eight scales in the general U.S. adult population are also presented. These can be used to determine whether a group or individual in question scores above or below the U.S. average. Reliability estimates and confidence intervals for individual scores are also presented.

The remaining information in Table 8.2 under the headings "Validity," "Range," and "Definitions of Lowest and Highest Scores" is presented to facilitate interpretation. The summaries and graphics in Table 8.2 are based on all available evidence described in Chapters 8 through 10. The following section discusses three distinct interpretation issues: (1) the extent to which each scale measures the physical or mental component of health ("Validity"), (2) whether both limitations and well-being are measured by each scale ("Range"), and (3) the description of the health states that are assigned at the lowest and highest possible score on each scale. These summaries will help determine what results mean.

Three of the scales (PF, RE, and BP) have substantial validity as measures of physical health status. Each scale, however, addresses a different aspect of physical health. PF measures limitations in behavioral performance of

TABLE 8.2 SUMMARY OF INFORMATION ABOUT SF-36 SCALES

Scale	Label	Number of Items	Number of Levels	Mean[a]	SD[a]	Reliability[a]	CI[b]	Validity[c] P	Validity[c] M	Range[d]	Definitions of Lowest and Highest Scores: Lowest Possible *(Floor)*	Definitions of Lowest and Highest Scores: Highest Possible *(Ceiling)*
Physical Functioning	PF	10	21	84.2	23.3	.93	±10	●	○	▬	Limited a lot in performing all physical activities including bathing or dressing due to health	Performs all types of physical activities including the most vigorous without limitations due to health
Role–Physical	RP	4	5	81.0	34.0	.89	±19	●	○	▬	Problems with work or other daily activities as a result of physical health	No problems with work or other daily activities as a result of physical health
Bodily Pain	BP	2	11	75.2	23.7	.90	±12	●	○	▬	Very severe and extremely limiting pain	No pain or limitations due to pain
General Health	GH	5	21	72.0	20.3	.81	±15	◑	◑	▬▭	Evaluates personal health as poor and believes it is likely to get worse	Evaluates personal health as excellent
Vitality	VT	4	21	60.9	21.0	.86	±13	◑	◑	▬▭	Feels tired and worn out all of the time	Feels full of pep and energy all of the time
Social Functioning	SF	2	9	83.3	22.7	.68	±21	◑	●	▬	Extreme and frequent interference with normal social activities due to physical or emotional problems	Performs normal social activities without interference due to physical or emotional problems
Role–Emotional	RE	3	4	81.3	33.0	.82	±23	○	●	▬	Problems with work or other daily activities as a result of emotional problems	No problems with work or other daily activities as a result of emotional problems
Mental Health	MH	5	26	74.7	18.0	.84	±12	○	●	▬▭	Feelings of nervousness and depression all of the time	Feels peaceful, happy, and calm all of the time

[a] Mean, standard deviation, and reliability for the general U.S. population (total sample).

[b] CI = 90% confidence interval around the score for an individual, general U.S. population.

[c] P = physical health, M = mental health, ● = substantial, ◑ = moderate, and ○ = weak validity.

[d] ▬ = limitations/disability range; ▭ = well-being range.

everyday physical activities, RP measures the extent of disability in everyday activities due to physical problems, and BP focuses specifically on the severity of bodily pain and resulting limitations in activities.

The three best measures of the mental component of health status are the MH, RE, and SF scales. All have substantial validity for measuring mental health. They differ in the range of mental health measured, with perfect SF and RE scores earned by those reporting no limitations or disability due to personal or emotional problems. In contrast, the MH scale is a bipolar scale with a mid-range score earned by those reporting no symptoms of psychological distress. A score of 100 on this scale requires reports of frequently feeling happy, calm, and peaceful.

The SF-36 scales most sensitive to both physical *and* mental health outcomes are the VT and GH scales, which have moderate empirical validity for these two components. They are relatively precise scales, with 21 scale levels. A mid-range score on the VT scale is achieved by those who do not report feeling tired or worn out; a score of 100, in addition to indicating an absence of these symptoms, is only earned by those who report feeling full of pep and energy all of the time. A mid-range score is obtained on the bipolar GH scale by reporting no unfavorable evaluations of health in general.

As indicated in the "Range" column of Table 8.2, the highest possible score on all but three scales indicates the absence of a negative state (limitation, disability, pain).

Content-referenced interpretation, criterion studies of validity, and norm-referenced scoring all support the summary of guidelines for scale interpretation presented in Table 8.2. Chapters 9 and 10 presents detailed evidence of each of the three types.

Although a great deal of information has accumulated about SF-36 score interpretation, a great deal remains to be learned. Users of this manual are encouraged to use the information presented here with a healthy degree of caution and to publish their own findings and interpretation guidelines whenever possible.

9. VALIDITY: CONTENT- AND CRITERION-BASED INTERPRETATION

This chapter explains two different empirical strategies — approaches based on item content and external criteria — for validating and interpreting SF-36 scales. The first section explains content-based interpretation and presents results for the Physical Functioning (PF), General Health (GH), and Vitality (VT) scales for which this strategy has proven very useful. The second section explains criterion-based approaches to validation and presents results for six of the eight scales (PF; Role-Physical [RP], Bodily Pain [BP], GH, Role–Emotional [RE], and Mental Health [MH]) and for the self-reported health transition item.

The remaining six sections are devoted to empirical studies involving all eight scales. These sections are organized by type of study and include: factor analytic studies and a comparison of the factor content and validity of each scale in relation to clinical criteria; correlations with general health and quality of life measures; correlations with specific symptoms; correlations with Medical Outcomes Study (MOS) measures of concepts not represented in the SF-36; correlations with measures developed by others; and, finally, a section documenting the size of differences in scale scores that should be considered large.

Content-based interpretation

Table 9.1 describes the health states associated with the lowest and highest possible score, which define the "floor" and "ceiling" for each SF-36 scale. These descriptions are based on item content and the pattern of responses across items necessary to achieve these extreme scores.

The content of individual items can also be used to better understand differences in scale scores between the extremes. This is accomplished by plotting responses to a given item across the levels of the scale containing that item. For example, it is useful to know that about 90% of the population can walk one block without limitations at a score of 75 on the PF scale, whereas only about 32% can do so at a PF score of 45. This is an example of content-based interpretation.

TABLE 9.1 CONTENT-BASED DESCRIPTIONS OF LOWEST AND HIGHEST SCALE SCORES

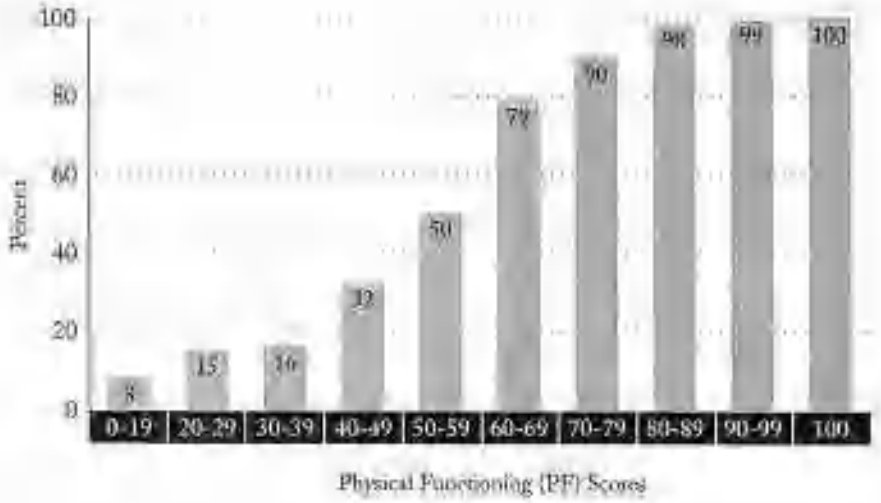
	Meaning of Scores	
Concepts	Lowest Possible *(Floor)*	Highest Possible *(Ceiling)*
Physical Functioning	Limited a lot in performing all physical activities including bathing or dressing due to health	Performs all types of physical activities including the most vigorous without limitations due to health
Role-Physical	Problems with work or other daily activities as a result of physical health	No problems with work or other daily activities as a result of physical health
Bodily Pain	Very severe and extremely limiting pain	No pain or limitations due to pain
General Health	Evaluates personal health as poor and believes it is likely to get worse	Evaluates personal health as excellent
Vitality	Feels tired and worn out all of the time	Feels full of pep and energy all of the time
Social Functioning	Extreme and frequent interference with normal social activities due to physical or emotional problems	Performs normal social activities without interference due to physical or emotional problems
Role-Emotional	Problems with work or other daily activities as a result of emotional problems	No problems with work or other daily activities as a result of emotional problems
Mental Health	Feelings of nervousness and depression all of the time	Feels peaceful, happy, and calm all of the time

Physical Functioning

Figure 9.1 shows the percentage of the general U.S. population (N=2,474) that reports being able to walk one block without a limitation due to health (Item 3i) at various PF scale levels. For this graph, the 21 PF scale levels were collapsed to 10 levels (as defined in the figure). At the top three levels, nearly everyone (98 to 100%) reported walking without a limitation. At the bottom three levels, only 8% to 16% reported walking without a limitation. Across the middle four levels, 32%, 50%, 79%, and 90% reported being able to walk one block. Thus, a difference between PF scale scores of about 45 versus 55 represents about an 18 percentage point difference (50% -32% = 18%) in the ability to walk one block without limitation.

We also estimated the percentage reporting limitations for each of the other nine PF scale items (results not reported). These analyses yielded results that

FIGURE 9.1
PERCENTAGE THAT CAN WALK ONE BLOCK OR MORE



may be useful in the interpretation of the scale. For example, 98.2% of those at the highest scale level reported being able to perform vigorous activities without limitations. The percentage giving this response at the other nine scale levels ranged from a low of 2.5% for the lowest level to a high of 16.9% at the second highest scale level. Thus PF scale Item 3a (vigorous activities) determines whether or not a respondent scores at the highest level and does not vary much across the lower levels.

As expected, limitations in the "bathing and dressing" item (3j) were relatively rare (from 0 to 25%) across the top six categories (top 11 scale levels). A majority of endorsements was not observed for this item until the 9th and 10th PF scale levels (69% and 74%, respectively). Thus, the bathing and dressing item defines the bottom of the PF scale, as hypothesized. With only one exception, the percentage reporting performance without limitations for each of the 10 PF items increased from the lowest to the highest PF level.

General Health

Table 9.2 shows the percentage of the general U.S. population (N=2,450) that evaluates their health as "excellent," "good," and "fair or poor" at 10 levels of the GH scale. The 21 scale levels have been collapsed to 10 as defined in the table. The choice of "excellent" clearly defines the top of the scale, "fair" and "poor" the bottom, and "good" defines the middle scale levels.

TABLE 9.2 PERCENTAGE EVALUATING THEIR HEALTH AS EXCELLENT, GOOD, AND FAIR/POOR AT TEN LEVELS OF THE GENERAL HEALTH (GH) SCALE, GENERAL U.S. POPULATION (N=2,450)

	GH Scale		Percentage		
Levels	Scale Mean	(N)	Excellent*	Good*	Fair/Poor*
91-100	96.9	423	59.5	3.1	0.3
81-90	85.0	552	25.7	17.5	1.2
71-80	75.0	535	11.8	37.3	4.2
61-70	65.0	371	5.9	60.3	7.3
51-60	55.6	214	3.3	59.0	21.1
41-50	45.7	118	1.7	40.5	50.5
31-40	36.9	99	1.9	27.1	66.5
21-30	27.3	77	0.8	9.0	89.4
11-20	18.5	38	0.0	7.5	92.5
0-10	8.7	23	0.0	4.6	95.4

* Responses to Item 1: "In general, would you say your health is:"

Vitality

Table 9.3 shows the percentage of the general U.S. population (N=2,461) that reports "feeling tired" and having a "lot of energy" all or most of the time during the past 4 weeks. These two responses clearly define the bottom and top scale levels, respectively.

Criterion-based interpretation

Criterion-based tests of validity are based on measures and data that are independent of the scale in question. These include other measures and data that are gathered concurrently (concurrent validity) or after some interval of time (predictive validity). For some clinical "criteria" illustrated here (e.g., severity of disease), concurrent measures are most appropriate. In nearly all instances, these "criteria" are actually conceptually related variables that provide meaningful empirical tests for a particular clinical application of each scale. They are not criteria in the sense of being a "gold standard."

We discuss criterion-based tests of validity and interpretation guidelines

TABLE 9.3 PERCENTAGE REPORTING FEELING TIRED AND HAVING A LOT OF ENERGY AT TEN LEVELS OF THE VITALITY (VT) SCALE, GENERAL U.S. POPULATION (N=2,461)

	VT Scale		Percentage	
Levels	Scale Mean	(N)	Tired[a]	Lot of Energy[a]
91-100	96.9	73	0.0	100.0
81-90	86.1	225	0.0	95.9
71-80	76.9	541	0.2	86.9
61-70	67.5	449	2.1	45.7
51-60	57.5	401	5.0	14.4
41-50	49.0	308	8.3	5.7
31-40	38.3	194	30.2	4.0
21-30	29.9	126	62.3	2.3
11-20	19.6	85	89.2	0.0
0-10	10.6	65	96.5	0.0

[a] All or most of the time, past 4 weeks.

below for six scales and the health transition item. The "criteria" were selected because they: (1) are important (clinically, socially); (2) represent plausible outcomes of the variations in functioning and well-being measured by the scales; and (3) were measured independently of the scale in question. Chapter 8 presents a summary of findings and interpretation guidelines.

Physical Functioning and Ability to Work

An important consequence of limitations in physical functioning, as measured by the PF scale, is that they may prevent work at a paying job. Table 9.4 presents the percentages of MOS panel participants at each of 10 levels of the PF scale that reported that their health kept them from working at a paying job (N=2,192). These percentages range from a high of 68.9% for PF scores below 20 to a low of 3% to 6% for scores between 80 and 100. A neatly perfect ordering of disability is apparent between the highest and lowest PF scale levels. These percentages can be used in interpreting the PF scale and explaining the social relevance of differences in scale scores.

TABLE 9.4 PERCENTAGE OF MOS PATIENTS THAT CANNOT WORK BECAUSE OF HEALTH PROBLEMS, TEN LEVELS OF THE PHYSICAL FUNCTIONING (PF) SCALE (N=2,192)[a]

Levels[b]	PF Scale Range	PF Scale Mean	N	% Cannot Work
1	100	100	338	4.7
2	90-99.9	95.0	253	3.2
3	80-89.9	87.6	360	6.1
4	70-79.9	77.3	291	18.9
5	60-69.9	67.7	207	24.2
6	50-55.9	55.3	255	36.9
7	40-49.9	44.9	55	38.2
8	30-39.9	37.0	128	61.7
9	20-29.9	26.9	112	54.5
10	0-19.9	10.8	193	68.9

[a] MOS longitudinal sample eligible to work.

[b] 21 PF scale levels collapsed to 10.

Role–Physical Scale and General Health

The validity of the RP scale was evaluated by computing mean GH scale scores for the general U.S. population at each of the five scale levels (Table 9.5). As hypothesized, these means differed substantially and were ordered consistently with the scale levels (from a low of 46.4 to a high of 77.5, $F=56.1$, $p < 0.001$).

To test the "interval" scoring of the RP scale, the range of GH scores (77.5 – 46.4 = 31.1) was transformed by setting the highest and lowest means to 100 and 0, respectively, and transforming mean GH scores for the three intermediate levels to the percentage of the observed range. For example, the second level was transformed: 53.4 - 46.4 = 7 and 7 divided by 31.1 = 22.5. Transformed "criterion" scores are presented in the right-hand column of Table 9.5. These transformed values are very similar to the standard SF-36 scores assigned to the three intermediate scale levels as shown in the first column (25 versus 22.5, 50 versus 47.9, and 75 versus 71.1).

Different permutations of responses across the four RP scale items that are scored at the same level are currently being evaluated for equivalence. Table

TABLE 9.5 MEAN GENERAL HEALTH (GH) SCORES FOR RESPONDENTS AT FIVE LEVELS OF THE ROLE–PHYSICAL (RP) SCALE, GENERAL U.S. POPULATION (N=2,422)

			General Health Evaluation[a]	
Score	f	%	Mean	Transformed
100	1580	65.2	77.5	100.0
75	212	8.8	68.5	71.1
50	141	5.8	61.3	47.9
25	172	7.1	53.4	22.5
0	317	13.1	46.4	0.0

[a] Average GH scale score.

Note. $F = 56.1$, $p < 0.001$, for differences among means across levels.

9.5 documents substantial differences in overall evaluation of health across the five RP scale levels and supports standard SF-36 scoring of those levels.

Bodily Pain and Ability to Work

A social consequence of bodily pain is that it may prevent work at a paying job. Table 9.6 presents the percentages of MOS panel participants at each of 10 levels of the BP scale who reported that their health kept them from working at a paying job (N=2,187). These percentages ranged from a high of 74.9 for level 10 (BP scores between 0 and 19.9) to a low of 8.7% for the third scale level. Percentages for the top three levels ranged from 8.7 to 12.0. A perfect ordering of disability rates is apparent for the fourth through 10th BP scale levels. A very large increase in disability (60.8 to 74.9%) was observed for the three lowest levels. These percentages can be used in interpreting the BP scale and in explaining the social relevance of differences in scale scores.

General Health and Utilization of Health Care Services

The GH scale is often used as a "criterion" in validating other scales because it is a direct measure of the respondents' personal evaluation of his/her health. Various forms of the GH scale have been extensively evaluated (Davies & Ware, 1981; Manning et al., 1982; Ware, 1976; 1984b; Ware & Karnos, 1976). Although we expect experience with longer forms of the GH scale

TABLE 9.6 PERCENTAGE OF MOS PATIENTS THAT CANNOT WORK BECAUSE OF HEALTH PROBLEMS AT TEN LEVELS OF THE BODILY PAIN (BP) SCALE (N=2,187)[a]

Levels[b]	BP Scale Range	BP Scale Mean	(N)	Criterion (%)[c]
1	100	100.0	350	12.0
2	90-99.9	92.5	291	10.1
3	80-89.9	83.5	329	8.7
4	70-79.9	72.8	288	12.9
5	60-69.9	61.8	284	18.2
6	50-59.9	51.4	230	27.5
7	40-49.9	41.3	185	34.7
8	30-39.9	31.2	147	60.9
9	20-29.9	23.9	102	62.5
10	0-19.9	7.0	71	74.9

[a] MOS longitudinal sample eligible to work.

[b] 11 BP scale levels collapsed to 10; level 10 collapses two scale levels.

[c] *Criterion* = Does your health keep you from working at a paying job?

to generalize to GH scale items, we chose to test that assumption, as discussed below. GH scale and item scores have been linked to several indicators of the utilization of health care services in the MOS (Kravitz et al., 1992). Table 9.7 summarizes results for SF-36 Item 1, which will be useful in interpreting both that item and overall scores for the GH scale that contains it. These data were obtained from adults (18 years and older) who visited MOS providers during a 9-day period in 1986. Sample sizes for specific utilization indicators varied by design from 9,020 to 17,780 because patients were randomly assigned to forms requesting different information. Additional details are presented elsewhere (Kravitz et al., 1992).

As Table 9.7 shows, the utilization rates for all three indicators (hospitalizations, annual office visit rates, and prescriptions per visit) were greater for patients with less favorable general health perceptions. For example, the hospitalization rate during the past 3 months was nearly 10 times higher for those who evaluated their health as "poor" compared with "excellent." The annual visit rate was two and one-half times greater for the "poor" group, and the prescription rate per visit was nearly four times greater in the "poor"

TABLE 9.7 HEALTH CARE UTILIZATION RATES FOR PATIENTS DIFFERING IN GENERAL HEALTH EVALUATIONS

General Health (GH) Item 1	Scale Score	Percent Hospitalized Past 3 Months	Annual Visit Rate per Year	Prescriptions per Visit
Excellent	100	2.7	3.09	0.8
Very Good	84	3.5	3.84	1.1
Good	61	5.9	4.88	1.7
Fair	25	14.5	6.55	2.6
Poor	0	25.8	8.11	3.1

Note. From "Differences in the mix of patients among medical specialties and systems of care: Results from the Medical Outcomes Study" by R.L. Kravitz et al., 1992, *Journal of the American Medical Association, 267,* 1617-1623.

group as compared to the "excellent" group. These results are consistent with previous findings for other forms of the GH scale in predicting outpatient utilization (Manning et al., 1982).

MOS results are also consistent with unpublished estimates of hospital expenditures for adults differing in General Health Rating Index (GHRI) scores in the Health Insurance Experiment (HIE), based on models of insurance claims data (Newhouse et al., 1981). (The SF-36 GH scale was constructed to reproduce the GHRI.) Annual expenditures for hospital services for those scoring in the bottom 20% of the GHRI score distribution (scores of 0 to 57) totaled more than $900 in the following year (1981 dollars) compared with less than $300 for those scoring in the top 20% (scores of 84 to 100). This represents more than a three-fold difference. Expenditures for the three quintiles in between were approximately $300, $400, and $500 (from better to worse health).

Role–Emotional Scale and Mental Health

Table 9.8 presents MH scale scores for four levels of the RE scale. The range of differences in MH scores across the four levels was transformed to 0 to 100 to test the appropriateness of the standard SF-36 scoring of the second and third RE levels. (The logic is the same as described for the RP scale.) If SF-36 scoring for the two intermediate RE levels is appropriate (66.7 and 33.3 as shown in column 1), we would expect transformed MH scores for

TABLE 9.8 MEAN MENTAL HEALTH SCORES FOR RESPONDENTS AT FOUR LEVELS OF THE ROLE-EMOTIONAL SCALE, GENERAL U.S. POPULATION (N=2,419)

	Prevalence		Mental Health Scale	
Score	f	%	Mean	Transformed
100	1687	69.7	80.8	100.0
66.7	267	11.0	70.4	64.3
33.3	197	8.2	61.1	32.3
0	268	11.1	51.7	0.0

Note. $F = 113.2$, $p < 0.001$, for differences among means across levels.

individuals at those levels to be roughly equivalent, on average, to SF-36 scoring.

Large differences in average MH scale scores were observed for MOS patients across the four RE scale levels ($F=113.2$, $p < 0.001$), and means for the MH "criterion" were ordered exactly as hypothesized. Thus, according to this "criterion," the RE scale is at least an ordinal scale.

The differences in MH scores between RE levels are also approximately equal (roughly 10 units on the 0 to 100 MH scale). As would be expected from this pattern of results, the transformed MH scale scores were roughly evenly spaced, on average. The second best level, which is scored 66.7 using standard SF-36 scoring, is approximately 64.3% of the range between the lowest and highest levels. The average MH scale score for those at the RE scale level next to the lowest, which is scored 33.3, was 32.3% of the criterion range. These results support the scoring and interpretation of the RE scale as a roughly "interval" measure.

Three different permutations of responses across the three RE items were observed at each of the two intermediate RE levels. However, the great majority of respondents at each level earned their score with the same pattern. Further, mean "criterion" scores across these permutations were nearly identical. Thus, it is not likely that much will be gained from recalibrations of different permutations of response to the RE scale, at least not in relation to a general mental health "criterion."

The implication of these results for interpretation is that the RE scale defines substantial differences in mental health burden. The differences in MH scores between RE scale levels are large and have been shown to have noteworthy clinical and social consequences. Although the RE scale is the most coarse SF-36 scale, the differences among its levels appear to be substantial and should be interpreted accordingly.

Mental Health

Because it is the "flagship" SF-36 measure of mental health, much effort has been focused on the MH scale. To illustrate the criterion-based approach to score validation and to provide information useful in the interpretation of MH scale scores, eight mental health "criteria" were studied using data from the MOS. Results for all criteria are summarized following a brief summary of operational definitions.

Mental health criteria were scored dichotomously (1,0) to define an undesirable mental health state or outcome, as follows:

(1) **Dissatisfaction with Life** - respondent was generally dissatisfied or very dissatisfied with his/her personal life according to Dupuy's (1984) measure of quality of life (Ware et al., 1979);

(2) **Depressive Symptoms** - scored beyond the CES-D cut-off for clinical depression but did not satisfy DSM-III criteria for major depression or dysthymia during the 1 to 4 month period prior to completing the MH scale (Wells et al., 1989);

(3) **Diagnosis of Depression** - met DSM-III criteria for major depression, dysthymia, or both after screening CES-D positive during the 1 to 4 month prior period (Wells et al., 1989);

(4) **Suicide Ideation** - self-report of thinking of taking own life a couple of times or more during the past month, item from the Mental Health Inventory (MHI) (Ware et al., 1979);

(5) **Mental Health Care (Outpatient)** - self-report of one or more outpatient visits for mental health treatment during the past 6 months;

(6) **Mental Health Specialty Care** - respondent was sampled from a formally trained mental health specialist's office;

(7) **Mental Health Care (Inpatient)** - self-report of one or more overnight stays in a hospital for an emotional/mental health problem, during the past 12 months; and

(8) **Mental Health Scale** - average score (0 to 100 scale) on the 32-item MOS version of the long-form MHI (Stewart, Hays, & Ware, 1992).

For each criterion, percentages at each of the six response categories were averaged across the five MH items to summarize results. For negatively worded items, the first category was a choice of "all of the time;" for positively worded items, the first category was a choice of "none of the time" (see footnote to Table 9.9).

Table 9.9 presents the average percentage of MOS patients that scored positively on each criterion for each of the six response categories across the five MH scale items. Thus, the first table entry of 70.8% is the average percentage of respondents who reported dissatisfaction with their life at the lowest response category across the five MH items (i.e., a category scored 0). The second panel of the table averages these percentages across the seven independent criteria. The last panel presents average MHI scale scores (MOS long-form) for respondents at each of the six item response categories for the MH scale. Figure 9.2 illustrates these averages.

According to the theory underlying construction and interpretation of the MH scale, these percentages should vary considerably across response categories and across criteria. The worst outcomes (e.g., inpatient mental health care, suicidal ideation, diagnosis of depression) should be least prevalent at the highest scale levels (80 and 100). Less serious outcomes (dissatisfaction with life, depressive symptoms without a diagnosis) should be more prevalent. The percentage rates for all criteria should be a monotonical function of the ordered response categories; that is, these percentages should be ordinally consistent with the ordering of response categories. Further, ideally for the scaling method used, we expect the average percentage across the seven independent criteria to be roughly linearly related to the scores assigned to the six response categories (Nunnally, 1978). Without exception, percentages for independent criteria studied to date have been monotonically related to MH response categories/scale scores. The average prevalence (across criteria) decreases by approximately 10 percentage points (e.g., from 58.6 to 48.1 for the lowest and next lowest categories) from level to level (second panel

TABLE 9.9 AVERAGE PERCENTAGE OF INDEPENDENT CRITERION SCORES OBSERVED FOR THOSE CHOOSING SIX RESPONSE CATEGORIES IN THE MENTAL HEALTH (MH) SCALE (N=2,988)

	Item Response Categories (Scale Scores)[a]					
Independent Criteria	All (0)	Most (20)	Good Bit (40)	Some (60)	A Little (80)	None (100)
Dissatisfaction with Life	70.8	42.4	29.1	12.4	4.5	1.0
Depressive Symptoms	78.8	78.7	68.6	50.1	30.0	12.5
Diagnosis of Depression	54.9	46.2	30.9	18.4	7.6	2.2
Suicide Ideation	43.1	29.6	14.4	5.6	1.9	.5
Mental Health Care:						
Outpatient (past 6 months)	51.6	45.8	37.4	25.8	13.2	4.6
Specialist (current)	72.9	66.8	60.5	47.0	32.1	17.0
Inpatient (past 12 months)	39.3	27.2	16.6	12.2	6.9	3.1
Average % Across Criteria	58.6	48.1	36.8	24.5	13.7	5.9
		Δ=10.5	Δ=11.3	Δ=12.3	Δ=10.8	Δ=7.9
Mental Health Scale[b]						
Long Form Average	35.6	43.1	52.9	62.8	73.9	84.5

a These categories are scored 1 to 6 (all to none) for negatively worded items (e.g., downhearted and blue), and from 6 to 1 (all to none) for positively worded items (e.g., happy person). Scale scores are then transformed to range from 0 to 100 (see Chapter 6).

b Scored 0 to 100 with scores in between indicating the percentage of the total score.

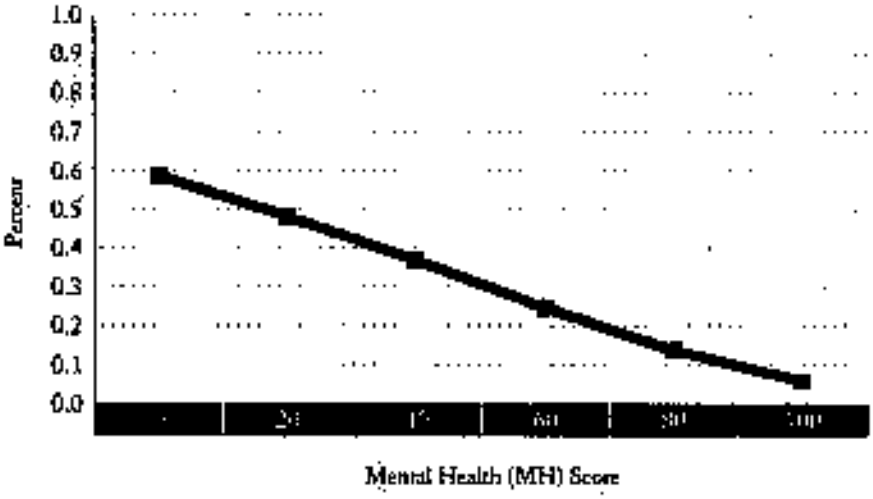
Note. Sample includes MOS patients with one or more chronic conditions and complete data for the MHI-5 (N=2,988).

of Table 9.9). These differences in average percentages between levels range from a low of 7.9 between scale values of 80 to 100 to a high of 12.3 for the difference between 40 and 60. The average of these percentages is roughly linearly related to the MH scale scores, as Figure 9.2 shows.

Finally, as would be expected given their very high intercorrelation (r=0.96), the long-form MOS MHI, which includes the five-item MH scale, yields approximately a 10-point difference between each of the six response categories (scale levels) shown in Table 9.9, slightly less for the 0 to 20 difference than for the other four intervals.

The clinical and social relevance of differences in MH scale scores is most apparent for the three lowest scale levels shown in Table 9.9 (0, 20, and 40

FIGURE 9.2
PLOT OF AVERAGE PERCENTAGE OF UNDESIRABLE CRITERION SCORES, SIX LEVELS OF THE MENTAL HEALTH (MH) SCALE



on the 100-point scale). Very large percentages of poor mental health status and outcome were observed across the seven criteria at the lowest scale level (from 39.3% for inpatient mental health care to 78.8% for serious depressive symptoms). These rates drop substantially as scale scores increase, but still remain high across the lowest three scale levels.

These rates should be useful in interpreting differences in MH scores and in explaining their clinical and social relevance. For example, a difference between 20 and 40 on the MH scale represents: (1) a 13 percentage point difference in dissatisfaction with life, (2) a 15 percentage point difference in the probability of a clinical diagnosis of depression, and (3) more than a 10 percentage point difference in rates of inpatient mental health treatment.

It is sometimes argued that differences in the psychological well-being range (at the top) of the MH scale are not clinically and socially relevant. For example, given that the mean for the MH scale is approximately 70, some would ignore differences in the 80 to 100 range. For all criteria presented in Table 9.9, substantial differences in outcomes are apparent for those scoring at these two levels.

Other interpretation guidelines for the MH scale have been published including the often quoted estimate of the social relevance of a three-point difference on the MH scale, which has been shown to be roughly equivalent to the psychological distress caused by being fired or laid off from one's job

(Brook et al., 1983; Ware et al., 1986). This estimate was based on longitudinal studies of stressful life events and mental health outcomes in the HIE (Williams et al., 1981).

Studies of insurance claims during the HIE linked MHI scale scores to the use of mental health services (Ware et al., 1984). These estimates were based on 12,435 person-years of data and utilization and other expenditure estimates from insurance claims forms. Table 9.10 presents estimates of utilization of mental health services for those scoring in the top, middle, and bottom thirds of the MHI score distribution. Substantially higher rates of utilization across all three indicators are apparent for those scoring in the bottom third of the MHI distribution compared with the top third. For example, more than a tripling in total mental health expenditures (from $10 to $34 per person per year) was observed for those who began the year at the top third versus the bottom third of the mental health distribution. Although these estimates were based on the 38-item MHI, the results can be generalized to the SF-36 MH scale with confidence because the latter includes the MH score, which also correlates 0.97 with the MHI.

These results support the assumptions underlying the "equal interval" scoring of the MH scale and provide useful guidelines for interpreting scale scores. Results for specific criteria at specific scale levels can be used in explaining the importance of differences in scale scores. For example, a difference of 20 points between 80 and 60 represents nearly a tripling (from 4.5% to 12.4%) in the percentage of people who report being dissatisfied with their life. Differences from 80 to 60 go hand-in-hand with a tripling in the probability of suicidal ideation (from 1.9% to 5.6%). A difference from 80 to 60 represents more than a doubling of the probability of a diagnosis of depression (from 7.6% to 18.4%). The latter rates increase much more dramatically in absolute terms at the lower levels of the scale.

Self-Reported Transitions

The "36th item" in the SF-36, actually Item 2, is not used to score any of the eight scales. This item was included to estimate change in health status from a cross-sectional administration of the SF-36. It is based on the hypothesis that self-reported transitions reflect true changes in health during the recall period. This hypothesis is being tested in the MOS; preliminary results for 1,698 patients measured twice, a year apart, are reported here.

TABLE 9.10 MENTAL HEALTH AND USE OF MENTAL HEALTH SERVICES

Mental Health Inventory Tertiles[a]	Mean[b] Score	Percent with Any Use of Services	Percent Who Saw a Specialist[c]	Total Mental Health Expenditures per Year[d]
Top Third	86	4.4%	2.4%	$10
Middle Third	76	5.4%	3.0%	$13
Bottom Third	61	11.4%	6.1%	$34

[a] Based on the 38-item MHI, which correlated 0.97 with the SF-36 MH scale.

[b] Mean scale score for each MHI tertile, shown to be equivalent to MH scale tertiles.

[c] Formally trained mental health specialist.

[d] Average expenditures estimated from insurance claim forms; 12,435 person-years of data, 1981 dollars.

Note. From "Health status and the use of outpatient mental health services" by J.E. Ware et al., 1984, *American Psychologist, 39,* 1090-1100.

As Table 9.11 shows, 976 of the 1,698 MOS patients (57.5%), all with one or more chronic conditions, reported that their health was "about the same" as it was a year earlier. An improvement during the 1-year follow-up period was reported by more respondents (17% + 11% = 28%) than a worsening (12.8% + 1.7% = 14.5%). Of course, these analyses were limited to survivors. If those who died had been added to the latter category, that percentage would increase.

The results in the "average Δ" column of Table 9.11 suggest that measured changes in general health ratings during a 1-year follow-up period tend to correspond, at least on average, to transitions reported in response to SF-36 Item 2. Thus, considerable useful information about actual change in health status can be gleaned from this item. The 13-item form of the GHRI scale (Stewart, Hays, & Ware, 1992) was used in these analyses. For example, those who reported that their health was "much better" after 1 year improved an average of 13.2 points on the 0 to 100 GHRI scale. Average changes for those who reported that they were "somewhat better" or "about the same" were 5.8 and 1.6, respectively. On average, measured changes for those who reported that their health was "worse" tended to be negative, as hypothesized; -10.8 for those in the "somewhat worse" category and -34.4 for those in the "much worse" category (note that the latter category included only 29 patients).

TABLE 9.11 MEASURED CHANGES IN GENERAL HEALTH FOR PATIENTS IN FIVE CATEGORIES OF SELF-REPORTED HEALTH TRANSITIONS (N=1,698)

Reported Transition	f	%	Average Δ[a]	Percent Improved[b]	Percent Worsened[b]
Much better	187	11.0	13.2	47.6	6.9
Somewhat better	288	17.0	5.8	35.1	9.4
About the same	976	57.5	1.6	20.4	14.6
Somewhat worse	218	12.8	-10.8	6.9	40.4
Much worse	29	1.7	-34.4	3.4	69.0

[a] Δ = General Health Rating Index (GHRI) score at follow-up minus GHRI score 1 year earlier.

[b] Improvements and worsenings based on change of two standard errors of measurement or larger (i.e., less than 5% expected by chance).

Note. $F = 106.2$, $p < 0.001$, for mean differences in GHRI scores across the five self-report transition categories.

Patients were placed in one of three conservatively defined categories in terms of whether they actually "improved," "worsened," or "stayed the same," during the 1-year follow-up period. Those whose follow-up GHRI scores were within two standard errors of measurement of their initial scores were placed in the "no change" group; those who improved an amount greater than two standard errors were placed in the "improved" group and those whose follow-up scores were lower an amount greater than two standard errors were placed in the "worsened" group. As explained in Chapter 7, a patient would be expected to change two standard errors of measurement or more only one time in 20 by chance; thus, two standard errors of measurement defines the 95% confidence interval around the score for each study participant.

Results from preliminary MOS analyses support the accuracy of self-reported transitions. Approximately 48% of those who reported that they were "much better" were categorized as improved, whereas only 6.9% were categorized as worsened. As shown in the bottom row of Table 9.11, 69% of those who reported that their health was "much worse" were categorized as worsened whereas only 3.4% were classified as improved.

Finally, Table 9.11 reveals important information useful in interpreting reports that health is "about the same" compared to 1 year ago. It appears that for more than one-third of those who make such reports, significant

changes actually occurred. Approximately 20% in the "about the same" category improved more than two standard errors and approximately 15% worsened over the 1-year follow-up period. Thus, about one-third of those who report no change actually changed an amount that would be expected by chance less than 5% of the time.

Results similar to those above for the GHRI were confirmed when repeated measurements used to categorize change scores were based on the PF and MH scales. It is possible that some of the disagreements between reported and measured changes are due to variations in the pattern of health outcomes across concepts. Analyses of the effect of various permutations of changes in specific health scales on self-reported transitions are currently under way in the MOS.

Factor analysis of the SF-36

Factor analysis provides an empirical test of the construct validity of the SF-36 in relation to its hypothesized structure. Factor analysis also has a second purpose. In the absence of agreed upon "criteria" for validating a scale, the validity of each scale can be tested using factor analytic methods. We used this methodology in a recently published study that included both psychometric and clinical tests of validity (McHorney et al., 1993).

The SF-36 was constructed to represent two major dimensions of health — physical and mental — that have been confirmed empirically in previous studies (Hays & Stewart, 1990; Ware, Davies-Avery, & Brook, 1980). Thus, we extracted two principal components from the correlations among SF-36 health scales and rotated them to orthogonal simple structure as summarized in Table 9.12. 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 and mental health is low (0.17 between the PF and MH scales in the MOS), an orthogonal solution was expected to also reproduce well the matrix of correlations. In fact, the two-factor orthogonal solution accounted for 82.4% of the reliable variance.

The two components were interpreted on the basis of their correlations with

the SF-36 scales. Because the pattern of results across scales was very consistent with expectations for physical and mental "dimensions" of health, they were labeled "physical" and "mental" accordingly. If the two-dimensional structure had not been confirmed or the interpretation of the factors turned out to be ambiguous, these components could not have been used as "criteria" in testing the validity of each scale.

Table 9.12 shows the factor loadings for the SF-36 scales. The PF scale has a strong association with the physical component of health and a weak association with the mental component of health, as hypothesized. At the other extreme, the MH scale has a strong association with the mental health component and a weak association with the physical component. Table 9.12 also lists the factor content of the other scales. Hypotheses about the factor content of the SF-36 scales, which were based on previous research, were largely confirmed by these psychometric tests of construct validity.

TABLE 9.12 SCALE VALIDITY AND CORRELATIONS WITH ROTATED PRINCIPAL COMPONENTS

	Hypothesized Association		Rotated Principal Components		
	Physical	Mental	Physical[a]	Mental[a]	h^2
Physical Functioning	●	○	.88	.04	.78
Role–Physical	●	○	.78	.30	.70
Bodily Pain	●	○	.77	.24	.65
Mental Health	○	●	.12	.90	.82
Role–Emotional	○	●	.19	.83	.69
Social Functioning	◑	●	.44	.71	.70
Vitality	◑	◑	.59	.57	.67
General Health Perceptions	◑	◑	.48	.32	.56

● Strong association ($r \geq .70$).

◑ Moderate to substantial association ($.30 < r < .70$).

○ Weak association ($r \leq .30$).

h^2 Proportion of total variance of each scale explained by the two extracted components.

[a] Correlation between each scale and rotated principal component.

Note. From "The MOS 36-Item Short-Form Health Survey (SF-36): II: Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs" by C.A. McHorney et al., *Medical Care*, 31, 247-263.

With the exception of the GH scale, a large proportion of the reliable variance in each scale was accounted for by these two rotated principal components. For the GH scale, only about 56% of the reliable variance was explained. Thus, much of the reliable variance in general health perceptions is due to things other than physical and mental health as defined by these two components. In fact, for all of the SF-36 scales, some reliable variance was unique to each scale and not explained by the two components.

The results shown in Table 9.12 constitute strong evidence for the conceptualization of health underlying the construction of the SF-36 and provide "psychometric" results useful in the interpretation of each scale. These results clearly indicate that some scales principally measure the physical dimension of health (PF, RP, and BP); others principally measure the mental component (MH, RE, and SF). Others (VT and GH) appear to measure both.

The results summarized in Table 9.12 greatly influenced the formulation of guidelines for the interpretation of the eight SF-36 scales. However, as discussed in the following section, it was the replication of these results across clinical criteria used to define physical morbidity versus mental morbidity that contributed most to our confidence in these interpretation guidelines, as they apply to clinical practice and research.

Clinical Tests of Validity

As documented in greater detail elsewhere (McHorney et al., 1992; 1993), clinical tests of validity were based on criteria used to form mutually exclusive patient groups. These groups differed in the severity of their conditions as defined by clinical measures of physical and mental (psychological) morbidity. The least severe comparison group was limited to patients with only a minor medical condition such as uncomplicated hypertension (N=638). The group with serious physical morbidity included patients with congestive heart failure (CHF) and complications (e.g., edema, orthopnea); myocardial infarction survivors with substantial morbidity (e.g., noteworthy and recurring angina and/or severe CHF symptomatology); hypertension patients with a history of a stroke; and diabetic patients with noteworthy complications (e.g., severe autonomic neuropathy). The group used to test validity in relation to clinical criteria of mental health was limited to patients with severe mental morbidity such as current unipolar affective disorder (major depression or dysthymia) or serious depressive symptoms.

Effect Sizes and Relative Validity

Table 9.13 compares results based on tests of factorial validity with results based on clinical "criteria." Mean differences in SF-36 scale scores (labeled Δ in the first column under each criterion) are for comparisons between the least severe group and the groups with either severe physical or mental morbidity. (The group with both is not analyzed here.) The second column under each clinical criterion presents each group difference expressed as a standardized effect size (ES), which is the group difference divided by the general population standard deviation (SD). The third column under each criterion reproduces estimates of relative validity (RV) (McHorney et al., 1993). RV is the ratio of pair-wise F-statistics, specifically the F for the comparison scale divided by the F for the most valid scale based on the same two-group comparison. The F-ratios analyzed in estimating RV are those for the difference between group means relative to the within-group (error) variance. The F-ratio is larger when the separation between group means is larger and the error term is small. Thus, a larger F reflects a greater discriminant validity and/or greater precision in estimating group means. RV estimates indicate how valid each scale is in discriminating between clinical groups, relative to the most valid SF-36 scale. RV is useful in addressing the issue of "conceptual relevance" and answers the question: how sensitive is each SF-36 health concept to differences in the levels of physical and mental morbidity defined by these clinical groups, relative to the best scale?

All eight SF-36 scales discriminated between groups differing in physical morbidity ($p < 0.01$), but not equally well. The best performing scale in the physical tests was the PF scale, which was the worst scale in the mental tests. This pattern of results was reported above in the psychometric tests. The best scale in the mental and worst in the physical health test was the MH scale, again following the pattern observed in the psychometric tests.

The PF and MH scales are relatively pure in terms of both their factor content and their validity in relation to these clinical criteria. Thus, the PF and MH scales can be interpreted as measures of physical and mental dimensions of health, respectively. Why include the other SF-36 scales? As discussed elsewhere (see Chapter 3 and McHorney et al., 1993), comprehensive health status assessment should represent more concepts than those measured by just the PF and MH scales. Understanding differences in health burden across these clinical groups requires broad assessment of limitations in the performance or capacity to perform everyday social and role activities, energy and fatigue, bodily pain, and general health perceptions.

TABLE 9.13 COMPARISON OF FACTORIAL VALIDITY AND TESTS BASED ON CLINICAL CRITERIA

	Factorial Validity[a]		Clinical Criteria					
			Physical			Mental		
Scales	Physical	Mental	Δ[b]	ES[c]	RV[d]	Δ[e]	ES[c]	RV[d]
Physical Functioning	●	○	-23.2*	1.00	1.00	0.1	.00	.00
Role–Physical	●	○	-26.4*	.76	.71	-14.7*	.43	.07
Bodily Pain	●	○	-11.0*	.46	.27	-12.8*	.53	.14
Mental Health	○	●	-4.9*	.27	.15	-29.7*	1.65	1.00
Role–Emotional	○	●	-8.1*	.24	.07	-43.5*	1.32	.54
Social Functioning	◑	●	-11.6*	.50	.35	-27.1*	1.19	.54
Vitality	◑	◑	-14.2*	.68	.67	-16.7*	.79	.29
General Health Perceptions	◑	◑	-17.9*	.89	.99	-9.1*	.46	.08

[a] ● = Strong association ($r \geq .70$), ◑ = moderate to substantial association ($.30 < r < .70$), and ○ = weak association ($r \leq .30$).

[b] Differences between minor medical group and group with serious medical conditions.

[c] Effect size (ES) = Δ/SD; where SD comes from the general population.

[d] RV = relative validity (see text).

[e] Difference between minor medical group and group with serious psychiatric conditions.

* $p < 0.001$.

The ES column estimates "clinical" validity in units of the score variance, actually SD units. This column calls attention to the fact that RV and ES estimates may be, in some instances, quite different for reasons that should be considered when designing a study and when choosing among health concepts and measures.

Discrepancies in results between ES and RV estimates of validity are most noteworthy for comparisons between clinical groups differing in mental health because the F-ratio for the MH scale proved to be nearly 50% better than the next best scale. RV estimates, which define the validity of the other scales relative to the MH scale, tend to be low (0.00 to 0.29) for five of the eight scales. However, the discriminant validity of only the PF scale was below .40 when group differences were estimated in SD units using the ES index. Thus, seven of the eight scales (all but PF) appear to be sensitive to clinically defined differences in mental health.

These and other results and their implications for the interpretation of SF-36 scales are discussed in much greater detail elsewhere; readers are encouraged to review the source articles (McHorney et al., 1992; 1993). Reasons for differences in results across psychometric and clinical tests of validity are also discussed. For example, the GH scale is known to be sensitive to differences in physical symptoms not represented well in either the factor analytic or clinical tests reported here (see Chapter 9). The validity of SF-36 scales in discriminating between groups of patients with coexisting physical and mental morbidities, a common but rarely studied situation, is discussed elsewhere (McHorney et al., 1993).

According to psychometric theory, all else being equal, scales that do best in the cross-sectional tests of "clinical validity" addressed in Tables 9.12 and 9.13 should also do best in tests of sensitivity in relation to clinical changes over time. Little evidence has been published on this important assumption, although findings to date suggest that the assumption may hold true (see Chapter 11 and Katz et al., 1992). Additional evidence from the MOS is forthcoming.

Health and quality of life

Is there a sound basis for interpreting SF-36 scales as measures of health and health-related quality of life? This important issue is being debated for the SF-36 and other general health measures. What is the relationship between SF-36 scales and general health and quality of life? To address these issues, we correlated SF-36 scales with the GH scale and with a general measure of quality of life (Table 9.14). All correlations were significant and positive, as hypothesized. Most were substantial in magnitude. All correlations between the GH scale and the other SF-36 scales were substantial, ranging from 0.43 for the RE scale to a high of 0.69 for PF and RP scales.

Correlations between SF-36 scales and ratings of specific dimensions of quality of life beyond health were also examined, including correlations with evaluations of living arrangements (house, apartment), neighborhood, standard of living, financial situation, family life, and friendships. All of these correlations were based on analyses of the general population sample, were in a positive direction, and all were statistically significant, although few were

TABLE 9.14 ASSOCIATIONS BETWEEN SF-36 SCALES AND GENERAL HEALTH AND QUALITY OF LIFE MEASURES, GENERAL U.S. POPULATION (N=2,474)

	General Health			Quality of Life		
		Percent Excellent by Scale Level[a]			Percent Satisfied by Scale Level[b]	
Scales	r	Lowest[c]	Highest[d]	r	Lowest[c]	Highest[d]
Physical Functioning (PF)	.69*	4.1	35.6	.19*	34.4	57.0
Role–Physical (RP)	.69*	3.8	25.7	.23*	30.3	55.1
Bodily Pain (BP)	.58*	0.0	34.7	.29*	22.4	61.7
General Health (GH)	—[e]	—	—	.35*	14.5	67.5
Vitality (VT)	.65*	1.2	47.0	.45*	20.3	83.2
Social Functioning (SF)	.57*	4.6	26.7	.38*	12.3	61.8
Role–Emotional (RE)	.43*	8.5	23.3	.38*	18.9	58.1
Mental Health (MH)	.49*	6.1	30.3	.60*	4.9	78.3

[a] Percent "excellent" in response to evaluation of health (excellent to poor) SF-36 Item 1.

[b] Percent "extremely" or "very" "happy, satisfied or pleased" with personal life during past month.

[c] The bottom of each scale distribution is PF (0-25), RP (0), BP (0-22), GH (0-25), VT (0-15), SF (0-25), RE (0), and MH (0-36).

[d] The top of each scale distribution is a score of 100 with the exception of VT and MH (96-100).

[e] Scale used as "criterion".

* $p < 0.001$.

noteworthy in magnitude (above 0.30), and only one exceeded 0.40 (quality of family life and the MH scale). These results indicate a significant correlation between health status and aspects of quality of life not specifically related to health.

For the analysis reported in this section, we relied on the widely used quality of life rating from the General Psychological Well-Being measure (Dupuy, 1984), which was studied in the HIE (Ware et al., 1979). This measure asks each respondent "how happy, satisfied, or pleased have you been with your personal life?" Six response choices are offered ranging from 'extremely happy, could not have been more satisfied or pleased,' to 'very dissatisfied or unhappy most of the time.' This item has been shown (Ware et al., 1979) to be substantially associated with the Cantrill ladder (Cantrill, 1965).

Correlations between the SF-36 scales and quality of life "criterion" measure were positive and significant (Table 9.14). These correlations are all positive

and statistically significant in the general U.S. population, ranging from a low of 0.19 for the PF scale to a high of 0.60 for the MH scale, with a median of about 0.36.

Table 9.14 also shows the range of differences in the percentage at the lowest and highest levels of each SF-36 scale that rate their general health as "excellent" and that report being "satisfied" with their quality of life. Low frequency levels at the bottom of the scales were collapsed to achieve a sample size of greater than 100 for all comparisons. These entries are for the general U.S. population (N=2,474). Clearly the SF-36 scale levels are directly related to personal evaluations of health status and to satisfaction with everyday life. The results presented in Table 9.14 can be useful in interpreting scores and explaining their implications for general health and quality of life.

Symptoms

Table 9.15 presents correlations between the self-reported frequency of symptoms and scores for the eight SF-36 scales. Symptoms were reported for the 4-week period prior to the survey containing the SF-36. Symptoms are grouped into four categories. Symptoms correlating highly (0.30 or better) with the PF scale are in the first category. The second category includes those symptoms that correlate 0.30 or better with the MH scale. Symptoms correlating 0.30 or better with both PF and MH scales are included in the third category, and symptoms correlating less than 0.30 with both are in the fourth category. Substantial correlations (those equal to 0.40 or greater) are in boldface. All table entries are product-moment correlations.

There appear to be a number of overall patterns in the results shown in Table 9.15. The symptoms most likely to be underlying differences in PF scale scores include shortness of breath; stiffness and pain in muscles, and chest pain. Only two substantial correlations were observed among the three best MH scales, but some are noteworthy including feeling drowsy or sedated, dizzy when standing up, light-headedness, headaches more than usual, and waking up early or unable to sleep.

Other substantial correlations involving other SF-36 scales are noteworthy, including substantial correlations involving the BP scale and such symptoms

TABLE 9.15 CORRELATIONS BETWEEN SF-36 SCALES AND SELF-REPORTED FREQUENCY OF SYMPTOMS IN FOUR CATEGORIES

			SF-36 Scales							
Symptoms	Mean[a]	SD	PF[b]	RP	BP[c]	GH	VT	SF[d]	RE	MH
Shortness of breath (climbing stairs)	1.89	1.2	-.63	-.46	-.32	-.45	-.44	-.34	-.26	-.25
Stiffness, pain in muscles	2.89	1.4	-.44	-.40	-.59	-.37	-.34	-.28	-.18	-.20
Chest pains brought on by activity	1.48	0.9	-.41	-.35	-.31	-.38	-.32	-.29	-.24	-.22
Pins and needles in your feet	1.75	1.2	-.39	-.28	-.32	-.34	-.28	-.24	-.20	-.18
Dry mouth	2.16	1.3	-.38	-.35	-.31	-.38	-.36	-.34	-.28	-.29
Backaches or lower back pains	2.41	1.4	-.35	-.33	-.49	-.30	-.36	-.27	-.18	-.24
Blurred vision	1.53	1.0	-.32	-.29	-.27	-.32	-.31	-.28	-.27	-.25
Headaches more than usual	1.69	1.1	-.19	-.30	-.36	-.27	-.40	-.37	-.31	-.38
Dizzy when standing up	1.61	0.9	-.27	-.31	-.28	-.28	-.37	-.34	-.29	-.34
Acid indigestion after meals	2.20	1.2	-.24	-.24	-.31	-.23	-.30	-.22	-.23	-.30
Lightheaded while on feet	1.63	0.9	-.39	-.38	-.35	-.36	-.41	-.40	-.31	-.34
Heart pounding or palpitations	1.58	1.0	-.33	-.32	-.29	-.35	-.33	-.30	-.28	-.31
Drowsy or sedated	1.87	1.1	-.32	-.39	-.34	-.39	-.48	-.40	-.36	-.38
Waking up early, trouble to sleep	2.34	1.3	-.30	-.34	-.31	-.33	-.35	-.36	-.32	-.39
Coughing producing sputum	1.82	1.2	-.22	-.21	-.16	-.23	-.23	-.23	-.13	-.16
Fainting or passing out	1.03	0.2	-.10	-.09	-.10	-.07	-.12	-.09	-.08	-.10
Sudden weakness relieved by eating	1.39	0.8	-.28	-.31	-.25	-.30	-.33	-.29	-.26	-.28
Trouble passing urine	1.22	0.7	-.14	-.18	-.13	-.17	-.15	-.18	-.16	-.13
Urinating more than usual	1.78	1.2	-.27	-.27	-.19	-.27	-.26	-.22	-.19	-.15

[a] 1 = never, 2 = once or twice, 3 = a few times, 4 = fairly often, 5 = very often.

[b] Short form version (k=8) of SF-36 PF.

[c] Short form version (k=1) of SF-36 BP.

[d] Short form version (k=1) of SF-36 SF.

as stiffness or pain in muscles, and backaches or lower back pains. The VT scale appears most likely to be affected by symptoms of feeling drowsy or sedated, shortness of breath when climbing stairs, light-headedness while on feet, and headaches more than usual. A number of symptoms have little or no association with scale scores, including cough producing sputum, fainting or passing out, sudden weakness, trouble passing urine, and urinating more than usual.

The results presented in Table 9.15 are useful in speculating about the symptoms most likely to be underlying differences in scores for SF-36 scales. These results are also useful in selecting an SF-36 scale for studies involving particular symptoms. Finally, these results suggest that the generic SF-36 health scales are sensitive to differences in self-reports of the frequency of a wide range of different symptoms, including all 19 symptoms shown in Table 9.15. Because some correlations were based on even shorter versions of the scales than those included in the SF-36 (BP and SF scales), the estimates of correlations may be attenuated.

Correlations with other measures

Published correlations between SF-36 scales and other measures, particularly those of known validity, can be useful in testing hypotheses about their validity and interpretation. Table 9.16 summarizes associations published for 29 measures. Not surprisingly, given that the MHI-5 scale was made available in the early 1980s, more findings have been published for that scale other SF-36 scales.

There are a number of noteworthy trends in results for the MH scale. The strongest associations are with other mental health concepts, including the psychological dimension of the shortened Arthritis Impact and Measurement Scales (sAIMS) ($r = 0.82$), the psychological dimension of the SIP ($r = -0.70$), the mental health dimension of the Duke UNC Health Profile ($r = 0.51$), and the emotional reactions scale for the Nottingham Health Profile (NHP) ($r = -0.67$). (The SIP and NHP correlations are negative because a high score indicates a poorer health state.)

Other correlations involving the SF-36 MH scale are in the substantial range

TABLE 9.16 CORRELATIONS BETWEEN SF-36 SCALES AND OTHER HEALTH MEASURES

Scales	PF	RP	BP	GH	VT	SF	RE	MH	Reference
General Health Rating Index (GHRI)								.46	Read et al., 1987
Quality of Well-being Scale (QWB)								.33	Read et al., 1987
				.52					Fryback et al., 1993
Physical Performance Test (PPT)									Reuben & Siu, 1990
Nine-Item Score								.24	
Seven-Item Score								.32	
Mini-Mental State Examination (MMSE)								.20	Reuben & Siu, 1990
Bosow-Breslow Scale								.41	Reuben & Siu, 1990
Katz Activities of Daily Living (ADL) Scale								.10	Reuben & Siu, 1990
Spector Scale								.13	Reuben & Siu, 1990
Sickness Impact Profile (SIP)[a]								-.57	Read et al., 1987
Physical Functioning Scale	-.78								Weinberger et al., 1991
	-.67								Katz et al., 1992
Social Functioning Scale						-.67			Weinberger et al., 1991
Psychological Dimension								-.70	Katz et al., 1992
Home Management Scale								-.46	Cleary et al., 1991
Work Functioning Scale								-.42	Cleary et al., 1991
Recreation Scale								-.51	Cleary et al., 1991
Duke Health Profile									Parkerson et al., 1991
Mental Health Dimension								.51	
Total Knee Replacement SF-36 Battery									Kantz et al., 1992
Physical Functioning: Knee	.85	.29	.34	.32	.53	.38	.38	.21	
Role Physical: Knee	.29	.36	.43	.18	.29	.50	.34	.34	
Pain: Knee	.29	.23	.41	.69	.51	.46	.41	.49	
Condition Specific Knee Society									
Function	.16	.12	.27	.34	.35	.32	.14	.24	
Pain	.65	.63	.41	.49	.51	.48	.41	.49	

continued on next page

TABLE 9.16 CONTINUED

Scales	PF	RP	BP	GH	VT	SF	RE	MH	Reference
Nottingham Health Profile (NHP)[a]									Brazier et al., 1992
Physical Morbidity	-.52		-.45		-.36	-.35		-.19	
Social Isolation	-.20		-.18		-.36	-.41		-.47	
Pain	-.47		-.35		-.33	-.35		-.21	
Emotional Reactions	-.18		-.28		-.55	-.53		-.67	
Energy	-.37		-.37		-.68	-.51		-.47	
Modified Health Assessment Questionnaire (MHAQ)	.50								Katz et al., 1992
Shortened Arthritis Impact Measurement Scales (sAIMS)									Katz et al., 1992
Physical Dimension	.60								
Psychological Dimension								.82	
Functional Status Questionnaire (FSQ)	.73								Katz et al., 1992

[a]Correlation coefficients are negative because the two scales run in opposite directions.

(above 0.40, absolute magnitude). Among them are correlations with measures of conceptually related concepts, including general health ratings, limitations in recreational activities, social isolation, and overall energy level.

Inspection of results in Table 9.16 for correlations that should be low can also be revealing. Among the published associations involving the MH scale that indicate weak associations (0.30 and less) are those with measures of activities of daily living, cognitive function, physical measures of performance, and physical morbidity (various clinical and generic measures).

Strong associations have been reported between the SF-36 PF scale and measures of physical functioning, including the SIP physical functioning scale ($r = -0.67$ to -0.78), a knee-specific adaptation of the PF scale ($r = 0.85$), the NHP physical morbidity scale ($r = -0.52$), the physical dimension of the sAIMS ($r = 0.60$), and the Functional Status Questionnaire ($r = 0.73$). Correlations involving the PF scale that should be lower, given the concepts and operational definitions involved, are apparent in Table 9.16, including low correlations with condition-specific measures of functioning, measures of emotional reactions, and social isolation.

For the other SF-36 scales, available information of this kind is very limited. Kantz and her colleagues (1992) published one orthopedics study, and Brazier and his colleagues in the United Kingdom published a study of the NHP (1992). Because there is some degree of match between the concepts being measured across these two methodological comparisons, their results can be inspected in a "convergent-discriminant" validation paradigm. The expectation is that the most conceptually related scales would be more highly correlated than those measuring different concepts. There is some support for this. For example, the highest correlation between an NHP scale and the SF-36 MH scale was observed for the NHP emotional reactions measure ($r = -0.67$). The strongest correlation involving the SF-36 VT scale was observed for the NHP energy scale ($r = -0.68$). As these investigators have noted in their publications, there are some exceptions to this pattern of results that may reflect on the validity of one or both of the methods being compared.

Overall, the associations summarized in Table 9.16 appear consistent with the SF-36 interpretation guidelines (see Chapter 8). However, it should be noted that there are many gaps in our understanding of associations between SF-36 scales and external criteria including, but not limited to, other measures of health status. As these results become more available, they should be evaluated in light of the summary of interpretation guidelines (see Chapter 8) and these guidelines should be adjusted as warranted.

Correlations with other MOS scales

Table 9.17 presents correlations between SF-36 scales and 27 other MOS measures. These measures are grouped into nine different categories, and the labels used to identify them, as well as their construction and scoring are documented elsewhere (Stewart & Ware, 1992, specifically in Chapter 20 [Tables 20-2 and 20-3, pp. 350-360]). It is important to keep in mind when interpreting this table that in many instances, the short-form SF-36 scale is included in the long-form MOS measure. The resulting correlations, which are inflated, are labeled in the table. These correlations indicate how well the short-form SF-36 scale reproduces the long-form measure. These correlations tend to be very high, for RP (r=0.96), SF (r=0.90), BP (r=0.93), GH (r=0.90), and MH (r=0.96). The scoring of the MOS scales is indicated in parentheses after each scale name. As would be expected, correlations between

TABLE 9.17 CORRELATIONS BETWEEN SF-36 SCALES AND MOS FUNCTIONING AND WELL-BEING MEASURES

MOS Measures	PF	RP	BP	GH	VT	SF	RE	MH
Physical Functioning								
Satisfaction w/Physical Ability (+)[a]	.62	.58	.56	.59	.60	.50	.38	.41
Mobility (+)	.58	.44	.46	.43	.41	.50	.30	.27
Role Functioning								
Role Limitations due to Physical Health (-)	-.65	-.96[b]	-.65	-.51	-.60	-.55	-.52	-.36
Unable to do Housework due to Health (-)	-.48	-.44	-.44	-.40	-.43	-.49	-.33	-.33
Social, Family, Sexual Functioning								
Social Activity Limitations due to Health (-)	.46	.52	.51	.51	.59	.90[b]	.55	.65
Sexual Problems (-)	-.14	-.21	-.19	-.24	-.24	-.25	-.26	-.29
Satisfaction with Family Life (+)	.06	.18	.17	.24	.29	.35	.34	.50
Psychological Distress/Well-Being								
Anxiety (-)	-.20	-.32	-.38	-.39	-.51	-.60	-.56	-.83[b]
Depression/Behavioral-Emotional Control (-)	-.18	-.34	-.36	-.39	-.55	-.68	-.64	-.92[b]
Positive Affect (+)	.21	.36	.36	.42	.61	.62	.56	.90[b]
Feelings of Belonging (+)	.08	.14	.12	.16	.24	.31	.26	.40
Psychological Distress (-)	-.20	-.35	-.39	-.40	-.56	-.67	-.63	-.94[b]
Psychological Well-Being (+)	.18	.33	.33	.39	.57	.60	.55	.87[b]
Mental Health Index (+)	.20	.36	.39	.42	.60	.68	.63	.96[b]
Cognitive Functioning								
Cognitive Functioning (+)	.28	.38	.36	.38	.51	.59	.59	.70
Health Perceptions								
Current Health (+)	.58	.58	.57	.88[b]	.65	.58	.44	.50
Prior Health (+)	.29	.28	.30	.45	.26	.25	.15	.19
Health Outlook (+)	.34	.31	.28	.69[b]	.37	.26	.21	.28
Health Concern (-)	-.21	-.19	-.19	-.29	-.17	-.17	-.16	-.13
Resistance to Illness (+)	.21	.25	.31	.57[b]	.35	.32	.26	.34
General Health Rating Index (+)	.51	.52	.52	.90[b]	.59	.53	.40	.48
Health Distress (-)	-.47	-.52	-.51	-.59	-.59	-.65	-.51	-.60
Sleep								
Sleep Problems Index (-)	-.35	-.44	-.47	-.47	-.60	-.56	-.46	-.61
Pain								
Effects of Pain (-)	-.55	-.63	-.88[b]	-.51	-.53	-.62	-.40	-.47
Pain Severity (-)	-.44	-.53	-.85[b]	-.43	-.44	-.43	-.28	-.34
Overall Pain Index (-)	-.55	-.63	-.93[b]	-.51	-.53	-.57	-.38	-.44
Physical/Psychophysiologic Symptoms								
Physical/Psychophysiologic Symptoms (-)	-.49	-.54	-.66	-.52	-.57	-.50	-.41	-.46

[a] (+) scale scores from low to high reflect positive status, (-) scale scores from low to high reflect negative status.

[b] Indicates measure includes SF-36 items (correlation is inflated).

unfavorably scored MOS scales and SF-36 scales, all of which are scored positively, are negative.

The entries in Table 9.17 can be very useful in judging the value of adding measures of other concepts. For example, based on substantial correlations (0.56 to 0.62), it is clear that the first five SF-36 scales, which are the best measures of physical health, well reflect overall satisfaction with physical ability. All of the SF-36 scales correlate substantially with the summary of eight physical and psychophysiologic symptoms measured in the MOS (from a high of -0.66 for the BP scale to a low of -0.41 for the RE scale). Thus, it appears that MOS measures reflect differences in the frequency of these symptoms.

These entries can also be useful in determining the extent to which concepts and measures that are not included in the SF-36 are nevertheless reflected in the scores for scales that are included. For example, substantial correlations between the MOS cognitive functioning scale and all four of the best mental health scales (MH, RE, SF, and VT) were observed. Also, variations in sleep as measured by the MOS sleep problems index correlate substantially with most scales and highly with the VT scale (r = -0.60) and the MH scale (r = -0.61). In contrast, correlations between the SF-36 scales and sexual functioning (problems) tended to be low (from only 0.29 to 0.14) suggesting that variations in sexual functioning are not well represented in the SF-36 scales. Thus, sexual functioning is a candidate for inclusion in a generic health battery designed to supplement the SF-36.

Large differences in scale scores

Many ongoing studies are examining the smallest differences in SF-36 scale scores that should be considered clinically and socially relevant. Guidelines for interpreting such differences must await the results of those studies. Chapter 7 documents the size of the smallest difference that can be interpreted with confidence for an individual patient for each SF-36 scale. More information is available about the size of large differences.

Very large differences in SF-36 scale scores have been reported from the MOS in comparisons between groups differing in the severity of medical

TABLE 9.18 SIZE OF THE LARGEST DIFFERENCE IN SF-36 SCALE SCORES REPORTED TO DATE FROM THE MOS

		Comparison[a]	Difference (0-100 Scale)	Difference[b] (SD Units)
Physical Functioning	PF	A	34.2	1.47
Role–Physical	RP	A	46.4	1.36
Bodily Pain	BP	A	25.8	1.09
General Health	GH	A	27.1	1.33
Vitality	VT	A	25.0	1.19
Social Functioning	SF	B	27.1	1.19
Role–Emotional	RE	B	43.5	1.32
Mental Health	MH	B	29.7	1.64

[a] Comparisons: A = patients with minor conditions versus serious medical and serious psychiatric conditions; B = patients with minor conditions versus serious psychiatric conditions (McHorney et al., 1993).

[b] Difference divided by general U.S. population standard deviation.

and psychiatric conditions (McHorney et al., 1993). Table 9.18 lists the size of the largest difference in each scale reported. Differences have been transformed into SD units to facilitate comparisons. For purposes of the latter transformation, we used SD estimates for the general U.S. population (see Chapter 10). We recommend use of a representative and consistent SD estimate for each scale so that such transformations can be compared across studies without variations due to both mean differences and SD units (Mosteller et al., 1989).

The largest differences to date for five of the eight scales were observed in comparisons between groups of relatively well patients (with only minor conditions) versus those with both serious medical and serious psychiatric conditions. These differences ranged from 1.09 SD units (BP scale) to 1.47 SD units (PF scale). For the remaining three scales (MH, RE, and SF), which have been shown to be the best measures of mental health, the largest differences have been observed in comparisons between relatively well patients versus those with serious psychiatric conditions. These differences, which range from 1.19 SD units (SF scale) to 1.64 SD units (MH scale), reflect substantial physical and mental morbidity.

Although these are very large differences in terms of SD units (Cohen, 1988),

even larger differences between groups would be expected, for example, in analyses of groups at even more severe levels of either physical or mental (emotional) morbidity. The differences reported above and any difference of 0.8 SD units or larger would be considered large according to conventional standards (Cohen, 1988).

10. VALIDITY: NORM-BASED INTERPRETATION

This chapter presents norms for SF-36 scales in the general U.S. population. The chapter is divided into seven sections. The first section explains norm-based interpretations and provides background information. The remaining six major sections present: (1) methods used to gather normative data, (2) norms for the general U.S. population, (3) norms for various medical conditions, (4) patient norms for the Mental Health (MH) scale, (5) dichotomous SF-36 limitations indicators, and (6) comparisons between Developmental and Standard SF-36 versions.

For ease in locating normative data tables at the end of this chapter, Figure 10.1 lists all tables with their page numbers.

Background

Normative data make it possible to interpret the scale score for an individual respondent or the average score for a group of respondents by comparison with scores for other individuals. This comparison is based on where they are in the distribution of scores in the norming sample. For example, it is useful to know that a change from a score of 70 to 80 on the SF-36 General Health (GH) scale is a change from the 38th percentile to the 60th percentile in the general U.S. population. The 38th percentile is the score that 38% of the population score at or below.

Norm-based comparisons require valid norms for a well-defined and representative sample of the population of interest. This chapter presents such data for the SF-36. Some view the publication of such norms as more than just a convenience. Accepted guidelines for the standardization, scoring, and the documentation of widely used psychological tests emphasize the publication of norms prior to their widespread use (APA, 1985).

Norm-based interpretation has only rarely been available as a methodology for interpreting even the most widely used health status measures. Noteworthy exceptions include some single-item health measures that have been normed

FIGURE 10.1
GUIDE TO TABLES PRESENTING NORMATIVE DATA

in the National Health Interview Survey (NCHS, 1991). This survey, however, is not comprehensive, and single-item measures are not likely to be satisfactory in assessing health for many purposes (McHorney et al., 1992). Average scores for Sickness Impact Profile (SIP) scales have been reported for health maintenance organization (HMO) enrollees and for patients with selected conditions (Bergner et al., 1981).

General population norms have provided a very useful basis for interpreting scores for many widely used tests. Likewise, it is clear from published examples that the interpretation of the SF-36 profile for patients with a particular condition is advanced by comparisons with normative profiles for the general population (Brazier et al., 1992; Jenkinson et al., 1993; Phillips & Lansky, 1992). Forthcoming results from the Medical Outcomes Study (MOS) will also demonstrate these advantages for both cross-sectional and longitudinal studies.

U.S. general population norms were gathered by Louis Harris and Associates (1989) for 18 of the items in the MOS 20-Item Short-Form Health Survey (SF-20). These norms are published in detail in Ware et al. (1992) and proved useful in interpreting profiles for patients with specific conditions in a number of MOS studies (Stewart et al., 1988, 1989; Wells et al., 1989).

Norms for the British adaptation of the SF-36 (U.K. developmental version) are documented in two studies fielded in Great Britain (Brazier et al., 1992; Jenkinson et al., 1993). There appears to be a high degree of correspondence between average U.K. and U.S. scores for males and females for most SF-36 scales in the younger age groups (34 and younger) (see Chapter 11). However, noteworthy differences favoring the United Kingdom are apparent for most scales for those in the older age groups (for example, 55 to 64 years). U.K. norms have not been published for those 65 years and older.

Finally, normative data for SF-36 scales (Developmental version) were made available by InterStudy (Wetzler & Radosevich, 1992). These estimates of mean scores do not appear to agree with either U.S. norms reported here or with U.K. norms (Brazier et al., 1992; Jenkinson et al., 1993). Discrepancies are particularly large for older age groups. One possible explanation may be InterStudy's reliance on samples drawn from a single health care plan and elderly households from rural areas (Wetzler & Radosevich, 1992).

This chapter presents norms for a representative sample (N=2,474) of the non-institutionalized general U.S. population, ages 18 to 94, gathered from October through December 1990. Because health status scores for some concepts differ significantly across age groups and for men and women, norms are presented for the total population and separately for seven age groupings and for males and females.

How the SF-36 was normed

The following section discusses methods used to norm the SF-36. Users of these norms are encouraged to review the sampling and data collection methods so that they will understand their scientific basis as well as methodological details that might affect the use and interpretation of the norms. For example, the mode of SF-36 administration (personal interview versus self-administered questionnaire) may affect norms for some SF-36 scales (McHorney, Kosinski, & Ware, in review).

U.S. general population norms were estimated from responses to the National Survey of Functional Health Status (NSFHS), a 1990 cross-sectional survey that included the SF-36. Respondents were drawn from the sample frames of the 1989 and 1990 General Social Survey (GSS), conducted by the National Opinion Research Center (Thalji et al., 1991). The GSS has surveyed the non-institutionalized adult U.S. population annually over the last 20 years.

The GSS used a two-stage area probability sample to the block level. In the first stage, quota sampling was used with quotas based on age, sex, and employment status at the block level. The primary sampling units (PSUs) used were Standard Metropolitan Statistical Areas or non-metropolitan counties. These PSUs were stratified by region, age, and race before selection. The units of selection at the second stage were block groups stratified according to race and income.

The sample frame for the NSFHS was based on 1,537 GSS households from 1989 and 1,372 GSS households from 1990, for a total base sample of 2,909 households. From this base sample, two categories of respondents were drawn. First, a single member of each household who was previously interviewed for the 1989 or 1990 GSS was selected. The second category of

respondents consisted of 342 elderly persons (age 65 or older) who were residing in GSS households but were not selected for the GSS interviews (supplemental Medicare group). Altogether, the total designated sample for the NSFHS was 3,251 persons residing in 2,909 households.

Respondents were randomly assigned to a mail survey (80%) or a telephone survey (20%). The mail survey was administered in two waves, with a mail postcard prompt occurring between the two waves. A $2 prepaid incentive was provided to all mail respondents. Telephone interviewing was used to follow-up mail non-respondents using the same protocol as that utilized in the telephone survey.

The telephone survey was a computer-assisted telephone interview. Prior to the interview, a letter describing the purpose of the survey and selection of individuals was sent. No incentive was provided. Telephone interviewers were trained to avert refusals to maximize response rates. Personalized letters were sent to those respondents who refused the interview explaining the importance of the NSFHS. A telephone call followed the personalized letter in an attempt to complete the interview. If these telephone conversation attempts failed, a self-administered mail survey was sent to non-responders.

Data for the NSFHS were collected for 10 weeks in the mail survey and 8 weeks in the telephone survey between October 15, 1990, and December 22, 1990 (Thalji et al., 1991). Field activities were designed so that at least 50% of the data collection period for each survey mode overlapped. Locating activities were carried out by telephone. The locating protocol for both mail and telephone surveys followed three steps: (1) a call to directory assistance; (2) a check of returned envelopes for forwarding addresses; and (3) a review of the 1990 GSS case, locator page, and call record for possible locating leads. Further measures were taken on problem cases through calls to criscross directories. Extra locating activities were dropped, however, when the yield proved low and not cost effective. Given that social security numbers were not available for GSS respondents, outside locating vendors were not used.

At the end of the data collection period, 89% of GSS respondents had been located. The unlocated rate overall was 11%, 10% for respondents assigned to the mail survey and 12% for respondents assigned to the telephone survey. The overall response rate for the survey was 77.1%.

Many factors affect the usefulness of norms, including the representativeness of the normative sample. These issues have been addressed elsewhere (Thalji et al., 1991). For example, the sociodemographic characteristics (age, sex, etc.) of respondents in the NSFHS were compared with those of the general U.S. population published by the National Center for Health Statistics. Results support the representativeness of the norming sample for the SF-36 reported here (Thalji et al., 1991).

Finally, note that the norms reported here were estimated for a "mixed-mode" survey much like that often used when self-administration is the primary mode and telephone interviews are utilized for backup. Slightly more than two-thirds of respondents (1,692/2,474 = 68.4%) self-administered the questionnaire containing the SF-36 and received and returned it by mail. About 18% of those assigned to this method completed the survey by telephone personal interview. The additional sample assigned to the telephone method has been analyzed to estimate the effects of mode of administration on responses, data quality, and costs (McHorney, Kosinski, & Ware, in review). The norms reported in this chapter are appropriate for self-administered surveys, which use mixed-mode follow-up. The additional telephone administrations, added for research purposes, are likely to have little effect (less than 1 to 2 points) on average scores because the great majority of respondents self-administered the SF-36 and because, consistent with previous studies, the effects of mode of administration do not appear to be large (McHorney, Kosinski, & Ware, in review).

U.S. norms for SF-36 scales

Table 10.1 presents descriptive statistics for each SF-36 scale in the general U.S. population (males and females combined and separately). These include the mean, median (50th percentile), 25th and 75th percentiles, standard deviation, observed range of scores, and the percentage scoring at the ceiling (highest possible score) and at the floor (lowest possible score) for each SF-36 scale. These descriptive statistics are first presented for the total sample (including males and females; N=2,474).

How can these normative tables be used in interpretation? To determine whether the score for an individual or the average (mean) for a particular

group is above or below the average for the U.S. general population, look at the first row of Table 10.1. For example, the mean Physical Functioning (PF) scale score in the general U.S. population is 84.15 (see the upper left-hand corner of Table 10.1). Scores above and below 84.15 are above and below average, respectively.

The median for the PF scale is 90.00 indicating that 50% of the population score at or below 90.00. The standard deviation of the PF scale is 23.28. This estimate is used in Chapters 7 and 9 to standardize scores for comparison. The entry for PF in the "Range" column from 0-100 indicates that all possible scores from 0-100 were observed; 38.79% received a perfect score (see "% Ceiling"); and less than 1% (only 0.84%) received the lowest possible score (see "% Floor"). Table 10.1 presents these statistics again for each SF-36 scale for the total population and separately for males and females.

Norms for Age Groups

Table 10.2 presents national SF-36 norms for seven different age groups for males and females combined. These age groupings were selected (1) to be large enough to satisfy minimum standards for precision; (2) to correspond with standard practices for defining age-specific groups; and (3) to correspond with age groupings used by others in reporting norms for the SF-36 (Brazier et al., 1992; Jenkinson et al., 1993) and that are forthcoming from other countries. Comparing results across age groups (Table 10.2) clearly shows that health status, particularly physical health, is related to age. For example, whereas the mean PF score for the total sample is 84.15, the mean for the 18 to 24 group is much higher (92.13) and the mean for the 75 and older group is much lower (53.20).

Norms for Males and Females

Tables 10.3 and 10.4 present SF-36 norms separately for males and females in the U.S. general population. These tables differ from previous tables in that the two older age groups have been collapsed into one group (age 65 and older) to maintain precision.

Patients with medical conditions

We have also estimated norms for SF-36 scales for patients with five tracer conditions studied in the MOS (hypertension, congestive heart failure, Type II diabetes, recent acute myocardial infarction, and clinical depression). These are the five "tracer" conditions selected for longitudinal follow-up (Stewart & Ware, 1992; Tarlov et al., 1989). Standardized forms were completed by physicians to determine these diagnoses. The prevalence of these conditions among the 18,762 screened patients varied across the five conditions: 30.2% had hypertension, 9.2% had diabetes, 3.2% had congestive heart failure, and 1.5% had recent acute myocardial infarction. We also screened 4,335 patients from the offices of formally trained mental health providers. Patients were screened for clinical depression using a two-stage process with the CES-D and telephone-administered diagnostic interview schedule (Wells et al., 1989).

These patients were first screened in February to October 1986 from the practices of 362 medical clinicians (including 194 general internists, 91 family practitioners, 40 cardiologists, 24 endocrinologists, and 13 nurse practitioners) and 161 mental health providers. Providers were practicing in prepaid group practice HMOs, multi-specialty groups, and solo and small single specialty group practices in Boston, Chicago, and Los Angeles. These clinicians averaged 40.1 years of age and 21% were female. The representativeness of this sample is discussed in detail elsewhere (Stewart & Ware, 1992).

SF-36 scales were included in a 250-item questionnaire self-administered and returned by mail approximately 4 months, on average, after participants were screened in their doctors' office. It is important to keep in mind that SF-36 questionnaires administered at the time of a doctor visit are likely to be lower than periodic surveys not keyed to specific visits. For example, those in the MOS hypertension group scored approximately 3.2 points ($\pm$ 0.7, $p < 0.001$) higher on the 0 to 100 GH scale on the survey reported here relative to a survey completed 4 months earlier during a doctor visit. Average improvements in SF-36 profiles are even larger for those respondents with acute conditions at the time of a doctor visit. These influences should be considered when interpreting the reported data. Again, scores would be expected to be lower, on average, than the norms reported in this chapter, for patients at the time of a patient-initiated doctor visit.

The following tables describe the characteristics of patients in each condition group. Overall, they ranged in age from 18 to 97 years, with a mean age just under 58 years. A slight majority were female (53%), and about one in five was nonwhite. Approximately one in five were at or below the poverty level, and approximately three-fourths had completed at least a 12th-grade education.

Tables 10.5 through 10.9 present norms for each of these five samples in the same format used previously to present data for the general U.S. population. In addition, however, we also report sample descriptions for each of these groups including sociodemographic characteristics and the five most prevalent comorbid conditions observed for these groups. Because the norms in Tables 10.5 through 10.9 make no adjustments for differences in sociodemographic characteristics and other comorbidities, which have been shown to affect these scale scores, it is important to keep these confounding variables in mind. These norms are presented for clinicians and clinical investigators to use in interpreting scores for patients as they are seen in medical practice.

For example, in interpreting norms for those with current hypertension, note that one-third also suffered from back pain/sciatica and that one-fourth had musculoskeletal complaints. These comorbidities impact the mean scores for the PF, Role–Physical (RP), Bodily Pain (BP), GH, and Vitality (VT) scales reported in Table 10.5. Forthcoming reports from the MOS will includes estimates of the unique contributions of different tracer conditions, comorbid conditions, and the differences in scale scores explained by sociodemographic characteristics on SF-36 scale scores.

Patients with Hypertension and Comorbid Conditions

Tables 10.10 through 10.17 present norms for MOS patients with hypertension and each of eight comorbid conditions. These conditions, which are defined in the footnote for each table, include: chronic obstructive pulmonary disease (COPD) (N=85), angina (recent with no history of MI) (N=256), back pain/sciatica (N=481), osteoarthritis (N=175), musculoskeletal complaints (N=341), benign prostatic hypertrophy symptoms (N=184), varicosities (N=222), and dermatitis (N=231).

Because only patients with one or more of the five tracer conditions were

followed longitudinally in the MOS, all of the patients described in Tables 10.10 through 10.17 have at least two conditions. To estimate the effects of these comorbid conditions, we focused on the hypertension tracer group. We selected patients with hypertension to estimate health scores for these comorbidities, because hypertension was the largest tracer group studied and because hypertension was the least morbid of the five tracer conditions studied (Stewart et al., 1989). For these reasons, the estimates in Tables 10.10 through 10.17 involve the largest samples available and are more likely to represent the effects of the comorbid conditions than the comorbid hypertension.

For the same reasons discussed above with regard to the results for the five MOS tracer conditions, it is very important to note the sociodemographic characteristics of patients with hypertension and these eight comorbid conditions and also the prevalence of other comorbid conditions that have been shown in previous studies to affect scale scores. For example, patients with hypertension and COPD had an average age of 62.4 years, and 55% also suffered from back pain/sciatica (sample description in 2nd panel of Table 10.10). These factors and other potential confounding factors should be noted when interpreting the results reported in Tables 10.10 through 10.17.

Patient norms for the Mental Health scale

In response to numerous requests from those using the SF-36 MH scale to assess the emotional burden of patients and to screen for the probability of a disorder, we also prepared "patient norms" for the MH scale. We estimated these norms for the 9,385 patients who completed the MH scale at the time of a visit to one of the 523 clinician who participated in the MOS. Sampling methods, practices, physicians and other clinicians, and patients are detailed elsewhere (Stewart et al., 1989; Tarlov et al., 1989; Wells et al., 1989).

These patients were sampled from practices of non-psychiatrist physicians (cardiologists, endocrinologists, diabetologists, family physicians, and those trained in general internal medicine) as well as from the practices of formally trained mental health specialists. Because MH scores tend to be much lower at the time of a visit to a mental health provider (Wells et al., 1989), we estimated norms for the total "patient" sample, as well as separately for those

sampled from the practices of non-psychiatrist physicians (medical providers), and those sampled from the practices of formally trained mental health providers. For each of these three samples, we estimated norms for six age groups as well as separately for males and females across age groups. The format of these tables is the same as in previous tables.

As Table 10.18 shows, male patients score higher than female patients in all age groups in the total sample (for example, 77.20 versus 71.86, respectively, for the 18 to 24 age group). These results are consistent with published studies based on other mental health measures (Ware et al., 1979). Age differences in MH scale scores are also apparent; scores tend to be more favorable for those in the older age groups.

Table 10.19 presents MH-scale norms for male and female patients sampled from medical practices excluding those of formally trained mental health providers. Table 10.20 presents these norms for patients sampled only from the offices of formally trained mental health providers (psychiatrists, clinical psychologists, etc.). As would be expected, patients sampled from the practices of non-psychiatrist physicians score much more favorably than patients sampled from the offices of mental health providers. For example, the mean for the 18 to 24 age group is 75.87 for medical patients compared with 51.27 for patients of mental health providers (a difference of more than one standard deviation). This pattern of results is consistent with previous MOS analyses (Wells et al., 1989) and with studies reported by others (Berwick et al., 1991; Weinstein et al., 1989).

National norms for the dichotomous limitations indicators

Users of SF-36 computer software, such as the RT-2000 processing system, have the advantage of a printout of four summary limitations measures. These *dichotomous* indicators identify patients with: (1) physical limitations, (2) emotional limitations, (3) role disability, and (4) an unfavorable personal evaluation of their health in general.

The *physical limitation* indicator identifies individuals reporting any physical limitation in response to the 10-item PF scale. An *emotional limitation* is counted whenever a patient scores at or below 52 on the 0 to 100 MH

scale. This cut-off is based on studies of the relationship between MH scores and clinical measures of the probability of any psychiatric disorder (Berwick et al., 1991; Weinstein et al., 1989). A *role disability* is counted when a respondent endorses any of the RP or Role–Emotional (RE) scale items. Finally, an *unfavorable personal evaluation* is counted when respondents choose the "fair" or "poor" category in response to the rating of health in general (SF-36, Item 1).

The first panel of Table 10.21 presents norms for the general U.S. population. Because the prevalence of these limitations varies considerably with age, norms are presented separately for each of seven age groups. For example, one or more physical limitations are reported by 38.1% of 18 to 24 year-olds and by 96% of those 75 and older (first row of Table 10.21). Emotional limitations (at 13.2%) and unfavorable personal evaluations of health (at 7.7%) were much less prevalent for the 18 to 24 year age group. A substantial proportion of those in both younger and older age groups reported one or more role-disability limitations (37.7% and 75%, respectively for 18 to 24 and 75+ age groups). Because the prevalence of these limitations also varies for males and females, the table presents national norms for six age groups for both males and females.

Comparison of Developmental and Standard SF-36 scoring

All norms presented in this manual are for the Standard version of the SF-36 (see Chapter 3) and for standard scoring algorithms (see Chapter 6), unless otherwise noted. Scoring algorithms for the Developmental version have been developed to maximize comparability with the Standard SF-36 (as recommended in Chapter 6). When these scoring algorithms are used, the norms reported in this manual are appropriate for use in interpreting scores for the Developmental version.

The difference between the developmental and standard scoring algorithms for the SF-36 BP scale appears to have a substantial effect on average scores and increases the skewness of the score distribution (resulting in more scores in the upper range) for the developmental scoring method. The size of the shift may vary according to age and is certain to vary depending on the prevalence of pain in the population under study. The average shift appears to be

about 2 to 4 points, but may be smaller for the oldest age groups. Results from the United Kingdom for the Developmental version published by Brazier et al. (1992) and by Jenkinson et al. (1993) can be compared with norms and others results reported in this chapter for all but the BP scale. Results from the United Kingdom for the BP scale and those from any other studies that rely on the developmental scoring for the BP scale are not comparable to results presented in this chapter.

In summary, because of the nature of the difference in the standard and developmental algorithms, the amount of bias in BP scale scores will vary depending on the extent of pain in the population being studied. For this reason, we caution against comparisons between results for the BP scale scores across studies using developmental versus standard scoring. We encourage investigators to document which scoring method they use so that readers will know when they can and cannot compare their results.

TABLE 10.1 NORMS FOR THE GENERAL U.S. POPULATION, TOTAL SAMPLE

Total Sample (N=2,474)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	84.15	80.96	75.15	71.95	60.86	83.28	81.26	74.74
25th Percentile	70.00	50.00	61.00	57.00	45.00	75.00	66.67	64.00
50th Percentile *(median)*	90.00	100.00	74.00	72.00	65.00	100.00	100.00	80.00
75th Percentile	100.00	100.00	100.00	85.00	75.00	100.00	100.00	88.00
Standard Deviation	23.28	34.00	23.69	20.34	20.96	22.69	33.04	18.05
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	0-100
% Ceiling	38.79	70.85	31.85	7.40	1.50	52.32	71.81	3.91
% Floor	0.84	10.33	0.58	0.00	0.52	0.64	9.61	0.00

Males (N=1,055)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	87.18	86.61	76.88	73.48	63.59	85.23	83.28	76.37
25th Percentile	80.00	75.00	62.00	62.00	50.00	75.00	66.67	68.00
50th Percentile *(median)*	95.00	100.00	84.00	75.00	65.00	100.00	100.00	80.00
75th Percentile	100.00	100.00	100.00	87.00	80.00	100.00	100.00	88.00
Standard Deviation	21.29	30.88	22.97	20.02	20.04	21.28	31.31	17.16
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	12-100
% Ceiling	45.18	95.10	34.26	8.79	2.25	55.00	73.44	4.75
% Floor	0.67	7.92	0.32	0.00	0.23	0.35	7.96	0.00

Females (N=1,412)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	81.47	77.77	73.59	70.61	58.43	81.54	79.47	73.25
25th Percentile	65.00	50.00	52.00	57.00	45.00	62.5	66.67	64.00
50th Percentile *(median)*	90.00	100.00	74.00	72.00	60.00	87.5	100.00	80.00
75th Percentile	100.00	100.00	100.00	85.00	75.00	100.00	100.00	88.00
Standard Deviation	24.60	36.20	24.25	21.50	21.47	23.74	34.43	18.68
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	0-100
% Ceiling	33.05	67.14	29.64	6.13	0.82	49.21	68.85	3.12
% Floor	0.99	12.37	0.81	0.00	0.74	0.86	11.06	0.06

TABLE 10.2 NATIONAL NORMS FOR SEVEN AGE GROUPS, MALES AND FEMALES COMBINED

Ages 18–24
Males & Females
(N=173)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	92.13	89.14	80.82	76.71	62.53	83.88	83.00	74.73
25th Percentile	95.00	100.00	72.00	67.00	50.00	75.00	66.67	64.00
50th Percentile *(median)*	100.00	100.00	84.00	82.00	65.00	87.50	100.00	80.00
75th Percentile	100.00	100.00	100.00	87.00	75.00	100.00	100.00	88.00
Standard Deviation	18.34	26.81	21.35	18.22	19.76	20.64	31.12	18.09
Range	0-100	0-100	12-100	15-100	10-100	12.5-100	0-100	20-100
% Ceiling	61.9	82.0	45.5	8.7	2.2	48.0	72.0	2.9
% Floor	0.6	5.4	0.0	0.0	0.0	0.0	7.8	0.0

Ages 25-34
Males & Females
(N=474)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	92.00	89.28	81.35	77.09	61.34	84.86	82.24	73.29
25th Percentile	90.00	100.00	72.00	67.00	50.00	75.00	66.67	64.00
50th Percentile *(median)*	100.00	100.00	84.00	77.00	65.00	100.00	100.00	76.00
75th Percentile	100.00	100.00	100.00	87.00	75.00	100.00	100.00	88.00
Standard Deviation	15.78	24.88	19.72	17.33	20.20	20.96	31.51	17.95
Range	5-100	0-100	0-100	10-100	0-100	0-100	0-100	8-100
% Ceiling	57.3	79.8	39.9	9.5	0.5	51.9	70.8	2.0
% Floor	0.0	3.7	0.1	0.0	0.2	0.1	8.0	0.0

Ages 35-44
Males & Females
(N=503)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	89.70	86.66	77.06	75.87	62.42	85.75	82.76	75.13
25th Percentile	85.00	100.00	62.00	67.00	50.00	75.00	66.67	64.00
50th Percentile *(median)*	95.00	100.00	84.00	77.00	65.00	100.00	100.00	80.00
75th Percentile	100.00	100.00	100.00	90.00	75.00	100.00	100.00	82.00
Standard Deviation	16.35	28.92	22.11	17.86	19.45	21.04	31.26	16.69
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	20-100
% Ceiling	42.5	77.5	51.5	8.9	2.0	56.1	72.1	3.2
% Floor	0.1	6.7	0.2	0.0	0.1	0.6	7.6	0.0

TABLE 10.2 CONTINUED

Ages 45-54
Males & Females
(N=338)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	84.61	82.65	73.12	71.76	61.79	84.07	83.60	75.33
25th Percentile	80.00	75.00	61.00	62.00	50.00	75.00	100.00	68.00
50th Percentile *(median)*	95.00	100.00	74.00	77.00	65.00	100.00	100.00	80.00
75th Percentile	100.00	100.00	100.00	85.00	75.00	100.00	100.00	88.00
Standard Deviation	21.13	33.08	24.04	19.39	20.91	21.84	31.44	17.86
Range	0-100	0-100	0-100	10-100	0-100	0-100	0-100	8-100
% Ceiling	29.7	73.7	26.7	7.6	1.5	54.7	75.8	3.3
% Floor	0.5	9.5	0.8	0.0	0.6	0.5	7.0	0.0

Ages 55-64
Males & Females
(N=269)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	76.24	73.66	67.51	64.62	60.37	81.37	80.26	75.01
25th Percentile	60.00	50.00	51.00	50.00	45.00	62.50	66.67	64.00
50th Percentile *(median)*	85.00	100.00	72.00	67.00	65.00	100.00	100.00	80.00
75th Percentile	95.00	100.00	84.00	82.00	80.00	100.00	100.00	92.00
Standard Deviation	26.32	38.39	25.63	23.37	22.59	24.81	34.29	19.30
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	0-100
% Ceiling	18.9	62.5	21.2	5.4	2.2	53.8	69.6	6.6
% Floor	1.7	15.0	1.1	0.0	0.7	0.8	11.8	0.3

Ages 65-74
Males & Females
(N=442)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	69.38	64.54	68.49	62.56	59.94	80.61	81.44	76.87
25th Percentile	55.00	25.00	51.00	47.00	45.00	62.50	66.67	68.00
50th Percentile *(median)*	75.00	75.00	72.00	67.00	65.00	100.00	100.00	80.00
75th Percentile	90.00	100.00	84.00	77.00	80.00	100.00	100.00	92.00
Standard Deviation	26.26	41.30	26.42	22.42	22.12	25.63	34.56	18.08
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	4-100
% Ceiling	9.0	50.0	24.7	2.3	1.2	53.1	73.8	7.2
% Floor	1.5	21.3	1.7	0.0	0.7	1.6	11.8	0.0

TABLE 10.2 CONTINUED

Ages 75 & over
Males & Females
(N=264)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	53.20	43.28	60.88	56.66	50.41	73.89	63.18	73.99
25th Percentile	26.40	0.00	41.00	42.00	35.60	50.00	0.00	64.00
50th Percentile *(median)*	55.00	25.00	62.00	58.50	50.00	87.50	100.00	80.00
75th Percentile	80.00	100.00	84.00	72.00	70.00	100.00	100.00	88.00
Standard Deviation	29.98	41.95	26.01	21.21	23.62	28.75	43.96	20.33
Range	0-100	0-100	0-100	10-100	0-100	0-100	0-100	4-100
% Ceiling	3.5	29.1	16.8	2.9	1.1	42.7	52.1	7.0
% Floor	5.2	35.7	1.9	0.0	3.6	2.4	26.1	0.0

TABLE 10.3 NATIONAL NORMS FOR MALES BY AGE GROUP

Ages 18-24
Males (N=71)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	94.14	93.50	79.62	76.95	65.41	86.09	87.49	78.02
25th Percentile	95.00	100.00	62.00	62.00	55.00	75.00	100.00	68.00
50th Percentile *(median)*	100.00	100.00	84.00	80.00	70.00	100.00	100.00	80.00
75th Percentile	100.00	100.00	100.00	90.00	80.00	100.00	100.00	88.00
Standard Deviation	16.30	21.39	21.47	17.87	19.11	20.18	27.50	16.05
Range	15-100	0-100	12-100	25-100	25-100	25-100	0-100	20-100
% Ceiling	68.7	89.0	41.8	12.1	2.1	54.4	77.9	4.7
% Floor	0.0	3.5	0.0	0.0	0.0	0.0	6.4	0.0

Ages 25-34
Males (N=199)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	94.90	91.86	83.10	79.40	64.66	85.66	82.15	74.13
25th Percentile	95.00	100.00	72.00	72.00	55.00	75.00	66.67	68.00
50th Percentile *(median)*	100.00	100.00	84.00	82.00	70.00	100.00	100.80	80.00
75th Percentile	100.00	100.00	100.00	92.00	80.00	100.00	100.00	88.00
Standard Deviation	11.44	21.04	18.22	17.17	18.99	19.83	31.78	17.24
Range	15-100	0-100	12-100	25-100	10-100	12-100	0-100	12-100
% Ceiling	65.9	82.0	42.0	12.9	1.1	52.7	71.0	2.4
% Floor	0.0	2.8	0.0	0.0	0.0	0.0	9.3	0.0

Ages 35-44
Males (N=239)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	91.39	89.76	79.36	77.55	65.50	88.54	85.52	76.98
25th Percentile	90.00	100.00	72.00	67.00	50.00	87.50	66.67	68.00
50th Percentile *(median)*	95.00	100.00	84.00	77.00	70.00	100.00	100.00	80.00
75th Percentile	100.00	100.00	100.00	90.00	80.00	100.00	100.00	88.00
Standard Deviation	14.68	24.75	21.21	15.92	18.67	18.09	28.12	16.42
Range	20-100	0-100	0-100	5-100	0-100	12-100	0-100	20-100
% Ceiling	47.1	79.9	35.2	7.8	5.0	61.1	75.1	4.1
% Floor	0.0	5.0	0.2	0.0	0.2	0.0	4.6	0.0

TABLE 10.3 CONTINUED

Ages 45-54
Males (N=145)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	86.50	85.58	74.18	73.16	63.05	85.54	85.42	76.38
25th Percentile	85.00	100.00	62.00	62.80	55.00	75.00	100.00	68.00
50th Percentile *(median)*	95.00	100.00	84.00	77.00	67.50	100.00	100.00	80.80
75th Percentile	100.00	100.00	100.00	85.00	80.00	100.00	100.00	88.00
Standard Deviation	20.37	30.24	24.81	17.89	20.44	22.85	29.25	17.62
Range	10-100	0-100	0-100	10-100	0-100	0-100	0-100	20-100
% Ceiling	32.9	76.8	29.6	8.6	1.8	62.8	77.3	4.2
% Floor	0.0	7.3	1.1	0.0	0.4	1.1	4.9	0.0

Ages 55-64
Males (N=105)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	79.88	76.03	68.50	66.58	63.00	83.56	81.14	76.87
25th Percentile	75.00	50.00	42.00	52.00	50.00	75.00	66.67	64.00
50th Percentile *(median)*	90.00	100.00	72.00	67.00	65.00	100.00	100.00	80.00
75th Percentile	95.00	100.00	90.00	85.00	80.00	100.00	100.00	92.00
Standard Deviation	25.47	36.66	26.10	23.27	21.38	31.96	34.02	18.71
Range	0-100	0-100	10-100	5-100	10-100	25-100	0-100	20-100
% Ceiling	25.4	65.6	25.7	5.8	4.8	52.9	71.8	8.2
% Floor	2.3	11.7	0.0	0.0	0.0	0.0	10.9	0.0

Ages 65 & over
Males (N=293)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	65.79	59.72	68.76	58.62	57.80	79.66	76.94	77.37
25th Percentile	45.00	25.00	51.00	47.00	48.00	62.50	66.67	68.00
50th Percentile *(median)*	75.00	75.00	72.00	62.00	60.00	100.00	100.00	84.00
75th Percentile	90.00	100.00	84.00	77.00	75.00	100.00	100.00	92.00
Standard Deviation	28.31	42.51	25.37	22.05	22.55	26.00	37.48	17.42
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	16-100
% Ceiling	6.6	45.9	22.8	1.8	1.7	50.7	68.3	7.6
% Floor	3.0	24.4	0.9	0.0	1.0	1.5	14.9	0.0

TABLE 10.4 NATIONAL NORMS FOR FEMALES BY AGE GROUP

Ages 18–24
Females (N=102)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	90.18	84.91	82.00	76.48	59.71	81.73	78.63	71.53
25th Percentile	90.00	75.00	72.00	72.00	45.00	75.00	66.67	64.00
50th Percentile *(median)*	100.00	100.00	84.00	82.00	60.00	87.50	100.00	76.00
75th Percentile	100.00	100.00	100.00	87.00	75.00	100.00	100.00	84.00
Standard Deviation	20.04	30.73	21.31	18.66	20.09	20.98	33.86	19.44
Range	0-100	0-100	22-100	15-100	10-100	12-100	0-100	24-100
% Ceiling	55.2	75.1	45.1	5.3	2.4	41.8	66.3	1.1
% Floor	1.1	7.3	0.0	0.0	0.0	0.0	9.2	0.0

Ages 25–34
Females (N=275)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	89.12	86.73	79.61	74.80	58.04	84.06	82.32	72.45
25th Percentile	85.00	100.00	62.00	65.00	45.00	75.00	66.67	60.00
50th Percentile	95.00	100.00	84.00	77.00	60.00	100.00	100.00	76.00
75th Percentile	100.00	100.00	100.00	87.00	75.00	100.00	100.00	84.00
Standard Deviation	18.72	27.99	20.94	17.24	20.85	21.66	31.30	18.62
Range	5-100	0-100	0-100	10-100	0-90	0-100	0-100	8-100
% Ceiling	48.8	77.5	37.9	6.0	0.0	53.1	70.6	1.7
% Floor	0.0	4.7	0.2	0.0	0.4	0.2	7.8	0.0

Ages 35–44
Females (N=264)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	88.06	83.65	74.85	74.25	59.43	83.07	80.08	73.32
25th Percentile	85.00	75.00	62.00	62.00	45.00	75.00	66.67	64.00
50th Percentile *(median)*	95.00	100.00	74.00	77.00	60.00	93.75	100.00	76.00
75th Percentile	100.00	100.00	100.00	87.00	75.00	100.00	100.00	84.00
Standard Deviation	17.70	32.21	22.74	19.44	19.72	23.27	33.88	16.79
Range	0-100	0-100	0-100	10-100	5-100	0-100	0-100	20-100
% Ceiling	38.0	75.2	28.0	10.0	1.1	51.3	69.3	2.2
% Floor	0.2	8.4	0.2	0.0	0.0	1.2	10.4	0.0

TABLE 10.4 CONTINUED

Ages 45–54
Females (N=193)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	82.86	79.93	72.14	70.48	60.62	82.71	81.92	74.36
25th Percentile	75.00	75.00	52.00	62.00	50.00	62.50	83.00	68.00
50th Percentile *(median)*	90.00	100.00	74.00	72.00	65.00	87.50	100.00	80.00
75th Percentile	100.00	100.00	84.00	85.00	75.00	100.00	100.00	88.00
Standard Deviation	21.72	35.38	23.34	20.58	21.32	20.84	33.34	18.08
Range	0–100	0–100	0–100	10–100	0–100	12.5–100	0–100	8–100
% Ceiling	26.7	70.9	24.1	6.7	1.1	47.1	74.4	2.4
% Floor	0.9	11.6	0.5	0.0	0.8	0.0	8.9	0.0

Ages 55–64
Females (N=164)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	73.09	71.61	66.64	62.87	58.08	79.43	79.51	73.40
25th Percentile	60.00	33.33	51.00	46.00	40.00	62.50	66.67	64.00
50th Percentile *(median)*	85.00	100.00	72.00	67.00	65.00	100.00	100.00	76.00
75th Percentile	95.00	100.00	84.00	82.00	77.50	100.00	100.00	88.00
Standard Deviation	26.73	39.84	25.26	23.37	23.42	27.02	34.64	19.74
Range	0–100	0–100	0–100	5–100	0–95	0–100	0–100	0–100
% Ceiling	13.4	59.8	17.2	5.5	0.0	54.5	67.7	5.2
% Floor	1.2	17.9	2.1	0.0	1.2	1.5	12.7	0.5

Ages 65 & over
Females (N=413)

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	61.86	56.11	63.44	61.64	55.46	77.00	73.38	74.71
25th Percentile	40.00	0.00	41.00	45.00	40.00	62.50	33.33	64.00
50th Percentile *(median)*	66.70	75.00	62.00	62.00	55.00	87.50	100.00	80.00
75th Percentile	85.00	100.00	84.00	77.00	75.00	100.00	100.00	88.00
Standard Deviation	28.95	42.53	27.12	22.08	23.51	27.69	39.66	19.88
Range	0–100	0–100	0–100	10–100	0–100	0–100	0–100	4–100
% Ceiling	7.3	40.0	21.0	3.1	0.8	48.3	64.3	6.8
% Floor	2.8	27.9	2.4	0.0	2.2	2.2	18.5	0.0

TABLE 10.5 NORMS FOR FIVE MEDICAL CONDITIONS: HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	73.43	62.01	72.31	63.30	58.34	86.70	76.69	77.86
25th Percentile	60.00	25.00	52.00	50.00	45.00	75.00	66.67	68.00
50th Percentile	80.00	75.00	74.00	67.00	60.00	100.00	100.00	84.00
75th Percentile	95.00	100.00	94.00	77.00	75.00	100.00	100.00	92.00
Standard Deviation	26.41	39.40	24.44	19.69	21.38	20.67	35.74	17.39
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	17.14	37.34	19.53	1.58	1.10	55.62	62.37	6.13
% Floor	1.20	21.88	0.81	0.10	0.72	0.57	11.82	0.05

Sample Description

N.	2089	Five Most Prevalent Comorbidities	
Mean Age	59.1	Back Pain/Sciatica	34.0%
Percent Over 65	35.7	Musculoskeletal Complaints	24.6%
Percent Female	58.5	Recent Angina	16.3%
Mean Education	12.5	Diabetes-Type II	16.2%
Percent Poverty	19.2	Varicosities	13.1%

Definition: Physician report of current hypertension.

TABLE 10.6 NORMS FOR FIVE MEDICAL CONDITIONS: CONGESTIVE HEART FAILURE

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	47.54	34.37	62.67	47.05	44.29	71.31	63.67	74.68
25th Percentile	20.00	0.00	42.00	30.00	25.00	50.00	0.00	66.67
50th Percentile	50.00	25.00	64.00	46.25	45.00	87.50	100.00	80.00
75th Percentile	75.00	75.00	90.00	65.00	65.00	100.00	100.00	90.00
Standard Deviation	31.00	39.72	30.97	24.17	24.41	33.06	43.00	21.29
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	.93	14.81	16.67	0.46	0.46	38.43	53.24	5.09
% Floor	4.63	43.52	1.39	0.93	2.31	2.31	19.91	0.46

Sample Description

N	216	Five Most Prevalent Comorbidities	
Mean Age	67.4	Hypertension	52.8%
Percent Over 65	59.7	Back Pain/Sciatica	47.4%
Percent Female	52.3	Past MI	45.7%
Mean Education	12.2	Angina-Recent	40.2%
Percent Poverty	37.3	Musculoskeletal Complaints	32.7%

Definition: Physician report of current congestive heart failure.

TABLE 10.7 NORMS FOR FIVE MEDICAL CONDITIONS: DIABETES-TYPE II

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	67.69	56.75	68.52	56.11	55.73	82.04	75.60	76.74
25th Percentile	50.00	0.00	51.00	40.00	40.00	62.50	66.67	68.00
50th Percentile	75.00	75.00	74.00	52.00	55.00	100.00	100.00	84.00
75th Percentile	90.00	100.00	90.00	72.00	75.00	100.00	100.00	88.00
Standard Deviation	28.66	41.72	26.48	21.12	21.58	24.96	36.63	18.32
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	12.38	31.98	18.30	0.74	0.55	48.06	58.23	6.10
% Floor	1.66	27.91	1.66	0.37	0.55	0.74	14.23	0.18

Sample Description

N	541	Five Most Prevalent Comorbidities	
Mean Age	60.2	Hypertension	64.3%
Percent Over 65	37.7	Back Pain/Sciatica	31.0%
Percent Female	55.6	Musculoskeletal Complaints	25.6%
Mean Education	12.5	Angina-Recent	18.6%
Percent Poverty	22.7	Dermatitis	17.0%

Definition: Physician report of diabetes with age of onset 30 years or older.

TABLE 10.8 NORMS FOR FIVE MEDICAL CONDITIONS: RECENT ACUTE MYOCARDIAL INFARCTION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	69.68	51.41	72.75	59.17	57.68	84.64	73.49	75.78
25th Percentile	60.00	25.00	52.00	45.00	45.00	75.00	66.67	68.00
50th Percentile	75.00	50.00	82.00	62.00	60.00	100.00	100.00	80.00
75th Percentile	95.00	100.00	94.00	77.00	75.00	100.00	100.00	88.00
Standard Deviation	26.12	39.35	25.25	19.34	19.97	22.23	38.01	15.69
Range	0-100	0-100	12-100	10-100	0-100	0-100	0-100	12-100
% Ceiling	6.54	30.84	18.69	0.93	0.93	44.86	57.01	0.93
% Floor	0.93	26.17	0.00	0.00	0.93	0.93	16.82	0.00

Sample Description

N	107	Five Most Prevalent Comorbidities	
Mean Age	59.2	Angina-Ever	55.8%
Percent Over 65	29.0	Angina-Recent	50.7%
Percent Female	30.8	Hypertension	42.5%
Mean Education	12.8	Back Pain/Sciatica	28.7%
Percent Poverty	14.0	Diabetes Type II	24.3%

Definition: Physician report of myocardial infarction within the past year.

TABLE 10.9 NORMS FOR FIVE MEDICAL CONDITIONS: CLINICAL DEPRESSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	71.58	44.39	58.84	52.94	40.12	57.16	38.90	46.26
25th Percentile	55.00	0.00	43.00	35.00	25.00	37.50	0.00	32.00
50th Percentile	80.00	50.00	61.00	52.00	40.00	62.50	33.33	44.00
75th Percentile	95.00	100.00	84.00	72.00	55.00	75.00	66.67	60.00
Standard Deviation	27.17	40.26	26.74	22.98	21.08	27.67	39.80	20.83
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	22.66	23.26	10.74	1.39	0.20	13.12	21.87	0.80
% Floor	0.80	34.00	2.19	0.99	2.39	2.58	39.96	0.40

Sample Description

N	502	Five Most Prevalent Comorbidities	
Mean Age	41.6	Back Pain/Sciatica	45.9%
Percent Over 65	6.0	Angina-Recent	25.0%
Percent Female	75.8	Hypertension	20.9%
Mean Education	13.4	Musculoskeletal Complaints	17.6%
Percent Poverty	23.3	Dermatitis	17.5%

Definition: NIMH (DIS) criteria met for major depression and/or dysthymia.

TABLE 10.10 NORMS FOR COMORBID CONDITION: CHRONIC OBSTRUCTIVE PULMONARY DISEASE (COPD) WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	56.91	34.38	54.82	45.29	54.95	71.82	59.55	68.06
25th Percentile	35.00	0.00	31.00	35.00	30.00	62.30	0.00	56.00
50th Percentile	61.11	25.00	52.00	42.00	50.00	75.00	66.67	72.00
75th Percentile	90.00	75.00	72.00	60.00	55.00	100.00	100.00	84.00
Standard Deviation	29.14	38.73	26.14	18.94	19.55	31.40	44.61	39.66
Range	0-100	0-100	10-100	5-87	0-90	0-100	0-100	13-100
% Ceiling	5.88	14.12	10.59	0.00	0.00	32.94	48.24	2.35
% Floor	2.35	42.35	0.00	0.00	1.18	2.35	27.06	0.00

Sample Description

N	85	Five Most Prevalent Comorbidities	
Mean Age	62.4	Back Pain/Sciatica	53%
Percent Over 65	42.4	Angina-Recent	38%
Percent Female	63.5	Angina-No MI	36%
Mean Education	11.6	Varicosities	31%
Percent Poverty	34.7	Musculoskeletal Complaints	27%

Definition: Lung disease diagnosed by physician as COPD (like chronic bronchitis or emphysema) in past 6 months.

TABLE 10.11 NORMS FOR COMORBID CONDITIONS: RECENT ANGINA WITHOUT MYOCARDIAL INFARCTION, WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	63.24	44.22	61.56	52.00	48.45	80.28	70.16	73.04
25th Percentile	45.00	0.00	41.00	40.00	35.00	62.50	33.33	64.00
50th Percentile	65.00	50.00	62.00	52.00	50.00	87.50	100.00	76.00
75th Percentile	85.00	75.00	84.00	67.00	60.00	100.00	100.00	88.00
Standard Deviation	26.74	39.03	24.53	18.87	20.34	22.95	36.63	18.67
Range	0–100	0–100	0–100	0–97	0–100	12–100	0–100	8–100
% Ceiling	6.64	16.41	4.69	0.00	0.39	35.55	49.22	2.34
% Floor	1.56	37.50	0.78	0.39	1.17	0.00	15.23	0.00

Sample Description

N	256	Five Most Prevalent Comorbidities	
Mean Age	59.7	Back Pain/Sciatica	50%
Percent Over 65	39.4	Musculoskeletal Complaints	29%
Percent Female	55.1	Past MI	24%
Mean Education	12.7	Dermatitis	21%
Percent Poverty	23.0	Osteoarthritis	18%

Definition: Symptoms of angina in past 6 months in the absence of an MI within 1 year.

TABLE 10.32 NORMS FOR COMORBID CONDITIONS: BACK PAIN/SCIATICA WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	66.32	46.71	59.34	58.45	52.39	81.48	70.90	74.93
25th Percentile	45.00	0.00	41.00	42.00	35.00	75.00	33.33	64.00
50th Percentile	75.00	50.00	61.00	60.00	55.00	87.50	100.00	80.00
75th Percentile	93.75	100.00	82.00	77.00	70.00	100.00	100.00	88.00
Standard Deviation	28.60	40.51	24.63	21.63	22.74	24.38	38.97	18.62
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	13-100
% Ceiling	9.98	21.41	7.48	1.04	0.42	44.07	54.89	4.37
% Floor	1.46	34.30	1.04	0.21	0.42	0.83	16.84	0.00

Sample Description

N	481	Five Most Prevalent Comorbidities	
Mean Age	60.4	Musculoskeletal Complaints	30%
Percent Over 65	35.8	Angina-Recent	28%
Percent Female	64.2	Angina-No MI	27%
Mean Education	12.2	Varicosities	21%
Percent Poverty	20.6	Osteoarthritis	21%

Definition: Attacks of back pain or sciatica in last 6 months.

TABLE 10.13 NORMS FOR COMORBID CONDITIONS: OSTEOARTHRITIS WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	57.44	38.17	55.04	58.96	49.54	79.74	74.84	78.01
25th Percentile	30.00	0.00	32.00	42.00	35.00	62.50	33.33	68.00
50th Percentile	65.00	25.00	52.00	60.00	45.00	100.00	100.00	84.00
75th Percentile	80.00	75.00	72.00	77.00	65.00	100.00	100.00	92.00
Standard Deviation	29.19	39.40	26.27	21.42	22.19	27.12	37.36	19.02
Range	0-100	0-100	0-100	5-100	0-100	0-100	0-100	15-100
% Ceiling	3.43	13.71	5.29	0.57	0.57	76.86	58.86	5.14
% Floor	1.71	44.00	2.29	0.00	1.71	1.14	13.14	0.00

Sample Description

N	175	Five Most Prevalent Comorbidities	
Mean Age	67.8	Back Pain/Sciatica	57%
Percent Over 65	58.9	Varicosities	27%
Percent Female	74.3	Angina–No MI	26%
Mean Education	11.9	Angina–Recent	28%
Percent Poverty	23.8	Musculoskeletal Complaints	19%

Definition: Now have active condition physician ever diagnosed as arthritis and physician ever labeled it osteoarthritis or degenerative arthritis and patient is ≥ 55 years old.

TABLE 10.14 NORMS FOR COMORBID CONDITIONS: MUSCULOSKELETAL COMPLAINTS WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	67.58	56.15	66.57	59.85	56.82	87.17	73.14	78.23
25th Percentile	50.00	0.00	51.00	45.00	40.00	75.00	33.33	68.00
50th Percentile	75.00	75.00	72.00	62.00	60.00	100.00	100.00	84.00
75th Percentile	85.00	100.00	84.00	77.00	75.00	100.00	100.00	92.00
Standard Deviation	25.66	41.09	24.39	20.55	21.55	20.25	37.95	17.61
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	8-100
% Ceiling	8.80	31.67	12.02	0.88	0.59	53.37	58.65	6.74
% Floor	0.88	26.39	0.29	0.29	0.59	0.88	13.20	0.00

Sample Description

N	341	Five Most Prevalent Comorbidities	
Mean Age	61.1	Back Pain/Sciatica	43%
Percent Over 65	41.6	Angina Pectoris	23%
Percent Female	63.0	Varicosities	22%
Mean Education	12.0	Angina No MI	21%
Percent Poverty	22.7	Dermatitis	18%

Definition: Active condition physician *ever* diagnosed as arthritis but criteria for osteoarthritis or rheumatoid arthritis not met.

TABLE 10.15 NORMS FOR COMORBID CONDITIONS: BENIGN PROSTATIC HYPERTROPHY SYMPTOMS WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	73.35	61.48	73.98	63.41	61.54	88.22	78.04	82.48
25th Percentile	56.00	25.00	60.00	50.00	50.00	75.00	66.67	76.00
50th Percentile	81.25	75.00	82.00	67.00	65.00	100.00	100.00	88.00
75th Percentile	95.00	100.00	100.00	80.00	75.00	100.00	100.00	95.00
Standard Deviation	25.11	39.41	25.03	23.24	21.17	19.48	34.95	17.20
Range	0-100	0-100	0-100	5-100	0-100	12-100	0-100	24-100
% Ceiling	9.24	34.24	20.11	1.09	1.09	53.80	61.43	12.50
% Floor	2.17	24.46	1.09	0.00	1.09	0.00	13.04	0.00

Sample Description

N	184	Five Most Prevalent Comorbidities	
Mean Age	67.1	Back Pain/Sciatica	32%
Percent Over 65	53.8	Musculoskeletal Complaints	27%
Percent Female	0.0	Post MI	23%
Mean Education	12.5	Angina-Recent	22%
Percent Poverty	17.6	Angina-No MI	21%

Definition: Male, age ≥ 50 years, history of nocturia in past six months, no earlier kidney disease ever diagnosed, and no report of prostatic cancer.

TABLE 10.16 NORMS FOR COMORBID CONDITIONS: VARICOSITIES WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	69.12	50.24	67.58	61.49	54.64	85.59	69.36	77.29
25th Percentile	50.00	25.00	51.00	45.00	40.00	75.00	33.33	68.00
50th Percentile	75.00	50.00	72.00	62.00	55.00	100.00	100.00	84.00
75th Percentile	90.00	100.00	90.00	77.00	75.00	100.00	100.00	92.00
Standard Deviation	26.84	38.44	24.34	19.94	21.45	20.85	40.23	18.57
Range	0-100	0-100	12-100	5-100	0-100	0-100	0-100	16-100
% Ceiling	9.01	28.38	17.12	1.35	0.45	54.05	57.21	8.56
% Floor	0.90	25.68	0.00	0.00	0.45	0.45	14.86	0.00

Sample Description

N	222	Five Most Prevalent Comorbidities	
Mean Age	61.8	Back Pain/Sciatica	45%
Percent Over 65	39.2	Musculoskeletal Complaints	34%
Percent Female	72.1	Angina-Recent	25%
Mean Education	12.2	Angina-No MI	25%
Percent Poverty	20.1	Osteoarthritis	21%

Definition: Now have condition that physician ever diagnosed as varicose veins/deep varicosities.

TABLE 10.17 NORMS FOR COMORBID CONDITIONS: DERMATITIS WITH HYPERTENSION

	PF	RP	BP	GH	VT	SF	RE	MH
Mean	74.80	56.82	67.00	60.78	56.87	84.22	78.71	76.46
25th Percentile	65.00	25.00	51.00	47.00	45.00	75.00	66.67	68.00
50th Percentile	85.00	75.00	72.00	62.00	60.00	100.00	100.00	80.00
75th Percentile	95.00	100.00	90.00	77.00	70.00	100.00	100.00	88.00
Standard Deviation	25.85	39.36	25.62	18.40	21.27	23.78	34.94	17.66
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	13-100
% Ceiling	14.72	29.87	15.58	0.43	0.87	52.81	62.34	6.93
% Floor	1.30	26.41	1.30	0.43	1.30	0.43	9.96	0.00

Sample Description

N	231	Five Most Prevalent Comorbidities	
Mean Age	57.6	Back Pain/Sciatica	53%
Percent Over 65	35.5	Musculoskeletal Complaints	38%
Percent Female	48.5	Angina-Recent	36%
Mean Education	13.2	Angina-No MI	31%
Percent Poverty	16.3	Varicosities	27%

Definition: Repeated episodes of dermatitis or skin rash in past 6 months.

TABLE 10.18 PATIENT NORMS FOR THE MENTAL HEALTH (MH) SCALE

Males & Females Combined
Various age groups

	18-24	25-34	35-44	45-54	55-64	65-74	75+
Mean	73.47	71.03	70.15	72.29	74.48	76.52	75.50
25th Percentile	64.00	60.00	56.00	60.00	64.00	68.00	64.00
50th Percentile *(median)*	80.00	76.00	76.00	80.00	80.00	80.00	80.00
75th Percentile	88.00	88.00	84.00	88.00	92.00	92.00	92.00
Standard Deviation	18.48	19.76	20.36	20.67	20.43	19.75	20.88
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	2.56	2.46	2.33	5.09	6.73	8.94	12.35
% Floor	0.20	0.08	0.13	0.28	0.20	0.30	0.74
N	1015	2563	2358	1415	1484	1330	680

Males
Various age groups

	18-24	25-34	35-44	45-54	55-64	65+
Mean	77.20	74.03	73.04	75.22	77.78	79.53
25th Percentile	68.00	64.00	64.00	64.00	68.00	72.00
50th Percentile *(median)*	80.00	80.00	80.00	80.00	84.00	84.00
75th Percentile	88.00	88.00	88.00	88.00	92.00	92.00
Standard Deviation	16.67	18.32	19.25	20.12	19.40	17.91
Range	12-100	8-100	0-100	0-100	0-100	0-100
% Ceiling	3.92	2.86	2.76	6.87	8.44	13.60
% Floor	0.00	0.00	0.11	0.53	0.31	0.12
N	306	908	941	568	652	809

Females
Various age groups

	18-24	25-34	35-44	45-54	55-64	65+
Mean	71.86	69.38	68.23	70.33	71.88	73.92
25th Percentile	60.00	56.00	56.00	56.00	56.00	60.00
50th Percentile *(median)*	76.00	76.00	72.00	76.00	76.00	80.00
75th Percentile	88.00	84.00	84.00	88.00	88.00	92.00
Standard Deviation	18.99	20.24	20.86	20.82	20.87	21.23
Range	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	1.97	2.24	2.05	3.90	5.41	7.74
% Floor	0.28	0.12	0.14	0.12	0.12	0.67
N	709	1655	1417	847	832	1201

TABLE 10.19 PATIENT NORMS FOR THE MH SCALE: MEDICAL PROVIDERS

Males & Females Combined

Various age groups

	18-24	25-34	35-44	45-54	55-64	65-74	75+
Mean	75.87	74.11	73.81	74.83	76.16	77.18	76.13
25th Percentile	68.00	64.00	64.00	64.00	64.00	68.00	64.00
50th Percentile *(median)*	80.00	80.00	80.00	80.00	80.00	84.00	80.00
75th Percentile	88.00	88.00	88.00	88.00	92.00	92.00	92.00
Standard Deviation	16.57	17.84	18.30	19.46	19.27	19.27	20.30
Range	0-100	0-100	0-100	0-100	0-100	0-100	0-100
% Ceiling	2.84	2.94	2.91	5.65	7.14	9.17	12.61
% Floor	0.11	0.05	0.11	0.25	0.07	0.31	0.60
N	916	2142	1855	1221	1385	1286	666

Males

Various age groups

	18-24	25-34	35-44	45-54	55-64	65+
Mean	79.65	76.67	76.29	77.49	79.55	80.09
25th Percentile	72.00	68.00	68.00	68.00	72.00	72.00
50th Percentile *(median)*	84.00	80.00	80.00	84.00	84.00	84.00
75th Percentile	88.00	88.00	88.00	92.00	92.00	92.00
Standard Deviation	14.53	16.57	17.40	18.74	17.80	17.51
Range	12-100	10-100	0-100	0-100	0-100	0-100
% Ceiling	4.36	3.40	3.29	7.46	8.87	13.91
% Floor	0.00	0.00	0.13	0.40	0.16	0.13
N	275	764	759	496	609	791

Females

Various age groups

	18-24	25-34	35-44	45-54	55-64	65+
Mean	74.25	72.68	72.11	73.01	73.48	74.60
25th Percentile	64.00	64.00	60.00	60.00	60.00	64.00
50th Percentile *(median)*	80.00	76.00	76.00	80.00	80.00	80.00
75th Percentile	88.00	88.00	86.5	88.00	88.00	92.00
Standard Deviation	17.13	18.37	18.72	19.74	19.97	20.67
Range	0-100	0-100	0-100	0-100	10-100	0-100
% Ceiling	2.18	2.69	2.65	4.41	5.80	7.92
% Floor	0.16	0.07	0.09	0.14	0.00	0.60
N	641	1378	1096	735	776	1261

TABLE 10.20 PATIENT NORMS FOR THE MH SCALE: MENTAL HEALTH PROVIDERS

Total Sample (MOS)

Various age groups

	18-24	25-34	35-44	45-54	55-64	65-74	75+
Mean	51.27	55.38	56.61	56.31	51.03	57.14	45.71
25th Percentile	36.00	40.00	40.00	44.00	36.00	36.00	32.00
50th Percentile *(median)*	52.00	56.00	56.00	56.00	52.00	56.00	38.00
75th Percentile	68.00	72.00	72.00	72.00	68.00	76.00	64.00
Standard Deviation	20.42	21.59	21.83	20.96	21.95	23.83	26.63
Range	0-96	0-96	0-100	0-100	0-100	16-100	0-92
% Ceiling	0.00	0.00	0.20	1.55	1.01	2.27	0.00
% Floor	1.01	0.24	0.20	0.52	2.02	0.00	7.14
N	99	421	503	194	99	44	14

Males

Various age groups

	18-24	25-34	35-44	45-54	55-64	65+
Mean	55.48	60.03	59.46	59.61	52.72	55.22
25th Percentile	40.00	48.00	48.00	44.00	35.00	40.00
50th Percentile *(median)*	52.00	60.00	64.00	62.00	52.00	58.00
75th Percentile	68.00	80.00	72.00	76.00	76.00	68.00
Standard Deviation	18.84	20.73	20.67	22.36	23.62	18.65
Range	24-92	8-96	8-100	0-100	0-100	20-84
% Ceiling	0.00	0.00	0.55	2.78	2.33	0.00
% Floor	0.00	0.00	0.00	1.39	2.33	0.00
N	31	144	182	72	43	18

Females

Various age groups

	18-24	25-34	35-44	45-54	55-64	65+
Mean	49.35	52.97	55.00	54.36	49.73	54.00
25th Percentile	34.00	36.00	40.00	40.00	38.00	32.67
50th Percentile *(median)*	46.00	56.00	56.00	56.00	50.00	52.00
75th Percentile	68.00	72.00	72.00	68.00	62.5	78.00
Standard Deviation	20.96	21.68	22.34	19.92	20.69	27.31
Range	0-92	0-92	0-92	4-100	0-88	0-100
% Ceiling	0.00	0.00	0.00	0.82	0.00	2.50
% Floor	1.47	0.36	0.31	0.00	1.79	2.50
N	68	277	321	122	56	40

TABLE 10.21 NATIONAL NORMS FOR DICHOTOMOUS LIMITATIONS INDICATORS (N=2,474)

Total Sample
By age groups

		18-24	25-34	35-44	45-54	55-64	65-74	75+	Total
Any Physical Limitation	N	136	258	282	233	240	220	133	1502
	%	38.1	42.7	57.5	70.3	81.1	91.0	96.5	61.2
Any Role Disability Limitation	N	134	233	189	125	135	128	105	1049
	%	37.7	38.7	38.6	37.5	45.8	52.7	75.0	42.8
Emotional Limitation	N	47	83	61	47	42	25	24	329
	%	13.2	13.8	12.6	14.1	14.2	10.3	16.9	13.4
Fair/Poor Personal Evaluation	N	27	40	38	42	75	72	65	359
	%	7.7	6.6	7.8	12.6	25.2	29.5	45.3	14.6

Males
By age groups

		18-24	25-34	35-44	45-54	55-64	65-74	Total
Any Physical Limitation	N	55	102	127	106	102	147	639
	%	31.3	34.1	52.9	67.1	74.6	93.4	54.8
Any Role Disability Limitation	N	46	113	91	54	55	91	450
	%	26.5	37.9	37.7	33.9	40.7	57.3	38.7
Emotional Limitation	N	13	36	25	20	18	15	127
	%	7.4	12.2	10.5	12.4	13.3	9.6	10.9
Fair/Poor Personal Evaluation	N	14	20	12	19	26	58	149
	%	8.1	6.6	4.9	12.0	18.7	36.0	12.7

Females
By age groups

		18-24	25-34	35-44	45-54	55-64	65-74	Total
Any Physical Limitation	N	81	155	154	126	138	204	858
	%	44.8	51.2	62.0	73.3	86.7	92.7	67.0
Any Role Disability Limitation	N	87	119	98	70	79	140	593
	%	48.6	39.5	39.5	40.8	50.1	63.3	46.4
Emotional Limitation	N	34	46	36	27	24	33	200
	%	18.8	15.3	14.5	15.6	15.0	14.9	15.6
Fair/Poor Personal Evaluation	N	13	20	26	23	49	78	209
	%	7.4	6.6	10.5	13.1	31.0	34.7	16.3

II. APPLICATIONS OF THE SF-36

Of the many potential applications of the SF-36, four examples are briefly discussed below: (1) monitoring the health of the general population, (2) estimating the burden of different conditions, (3) clinical trials of treatment effects; and (4) monitoring outcomes in clinical practice.

Monitoring population health

The health of the general population in developed countries cannot be well understood from analyses of treatment survival rates or from population mortality statistics (Elinson & Mattison, 1984). Standardized measures of physical and mental functioning and well-being, social and role disability, and general health perceptions are necessary for comprehensive monitoring of the health of the general population. These concepts, however, are rarely measured in general population surveys (NCHS, 1991).

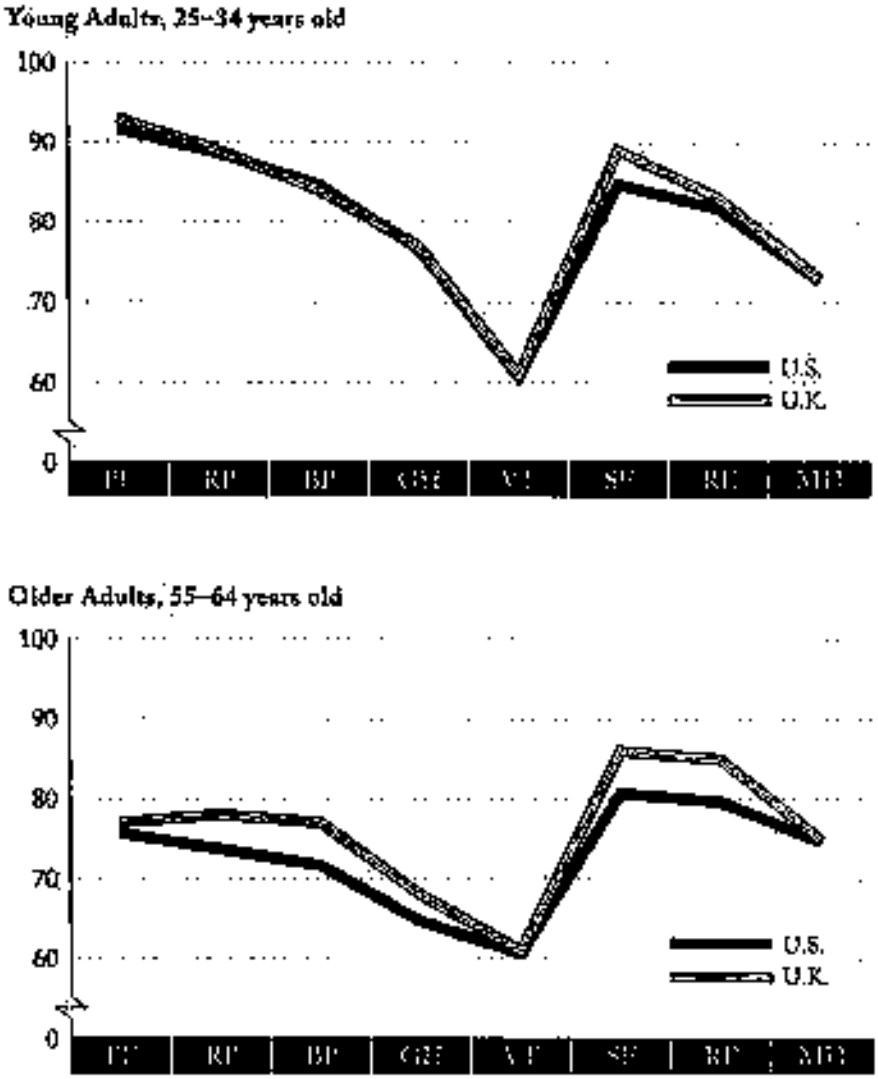
To illustrate the use of the SF-36 in measuring the health of the general population and in comparing different population groups, we compared SF-36 profiles for those between the ages of 25 to 34 and 55 to 64 in the general U.S. and U.K. populations. U.S. data come from the National Survey of Functional Health Status (Chapter 10); U.K. data come from Jenkinson et al. (1993). Men and women were included in both age groups for both countries. We chose these two groups because they are the youngest and oldest common age groups surveyed across the two countries.

The general population survey in the United Kingdom used mailout-mailback data collection methods for a sample of 13,042 randomly selected subjects. Jenkinson et al. reported a return rate of approximately 72% (1993). The U.S. survey, described in Chapter 10, used mailout-mailback as the primary data collection methodology and both mail and telephone follow-up to achieve a return rate of approximately 77% (McHorney, Kosinski, & Ware, in review). Because Jenkinson et al. (1993) used the Developmental version of the SF-36, we scored the Bodily Pain (BP) scale using the Developmental scoring method (see Chapter 6) in the United States to standardize comparisons shown in Figure 11.1. Chapter 10 compares results for

the Developmental and Standard scoring methods for the BP scale.

Plots of mean scores for younger and older age groups for the United States and the United Kingdom reveal interesting trends (Figure 11.1). Older respondents in both countries scored substantially lower than their younger counterparts, particularly across the first four scales, which are most sensitive to differences in physical functioning and well-being (Physical Functioning [PF], Role–Physical [RP], BP, and General Health [GH] scales). These differences in profiles across younger and older age groups can be seen by comparing the upper and lower panels of Figure 11.1.

FIGURE 11.1
COMPARISON OF HEALTH PROFILES FOR YOUNGER AND OLDER U.S. AND U.K. ADULTS



For the younger age groups, virtually identical profiles of scores across the two countries are apparent for seven of the eight scales (all but the Social Functioning [SF] scale). Younger aged respondents in the United Kingdom reported higher levels of social functioning, on average (89 versus 85 points).

As the lower panel of Figure 11.1 shows, elderly respondents in the United Kingdom scored higher than elderly respondents in the United States on five of the eight scales (RP, BP, GH, SF, and Role–Emotional [RE]). Differences between countries were much smaller for the PF, Vitality [VT], and Mental Health [MH] scales. We ignore statistical significance for purposes of this example, because standard errors are roughly equal to the thickness of the lines in Figure 11.1.

Many factors may account for the apparent differences across the two countries in profiles for the 55 to 64 age group. These analyses did not control for sociodemographic characteristics other than age, which may account for some of the differences. It is unlikely that changes in five SF-36 items necessary for the British adaptation (see Chapter 3) explain differences between countries. Mean scores for seven of eight scales are virtually identical across countries for younger respondents. Five of the British adaptations of SF-36 items involved the PF, VT, and MH scales, which yielded nearly identical mean scores for the two countries in both younger and older age groups. British adaptations of SF-36 items were not required for the RP, BP, and RE scales, which showed large differences between countries. Thus, adaptations in item wording are not a likely explanation for differences.

It is interesting that U.K. scores for five of eight scales were substantially higher than those in the United States for the 55 to 64 year age group in these preliminary analyses. Given that a larger proportion of the gross domestic product is believed to be devoted to treatment of this age group in the United States, compared with the United Kingdom, these preliminary trends raise questions about the health benefits of that investment. Standardization of the SF-36 for use in the United Kingdom and other countries will facilitate further study of population differences, specific treatment benefits, and various health care policy issues (Aaronson et al., 1992).

Estimating the burden of different conditions

The SF-36 and other standardized assessment methods offer a number of advantages to providers. Although most providers attempt to elicit information from their patients, the SF-36 can be used to obtain functioning and well-being information in a standardized way. By standardizing questions, answers, and scoring, reliable and valid comparisons can be made to determine the relative burden of different conditions. Examples from the Medical Outcomes Study (MOS) are discussed in the following section.

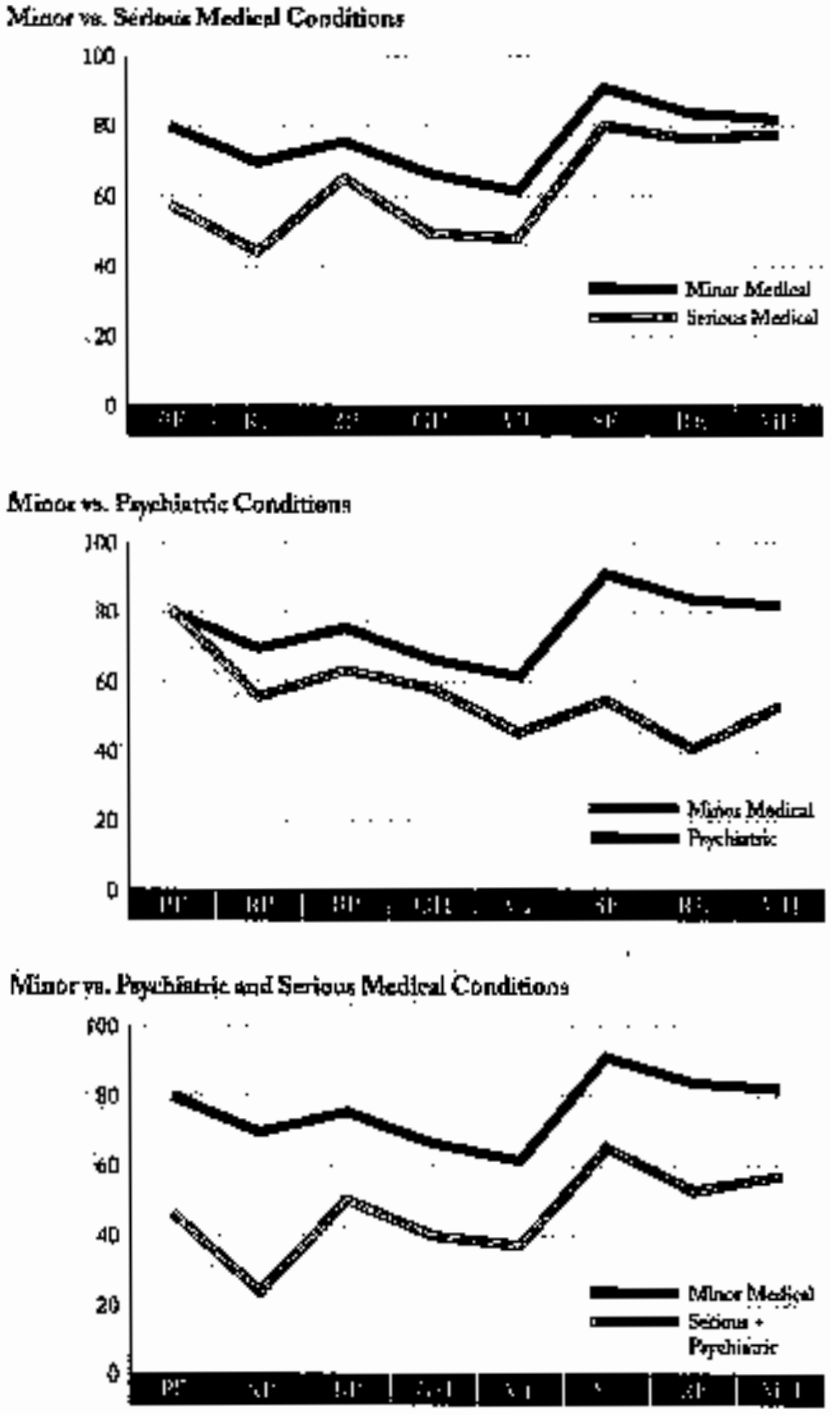
Comparing Health Profiles

The MOS team at The Health Institute (THI) published SF-36 profiles for groups of patients known to differ in the severity of their medical and/or psychiatric conditions (McHorney et al., 1993). Chronically ill patients who were relatively well (labeled "minor medical") included those with uncomplicated hypertension (N=638). Those with severe chronic medical conditions such as congestive heart failure, chronic obstructive pulmonary disease, and/or advanced diabetes were included in the "serious medical" group (N=168). Those with a psychiatric condition such as depressive disorder were placed in the "psychiatric" group (N=163).

The three panels in Figure 11.2 present SF-36 profiles for these groups. A unique feature of the SF-36 profile is that the eight scales are ordered from left to right according to the extent to which they measure physical or mental health. We expect a divergent pattern of profiles for the three MOS groups defined above. As expected for the first panel, which compares the minor and serious medical groups, the profile for the serious medical group is lowest (relative to the minor group) on the left side of the SF-36 profile, which includes the four scales most sensitive to differences in physical health status (including PF, RP, BP, and GH). Differences on the right side of the profile tend to be much smaller.

Comparing the minor medical group and the psychiatric group, the differences in profiles tend to be largest among scales on the right side of the profile (including MH, RE, SF, and VT). These scales are the most sensitive to differences in mental health status (Chapters 8 and 9). The VT scale, which is sensitive to the impact of both physical and mental conditions, reveals roughly equal separations between groups across the first two panels.

FIGURE 11:2
MINOR VERSUS SERIOUS MEDICAL AND PSYCHIATRIC CONDITIONS



Note. Data from "The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs," by C.A. McHorney et al., 1993, *Medical Care, 31*, 247-263.

The third panel of Figure 11.2 compares those with both serious medical and psychiatric conditions with those in the minor medical group. As would be expected, this group of patients scores low on all eight scales. Differences between these two groups are substantial and roughly equal on the left side and on the right side of the SF-36 profile.

The following section illustrates the use of norms in estimating the burden of specific conditions (see also Garratt et al., 1993). Cross-sectional SF-36 profiles and changes in those profiles over time will be reported later in 1993 for the five MOS tracer conditions and selected comorbid conditions. The forthcoming analyses control for differences in sociodemographic characteristics and the added "effects" of comorbid conditions.

Clinical trials of treatment effects

Evaluations of treatment in terms of general health outcomes have been rare, with some noteworthy exceptions (Bombardier et al., 1986; Croog et al., 1986; Fowler et al., 1988). Increasingly, alternative treatments are being evaluated in terms of their impact on patient functioning and well-being, in addition to traditionally defined medical endpoints (Chobanian, 1987). Reasons for this include the need to better understand trade-offs between changes in clinical parameters versus functioning and well-being and trade-offs between treatments that appear equally efficacious (e.g., in terms of blood pressure control) but differ in their impact on functioning and well-being. These and many other trade-offs must be documented to better inform the public about the implications of their treatment choices (Fowler et al., 1988).

Overview of Clinical Trials

Among the many clinical trials using the SF-36 are two NIH-sponsored studies: the National Breast Cancer Prevention Trial and the Prostate Cancer Prevention Trial. These trials are each administering the SF-36 at regular intervals over 5 to 7 year periods to 15,000 to 20,000 adults sampled from 100 to 300 sites. The goal is to better understand the trade-offs between the benefits of cancer prevention and undesirable treatment side effects. Standard U.S., Mexican-American, and French-Canadian SF-36 forms are being fielded along with supplemental generic and disease-specific measures.

These National Institutes of Health (NIH) trials exemplify state-of-the-art patient-based measurement models for treatment evaluation. In addition to a generic "core" subset of measures, both trials have added measures of social functioning and symptom/problem lists specific to each condition and to each treatment. Because of the large size, long duration, and comprehensiveness of their survey and clinical measures, these trials will yield a rich database for advancing understanding and interpretation of the SF-36.

We are aware of 260 clinical trials using the SF-36 to assess general health outcomes from the patient point-of-view. Table 11.1 lists 158 topics being studied in one or more clinical studies using the SF-36 and registered with the Medical Outcomes Trust as of June 1993. This list relies upon investigator reports regarding topics under study, and some categories have been collapsed because they were judged to be equivalent.

Preliminary findings from more than 36 clinical trials have been presented at professional meetings. To date, we are aware of more than two dozen publications reporting results from clinical studies that included the SF-36 (see the Annotated Bibliography). Plans to survey these investigators are under way, and we hope to summarize their experiences.

Published Outcomes Studies

This section summarizes results from two longitudinal studies of SF-36 profiles before and after treatment. Readers are encouraged to review the source publications for more complete details (Phillips & Lansky, 1992; Katz et al., 1992).

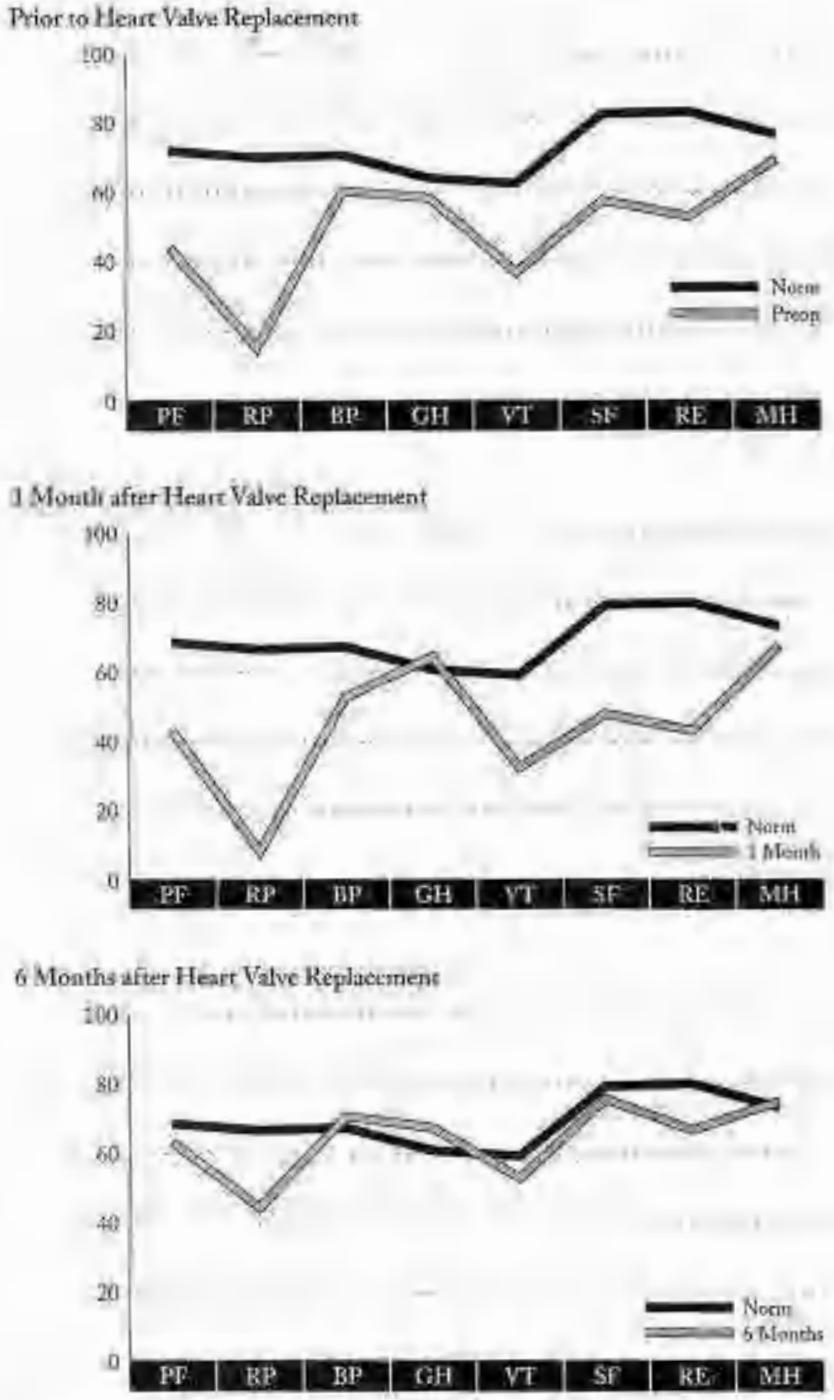
The Burden of Heart Disease and the Benefits of Heart Valve Replacement
Figure 11.3 illustrates results from a study that used the SF-36 profile to estimate the burden of heart disease and the benefit of surgical treatment. Phillips and Lansky (1992) administered the SF-36 to 100 patients before a heart valve replacement and again 1 and 6 months after surgery. The first panel of Figure 11.3 shows their SF-36 profile before surgery ("Preop"). To illustrate the value of norms in determining the health burden represented by this profile, the profile for the general population is also presented ("Norm").

Prior to heart valve replacement, patients scored below the norm on all eight scales and were particularly low in PF, RP, VT, SF, and RE. As shown in the

TABLE 11.1 LIST OF 158 TOPICS UNDER STUDY IN TRIALS USING THE SF-36 HEALTH SURVEY

Acetabular Fracture
Aerobic Exercise
Alcoholism
Allergic Bronchopulmonary Disease
Aspergillosis
Aneurysm
Angina
Angioplasty
Anxiety Disorder
Arterial Occlusive Disease
Arthritis, Osteoarthritis
Arthritis, Rheumatoid
Asthma
Atrial Fibrillation
Behavior Problems
Bone Loss
Bone Marrow Transplantation
Bone Mass
Bowel Cancer
Breast Cancer
Burns
Cancer, Solid Tumor
Cardiac Rehabilitation
Cardiac Surgery
Cardiomyoplasty Surgery
Cardiopulmonary Rehabilitation
Cardiovascular Disease
Caregiver Training
Chest Pain
Chronic Fatigue Syndrome
Chronic Illness, Effect on Family
Chronic Mental Illness
Chronic Obstructive Lung Disease
Chronic Pain, General
Cognition
Congestive Heart Failure
Coping & Support Resources
Coronary Artery Bypass Graft Surgery
Coronary Artery Disease
Coronary Revascularization
Critical Care
Cystic Fibrosis
Dementia Caregivers
Dental Conditions
Depression
Diabetes, Type I
Diabetes, Type II
Diabetic Foot Ulcers
Dialysis
Distal Humerus Fractures
Drug Abuse
End Stage Renal Disease
Endometriosis
Esophageal Cancer
Esophagectomy
Femoral Shaft Fracture
Fibromyalgia
Firearm Injuries
Functional Dyspepsia
Gall Bladder
Gastro-Esophageal Reflux
General Population Surveys
Geriatric Assessment, Evaluation & Management
Glaucoma
Growth Hormone
Head & Neck Cancer
Health Promotion
Heart Disorder, Congenital
Heart Transplant
Heel Pain Syndrome
Hemochromatosis
High-Risk for Cancer
Hip Fracture, Replacement
HIV Infection
Home Antibiotic Therapy
Home Blood Pressure Monitoring
Hospitalized Mental Patients
Hypercholesterolemia
Hypertension
Intermittent Claudication
Interstitial Lung Disease
Intra-Articular Calcaneal Fracture
Irritable Bowel Syndrome
Ischemic Leg Disease
Joint Reconstruction Procedures
Knee Replacement
Lifestyle Intervention
Liver Transplantation
Low Back Pain
Low Blood Pressure
Low Self Image
Lung Function
Lung Transplantation
Lymphoma
Macular Holes
Melanoma
Methadone Maintenance
Migraine
Mild Head Injuries
Mild High Blood Pressure
Mitral Stenosis
Multiple Sclerosis
Multiple Trauma Patients
Nephrology
Neurological Disease
Neuropsychologic Disease
Nutrition
Nutritional Supplements in Dialysis
Obstructive Sleep Apnea
Odorous Emissions, Effects of
Orthopaedic Implants
Pacemakers
Panic Attack
Parkinson's Disease
Partial Meniscectomy
Patients Admitted to Hospital
Pelvic Fracture
Peripheral Vascular Angioplasty
Post-Menopause
Prostate Disease, Benign
Prostate Cancer
Prostatic Hyperplasia
Psychiatric Outpatients
Psychiatric Services, Inpatient & Outpatient
Psychiatric Disorders
Psychosexual Problems
Psychosocial Interventions
Psychotherapy
Pulmonary Disease
Radial Keratotomy
Renal Failure, Chronic
Renal Transplant
Respiratory Tract Infection
Retinitis
Rheumatic Disease
Rhinitis
Schizophrenia, Chronic
Sclerosis
Short Bowel Syndrome
Sickle Cell Disease
Sinusitis
Skeletal Disease
Stress Incontinence
Stroke
Substance Abuse
Thoracic Outlet Syndrome
Thyroid Disease
Tibial Fractures
Total Knee Replacement
Transferring Patients
Trauma
Urinary Incontinence
Urologic Surgery
Varicose Veins
Vascular Rehabilitation
Vertebral Fractures
Violent Relationships
Weight Loss

FIGURE 11.3
NORM AND PROFILE FOR PATIENTS UNDERGOING HEART VALVE REPLACEMENT



Note. Data from "Outcomes Management in Heart Valve Replacement Surgery: Early Experience" by R.C. Phillips & D.J. Lansky, 1992, *Journal of Heart Valve Disease, 1,* 42-50.

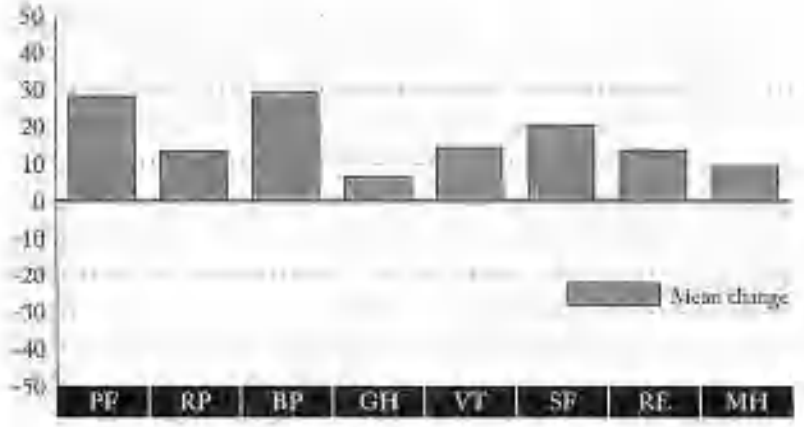
second panel, they remained below the norm for seven of eight concepts, all but GH, 1 month after heart valve replacement. It is interesting to note that their confidence in their health (GH scale) improved to equal the norm 1 month after surgery although they were still recovering. Six months after heart valve replacement, the patients scored at or above the norm for all but the RP and RE scales (third panel of Figure 11.3).

Sensitivity to the Benefits of Hip Replacement

Katz and his colleagues (1992) administered the SF-36 and other widely used health measures to 54 patients before and after total hip arthroplasty. Figure 11.4 presents SF-36 difference scores (follow-up score minus pre-surgery score). All differences are above the zero axis indicating that scores for all SF-36 scales were higher, on average, following surgery.

To compare measures of outcomes, Katz and his colleagues standardized each change score in terms of standard deviation (SD) units, using the SD from the distribution of change scores. The SF-36 scale most sensitive to the benefit of surgery was the BP scale, which improved 1.28 SD units, followed by the PF scale (1.26 SD units), the VT scale (0.79 SD units), and SF (0.73 SD units). Standardization did not dramatically change the graphical presentation of unstandardized change scores (Figure 11.4), with a few exceptions. Improvements in GH and MH, however, were larger (in SD units) than they appear in Figure 11.4. Both scores improved 0.5 SD units or more, which was larger than the change in the two role disability scales

FIGURE 11.4
CHANGES WITH HIP REPLACEMENT



Note. Data from "Comparative Measurement Sensitivity of Short and Longer Health Status Instruments" by J.N. Katz et al., 1992, *Medical Care, 30*, 917–925.

(which changed 0.32 and 0.31 SD units).

Katz and his colleagues compared these results with those for the 136-item Sickness Impact Profile (SIP) and other short-form measures. Despite their brevity, the SF-36 PF and MH scales did at least as well as any other scales studied in detecting the physical and mental health outcomes of this surgery. The four scales that appeared to be most sensitive, from the roughly three dozen scales that were compared, included the SF-36 BP scale, the SF-36 PF scale, the MHAQ Physical Activities scale, and the AIMS Bodily Pain scale. Scores for these scales improved by roughly one and a quarter or more SD units.

This prospective study suggests that, at least in relation to the benefits of hip replacement, very short scales may be sensitive to outcomes and that it may be more useful to measure the right concept with brevity than to measure the wrong concept in great detail. The principal benefit of hip replacement appears to be physical. Measures of mental health studied, including SF-36 measures, were less sensitive to these benefits, with few exceptions. These results are consistent with guidelines recommended for interpreting the SF-36 scales (see Chapter 8).

Figures 11.3 and 11.4 present empirical evidence of "clinical" validity and precision of SF-36 scales, their sensitivity to the burden of specific conditions, and their sensitivity to change over time. For example, patients known to be worse off clinically were shown to score below the SF-36 norm. Patients who received treatments known to be beneficial in clinical terms experienced improvements in SF-36 scores, on average. The results discussed above also provide useful information about the studied conditions and the treatment benefits.

Monitoring outcomes in clinical practice

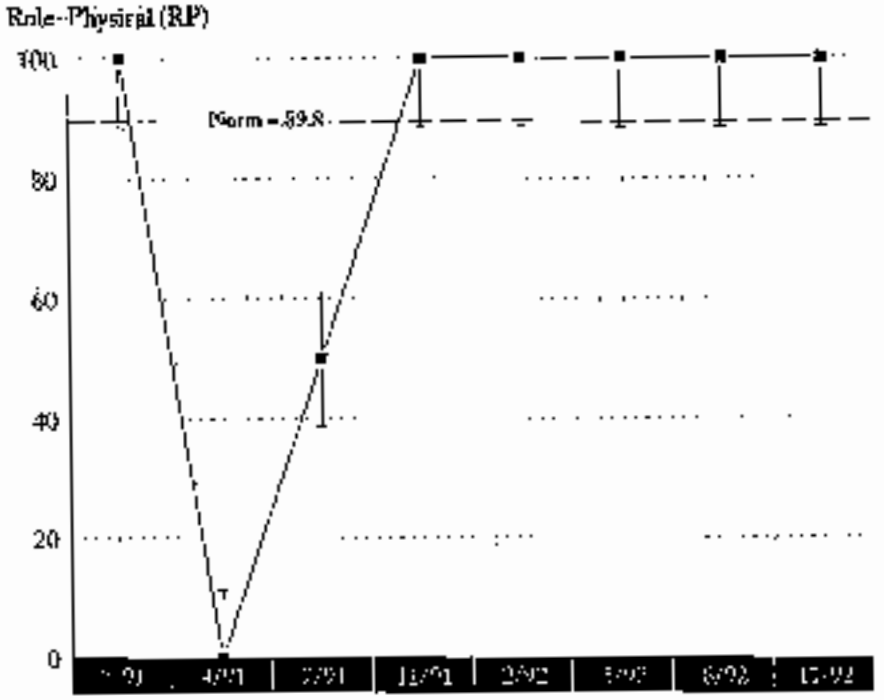
The SF-36 and other patient-based instruments have the potential to serve as "laboratory tests" of functioning and well-being in everyday medical practice (ACP, 1988). Their routine administration would be useful in: detecting and explaining decreased functional capacity and well-being, keeping track of changes in function over time, making it possible to consider the patient's total functioning in choosing among therapies, guiding the efficient use of community resources and social services, and predicting more accurately the course of chronic disease.

Three years ago the New England Medical Center (NEMC) Quality Assessment department began working with several outpatient clinics to develop patient-based systems for monitoring and improving health outcomes (Kurtin et al., 1992; Kantz et al., 1992). One project launched quarterly administrations of the SF-36 to expand the definition of the "adequacy" or "quality" of dialysis beyond traditional laboratory test values (Kurtin et al., 1992).

Figure 11.5 plots the results of eight quarterly administrations of the SF-36 for one patient over a 2-year period that included his sixth and seventh years on dialysis. The patient is a middle-aged married male and an employed parent. He completed SF-36 forms at the time of regularly scheduled outpatient visits for hemodialysis. The solid horizontal line in each panel defines the "norm" for a general population male of the same age who is free of chronic conditions. That normative level varies across the four scales: RP (89.8), BP (79.4), VT (65.5), and SF (88.5). (Although the project routinely scores all eight SF-36 scales, available space limits Figure 11.5 to showing results for four scales.)

Initial scores (1/91) were at or above the norm for six of eight scales, all except VT and PF (data not shown for the latter). In conjunction with an adverse medical event, scores declined dramatically by the end of the first 3-month period (4/91), most notably for the BP, RP, and SF scales. Recovery to levels equal to the general population norm was observed for the BP scale by the third observation period (7/91), followed by recovery to the norm for the RP and SF scales by the fourth period (11/91). As the third panel of Figure 11.5 shows, the patient scored consistently below the national norm for VT, a pattern often observed for patients with chronic renal failure. Longitudinal

FIGURE 11.5
REPEATED MEASURES OF A SINGLE PATIENT



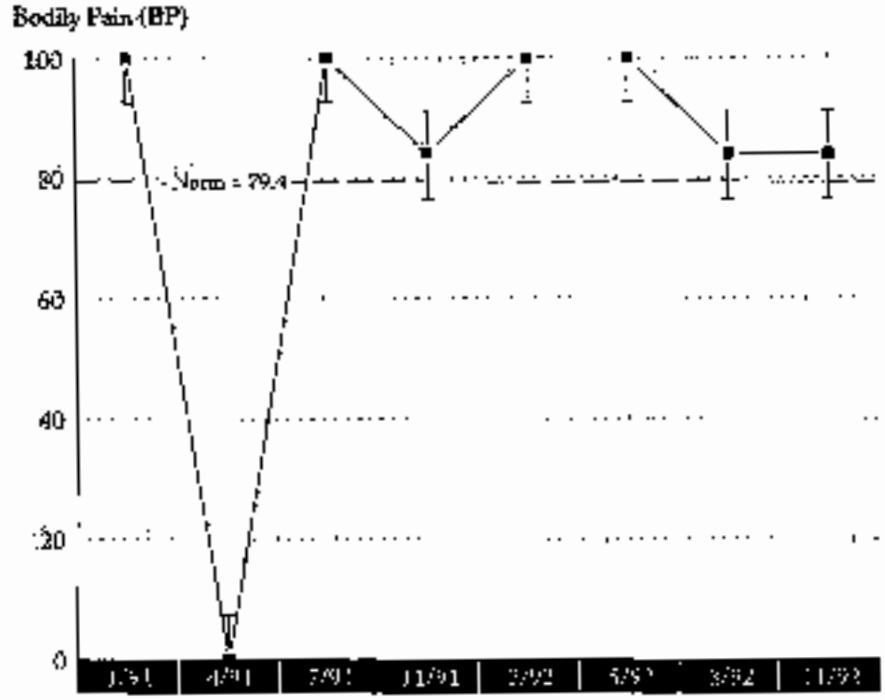
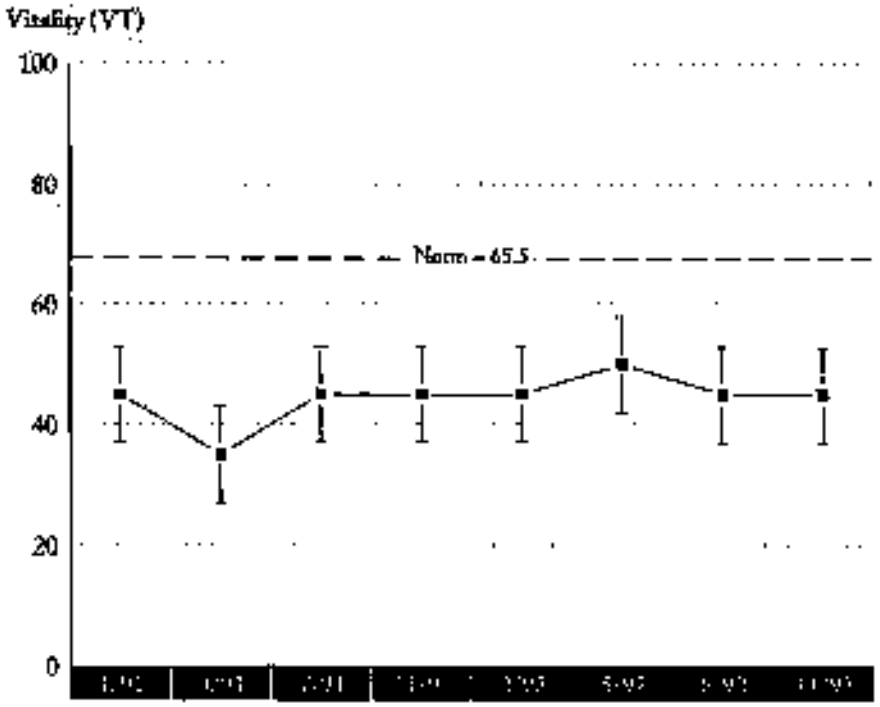
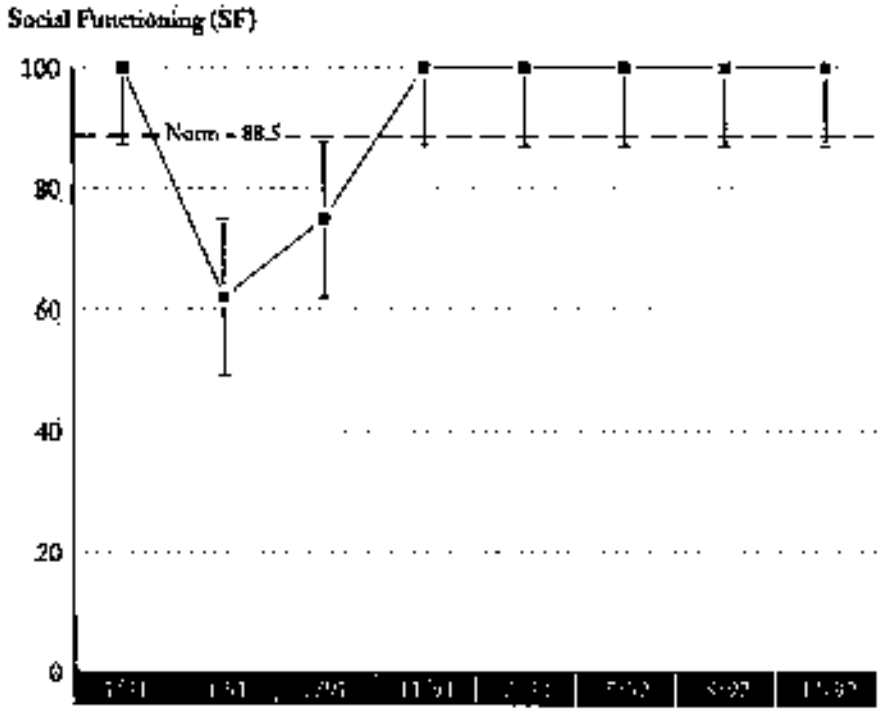


FIGURE 11.5
CONTINUED





monitoring revealed that this status remained stable throughout most of the observation period.

During the 2-year observation period, the patient showed: stable physical functioning (PF scale) below the norm (data not shown); health perceptions (GH scale) at or below the norm with noticeable drops during the third, fifth, and ninth periods (data not shown); RE scale scores consistently equal to or above the norm (data not shown); and stable mental health (MH scale) scores above the norm except for a drop during the second observation period (data not shown).

These observations illustrate several important lessons. First, general population norms are useful in understanding individual patient scores. In this case, the patient consistently scored below the norm for physical health but above the norm for mental health. Second, this example calls attention to noteworthy differences in the size of confidence intervals (CI) around individual patient scores across the eight SF-36 scales. CIs are much larger for the RP and SF scales relative to the other scales. These differences reflect variations in the number of levels defined by each scale and the reliability with which scale levels are measured (see Chapter 7). This example also illustrates the value of establishing a personal norm of functioning and well-being for each individual patient. That norm was consistently below the national norm for some concepts (e.g., the VT scale) and above the national norm for others (e.g., the MH scale). Changes over time for an individual patient may be best judged in relation to what is "normal" for that patient.

Finally, our work with the Quality Assessment Department and the Department of Medicine at NEMC has taught us that it takes time to convert patient-based assessments of health into information that is clinically useful. This conversion occurs when the following becomes routine: (1) processing of assessments with minimal staff disruption, (2) placement of SF-36 profiles in the medical record, and (3) review of profiles by staff.

RT-2000 Processing System for the SF-36

Along with using more practical instruments like the SF-36, routine health assessments in clinical practice are facilitated by systems that rapidly process survey forms with a high degree of reproducibility and at a low cost. The RT-2000 is a unique system that combines a scanner originally developed

for lotteries, a standard computer with a hard drive containing SF-36 software; and a printer like that used in a fax machine. The RT-2000 scans and scores SF-36 forms and prints a standard profile in a matter of seconds.

Figure 11.6 shows the printout from the RT-2000 processing system for three repeated administrations of the SF-36 labeled "Initial" (I), "Previous" (P), and "Current" (C). The bar graphs and scores printed for the eight scales reveal that this patient improved substantially since the two previous administrations. The "limitations grid" makes clinically and socially relevant scores easy for providers to detect. Whereas substantial limitations were apparent in all four areas at the time of previous administrations, none are apparent for the current administration. Accordingly, the patient reported that his health is much better now than it was a year ago.

The RT-2000 printout also summarizes data quality. Specifically, it informs the provider that the Standard SF-36 was used for all three administrations, that the overall data quality was excellent or satisfactory across administrations, that 94 to 100% of the items were complete, and that the Response Consistency Index (see Chapter 7) yielded 100% consistency scores for all administrations. This system is currently being evaluated by various clinical services at NEMC and is in use at more than 12 health care facilities throughout the United States. Other systems are discussed in Chapter 12.

With advances in processing systems for the SF-36, a doctor and patient can monitor the patient's functioning and well-being on a regular basis, inexpensively, and without delay. Such systems remove the practical barriers to monitoring patient health outcomes routinely and on a large scale.

Final comment on applications of the SF-36

Debate about the uses of health assessment methods is spreading beyond the arcane realm of methodologists (Ware, 1990c; 1993). Policy analysts and health care managers — intent on getting the best value for their dollar — have joined the intellectual fray. Clinical investigators evaluating new treatments and technologies and practicing clinicians seeking better patient outcomes are also demanding useful assessment methods. Methodological developments have become newsworthy (Winslow, 1992).

FIGURE 11.6
EXAMPLE OF RT-2000 PRINTOUT FOR THE SF-36 HEALTH SURVEY

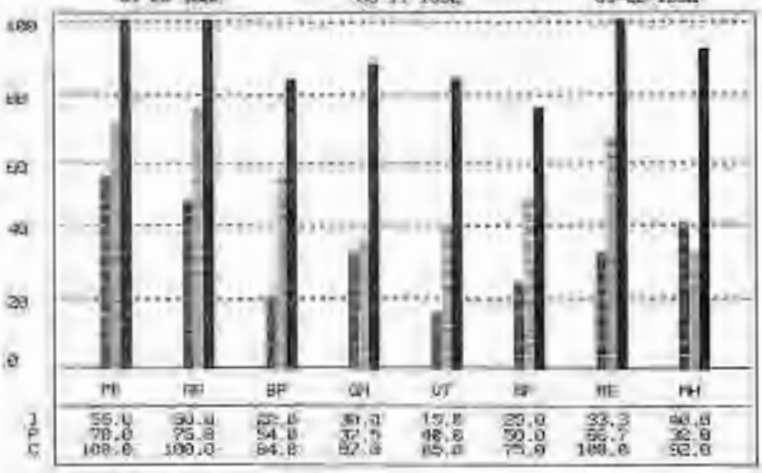
MOS SF-36™ HEALTH SURVEY

SITE: 1 DATE: 08-28-1992
ID: 3574221450 SEX: Male AGE: 45-54

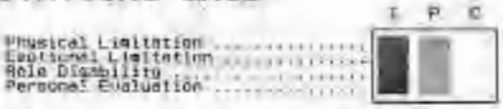
HEALTH SCORES

Physical Functioning (PF) Vitality (VT)
Role Physical (RP) Social Functioning (SF)
Bodily Pain (BP) Role Emotional (RE)
General Health (GH) Mental Health (MH)

■ = INITIAL (I) 01-20-1990 ■ = PREVIOUS (P) 06-11-1992 ■ = CURRENT (C) 08-28-1992



LIMITATIONS GRID



REPORTED CHANGE IN HEALTH

DATA QUALITY

	INITIAL 01-20-1990	PREVIOUS 06-11-1992	CURRENT 08-28-1992
VERSION USED	STANDARD	STANDARD	STANDARD
OVERALL QUALITY	EXCELLENT	SATISFACT	EXCELLENT
ITEMS COMPLETE (%)	100.0	97.2	100.0
CONSISTENCY OF RESPONSES (%)	100.0	100.0	100.0

Despite advances in measurement tools, the current state of health care monitoring is woefully deficient. 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. When it comes to health concepts, definitions, and measures, federal health agencies function independently and without coordination.

To meet the needs of the 1990s, information about general health outcomes must be added to the nation's health care database. Minimum standards of comprehensiveness should be adopted to monitor the health of the general population and evaluate health care policies. A core set of generic health outcomes measures 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 special needs of a given study. The resulting comparisons would greatly advance understanding of the interpretation of health measures for all applications. Adoption of a standardized core set of health measures should be a high priority.

In the 1990s, the SF-36 and other general health assessments will be used to discover whether health care cost containment strategies have harmful health effects. Unless the assessments relied upon achieve a high degree of precision, relevant effects may be missed. To make sure this does not occur, evaluations must be precise enough to rule out adverse health outcomes. In summary, health outcome assessment methods have come a long way but have not yet begun to be applied to their full potential.

12. FUTURE DIRECTIONS

Experience in constructing and evaluating the SF-36 Health Survey suggests that trade-offs between the breadth and depth of measurement can be managed successfully. To achieve *breadth*, we included measures believed to be most likely to detect important differences in health status, defined as physical and mental functioning and well-being, social and role disability, and general health perceptions (Ware & Sherbourne, 1992). There appears to be a growing consensus in the field that these concepts are necessary to satisfy a *minimum* standard for a *comprehensive* general health survey. To achieve *depth* of measurement in each conceptual area, we constructed a very short multi-item scale from items that best reproduced a lengthier measure of proven validity.

The SF-36 is proof that efficient health status surveys can be constructed, that they are useful for many purposes, and that short-form surveys lead to more widespread adoption of patient-based outcome assessment methods. Hopefully, there will be more efforts of this type. In this chapter we discuss some future directions in this field and some of the factors that are likely to influence progress.

Standardization

Standardization of *content* and *scoring* are essential to the interpretation of the SF-36 and for meaningful comparisons of results across studies. For these reasons we recommend adhering to the standard content of the SF-36 form and standard scoring, and using the "SF-36" label only in relation to these standards. By limiting use of the "SF-36" label to standard forms and scoring algorithms, it will be clear when data can and cannot be compared.

Medical Outcomes Trust

To protect the standardization of the SF-36 and assure widespread availability, the copyright for the SF-36 form and the trademark for the SF-36 label were placed with the Medical Outcomes Trust, a non-profit public

service trust. The Trust has the objectives of: (1) ensuring the availability of the SF-36, scoring information, and other documentation useful for interpretation, free of charge; (2) protecting the standardization of the content, scoring, and labeling of the SF-36 and other instruments to ensure their scientific validity and comparability across studies and countries; and (3) serving as a clearinghouse to facilitate the exchange of information and promote networking among users to advance state-of-the-art health assessment in general. Information is available from: The Medical Outcomes Trust, PO Box 1917, Boston, MA 02205.

Modifications

We encourage those who modify either the content of the SF-36 form or the scoring algorithms to study the effect on scores relative to those for the Standard form and scoring and to report their findings in the scientific literature (see Chapters 6 and 10). By doing so they will advance understanding of the consequences of their modifications and will make it possible for others to determine whether their modifications are worth the loss of comparability with the SF-36 database.

Use with disease-specific measures

The best solution to the choice between general health measures and traditional biomedical (i.e., disease-specific) measures of health outcomes appears to be the synchronous measurement of both (McHorney et al., 1992; Patrick & Deyo, 1989; Ware, 1984a). Although measures of general health represent patient health values better than biological measures of traditional clinical endpoints, the greatest progress will occur not from substituting one for the other, but from assessing and working to improve both kinds of outcomes in concert.

An important agenda for the future is the development and testing of assessment modules for specific medical and psychiatric conditions, such as those being tested by the AHCPR-sponsored Patient Outcomes Research Teams (PORTS) in the United States. Modules containing both generic health measures and disease-specific measures of symptoms and problems are also being tested in numerous clinical trials in the United States and in other

countries. Evaluation of disease-specific measures, generic core measures, and supplemental generic measures will lead to the development of measures that are most sensitive to the burden of each disease and to the benefits of treatment (see Chapter 11, "Clinical Trials of Treatment Effects").

An Empirical Comparison

It is often assumed that condition-specific measures are both more sensitive and more specific to the target condition. To test this assumption, Kantz and her colleagues analyzed the sensitivity and specificity of the SF-36, three condition-specific adaptations of the SF-36, and the Knee Society's Clinical Rating System in a study of osteoarthritis patients (Kantz et al., 1992). As expected, knee-specific role function and pain scales were more *specific* than the generic SF-36 scales among patients with other comorbid conditions. The SF-36 Physical Functioning (PF) scale and knee-specific adaptation of that scale were equally specific among patients with only knee problems. These patterns were observed for measures of variations in knee status and effects of treatment (knee replacement). However, the condition-specific adaptations were no more *sensitive* than the three SF-36 scales to variations in clinical measures of knee function. Finally, the condition-specific adaptations failed to discriminate between treated and untreated knees.

As expected, condition-specific adaptations tended to be more sensitive to variations in clinical status when *comorbidities were present.* Generic measures have been shown to be sensitive to the *combined* effects of conditions when multiple conditions are present (Stewart et al., 1989). It follows that condition-specific measures should reflect specific effects better than generic measures. However, the comparison between the SF-36 PF and the condition-specific adaptation for physical function did not confirm this.

Kantz and her colleagues offered two conclusions for assessing patients undergoing total knee replacement (TKR):

> First, for assessing knee pain and knee-related limitations in role function, measures that refer explicitly to the knee provide greater specificity. Second, comprehensive assessment of the health of TKR patients requires both generic and condition-specific measures of key health concepts, including physical function, role limitations, and pain. We believe a combined approach, using both generic and knee-specific measures, is likely to prove

best for investigations of TKR and related knee problems. We expect that these conclusions would generalize to other similar clinical conditions marked by changes in lower extremity function (Kantz et al., 1992).

Because it is short, well validated, and well normed, the SF-36 may constitute a good "generic core" for use in studies that include patient-based assessments of both generic and disease-specific health outcomes.

Health indexes

Medical Outcomes Study (MOS) projects have also led to the development of Physical and Mental Health Summary Indexes for the SF-36. These indexes are based on the physical and mental health components (factors) shown to account for about 82% of the reliable variance in scale scores. Further, factor loadings (weights) for each SF-36 scale for these two components, as evaluated using clinical criteria (see Chapter 9 and McHorney et al., 1993), go hand-in-hand with their validity.

The SF-36 Summary Indexes offer a number of advantages, including: (1) a reduction in the number of statistical comparisons necessary to capture differences in health status or health outcome, (2) adjustment for substantial correlations among the eight scales, (3) more straightforward interpretation of differences in physical and mental health scores, and (4) substantially greater precision for *general* physical and mental health outcomes. We are evaluating other strategies for constructing summary health indexes including a Survival Probability Index, a Health Utility Index, and a Consumption Index. Others have experimented with a summary measure based on a simple summation of the eight SF-36 scales (Katz et al., 1992), which we do not recommend. MOS studies to date indicate that this summary index is: (1) less sensitive to differences in health status (compared with SF-36 scales and Summary Indexes), and (2) not specific (i.e., not interpretable) in relation to an underlying health concept. Consistent with MOS findings, Katz and his colleagues reported their summary index was less sensitive than other scales (e.g., the two-item SF-36 Bodily Pain [BP] scale) to improvements in health following hip surgery. They also observed that their summary score for the SF-36 scales was no worse than other summary measures studied, including those for the SIP, sAIMS, and FSQ. The latter results may reflect problems with summary indexes in general.

Scoring advances

The scoring algorithms developed for the SF-36, as well as those for MOS "parent" (Stewart & Ware, 1992) and Health Insurance Experiment (HIE) "grandparent" measures (Ware, Brook, et al., 1980) were developed to be as simple as possible while satisfying assumptions underlying scaling methods. Although computers make complicated algorithms as easy to score as simple ones, simple algorithms are easier to understand, more likely to be reproduced without error, and are more readily accepted. The assumptions underlying simple algorithms are also more likely to be satisfied across populations (Armor, 1974; Ware, Brook, et al., 1980).

We are presently evaluating several potential improvements in the scoring of the SF-36, including: (1) improvements in the enumeration of scale levels, (2) construction of aggregate (summary) health indexes, and (3) norm-based scoring of scales and summary indexes. The most important advance in the scoring of the SF-36 is likely to be norm-based scoring algorithms that will be published later in 1993. Norm-based scoring will greatly improve the interpretability of the SF-36 Profile and summary indexes. With norm-based scoring, comparisons with tables of norms will not be necessary for many purposes because normative status will be apparent directly from the scale scores.

Recalibration of Items

Recalibrations that improve the linear fit between item and scale scores, a basic assumption of the SF-36 scaling method, have already been adopted for items in the BP and General Health (GH) scales (see Chapter 6). Such recalibrations are necessary when different numbers and different types of response choices are offered for items in the same scale, as they are for the BP and GH scales. From work in progress, we do not expect other recalibrations of item response choices to produce noteworthy improvements.

Additional Items

Addition of one item each to the Role-Physical (RP) and Role-Emotional (RE) scales greatly improves those scales by lowering the "floor" and thereby increasing precision. These additions do not sacrifice comparability because both standard and "enhanced" RP and RE scales and scoring can be analyzed

and reported in parallel. (Information about these two items and two other items in the "enhanced" version of the SF-36 will be included in a future *SF-36 Update* sent to those on the user mailing list.)

Item Response Scales

Experiments with categorical ratings rather than dichotomous response choices for items in the RP and RE scales have produced substantial gains in precision. Unfortunately, studies have not yet revealed a way to take advantage of these gains without sacrificing comparability with SF-36 norms. Hence, these and other advances that are achieved at the expense of comparability may best await a new short form that will follow the SF-36.

Scaling Methods

Improved methods for aggregating items within SF-36 scales are also on the horizon. For example, different permutations of responses to items in the same scale, which currently receive the same scale score, may define different levels of disability (see Chapter 9). This phenomenon appears to hold more true in the RP scale than for any other SF-36 scale. Currently advances in the scoring (weighting) of those response patterns are being evaluated. These algorithms can be adapted without loss of comparability if scored in parallel with the standard algorithms.

Scoring of the PF scale has also been improved by using the Rasch scoring method (Haley et al., in review). These gains may be particularly important in studies of respondents who score predominantly at the highest and lowest scale levels. The physical functioning of those who score just below the highest level of the PF scale appears to be overestimated by the current scoring method. An analogous problem has been identified for those who score just above the bottom of the PF scale. However, we do not recommend widespread adoption of the improved scoring pending successful peer review. It may then be best to analyze standard and improved PF scale scores in parallel to maintain comparability.

Translations for use in other countries

An international team of 15 investigators has been developing and evaluating translations of the SF-36 over the past 2 years for the International Quality of Life Assessment (IQOLA) Project (Aaronson et al., 1992; Gandek, 1992). The goal of the IQOLA Project is to culturally adapt, translate, validate, and norm the SF-36 for use in Argentina, Australia, Belgium,* Canada,* Denmark, France, Germany, Italy, Japan, the Netherlands, Norway, Spain, Sweden, the United Kingdom, and the United States (including the Mexican-American version) (*two languages per country).

Although only the U.S. and U.K. versions of the SF-36 have been published to date, IQOLA studies in progress have demonstrated the feasibility of achieving valid translations. These studies underscore the importance of careful adherence to proven procedures of translation and linguistic evaluation, as well as rigorous psychometric and clinical validation of forms to be used in multiple languages.

Within each country, the work is directed by an experienced senior investigator who is a native of that country. Each country follows a standard procedure of forward and backward translation, review by representative focus groups, and formal evaluation of the quality of each translation. Data from clinical trials and other studies are being used to evaluate the translated forms, using standard psychometric techniques of reliability and validity assessment, including empirical tests of validity in relation to clinical status. The measures are being refined in an iterative cycle to enhance their quality and to evaluate cross-cultural empirical validity.

Additional IQOLA Project goals include: collection of general population norms for each country and preparation of user's manuals and computerized scoring software that document how to administer, score, analyze, and interpret SF-36 scales and summary indexes in each country.

Sponsors of the project (including Glaxo Research Inc. and Schering-Plough Corporation) and participating scientists have agreed to place all completed translations and the U.K. version with the Medical Outcomes Trust. This will ensure availability without charge and will protect their standardization.

Data processing systems

Advances in systems for administering, processing, and scoring SF-36 scales have arrived, and more are on the horizon. Improved formats for forms that can be processed using either traditional key punch or scanning methods are available. Both "mark sense" scannable forms and optical character reading (OCR) methods have been used successfully. For example, the large NIH-sponsored Breast Cancer Prevention Trial is using the scannable SF-36 form (reproduced in Appendix B).

Another promising approach is the SF-36 Fax form (reproduced in Appendix B) which can be transmitted by any fax machine for centralized processing. This methodology is currently being used in the NIH-sponsored Prostate Cancer Prevention Trial. (Response Technologies in East Greenwich, Rhode Island, offers forms and processing services for these and other SF-36 formats.)

Response Technologies has also developed a self-contained processing system for the SF-36 now in use by a dozen health care delivery organizations, including numerous clinical services at NEMC. Their RT-2000 "box" includes a scanner originally developed for lotteries, an IBM-compatible computer processor with SF-36 software, and a fax-quality printer. This system requires less than 10 seconds to process a form and print out numerical results and profiles, and costs less than one dollar per administration in a busy practice.

Computer-based technologies we are aware of include interactive systems that administer and score the SF-36 on standard PCs and on notebook and notepad computers. The latter include an "electric pen" notepad system available from Response Technologies. One advantage of computer-based systems is that the data quality checks the SF-36 software performs (see Chapter 7) are performed in real time so that problems can be identified and corrected while the respondent is still available. It is likely that advances in these and other technologies will greatly reduce the costs of administering the SF-36, guarantee the reproducibility of results, and make even the most user-friendly displays of results available to health care providers and their patients almost instantaneously.

Final comments

The cycle of defining and measuring health begins with a conceptual framework, which leads to specific operational definitions such as those represented by the SF-36 Health Survey. For the past two decades the field has been simultaneously addressing issues about the meaning of health, searching for its major and minor components, exploring their interrelationships, and trying to apply the results to improving measurement strategies.

We still have much to learn about the nature of health, what health concepts are valued most, how to measure those concepts, and the extent to which conceptual and methodological lessons learned in one study are generalizable to another. Psychometric theory, used by the psychological testing movement in Europe and the United States for over 100 years, appears to ideally suit the task of conceptualizing and measuring health (Stewart & Ware, 1992).

Dissemination and Adoption

The MOS model of health and psychometric theory, which guided the construction of the SF-36, is now being applied and evaluated in Western Europe, Asia, and Australia (Aaronson et al., 1992; Garratt et al., 1993; Jenkinson et al., 1993). Since publication of the SF-36 in June 1992, 7,205 user information packets have been distributed. The rate of requests to use the SF-36 continues to increase; 2,968 requests have been received and processed since January 1993. Numerous provider organizations have established longitudinal monitoring systems to evaluate the feasibility of inpatient and outpatient assessments based on the SF-36. Within a year we expect to have preliminary results from psychometric evaluations of SF-36 scales from over 100 randomized trials and clinical observational studies. As of June 1993, the SF-36 user database maintained by The Medical Outcomes Trust included 260 such studies involving 158 different topics (see Chapter 11). With the widespread adoption of the SF-36 as a "core" generic health measure in clinical trials, meta-analyses of treatment effects will be possible on a larger scale and with a greater degree of standardization of definitions of health outcomes than ever before possible. With the permission and cooperation of the investigators involved, the Medical Outcomes Trust plans to report summaries of the characteristics of these studies (goals, size, duration, etc.) and other information in its database.

Next Steps

Priority should be placed on research that makes the SF-36 and other measurement methods easier to use and interpret. Since research findings may apply to use in one application and not another, research should focus on specific uses of measurement methods. Manuals, interpretation guides, and other supporting documents that are user-friendly should be made available to ensure their proper use and interpretation. We hope that this manual will help to meet these needs for users of the SF-36.

Health outcome assessment methods have come a long way, but they are not being used to their full potential. To keep pace with the "outcomes movement" in health care, two paths must be followed in tandem (1) methodologists must continue to advance the state-of-the-art of patient-based health outcomes assessment, and (2) health care decision-makers must use the best available assessment methods to identify the interventions most useful in improving health outcomes. In this imperfect world, decision-makers should not wait for methodologists to perfect their craft. To wait would be to hold progress hostage to perfection.

Our work in the HIE, the MOS, and the IQOLA Project is based on the assumption that one of the best ways to gain a better understanding of health is to try to measure it. The challenge of the future is to master understanding of how to *improve* health. Can better measures and better data lead to better outcomes? Time will tell.

APPENDIX A: ADDITIONAL NORMS: FREQUENCY DISTRIBUTION FOR TRANSFORMED SCALE SCORES

This appendix presents the frequency, percent, and cumulative percent of respondents in the general U.S. population at each observed score for each SF-36 scale (N=2,474). See Chapter 10 for details.

TABLE A.1 PHYSICAL FUNCTIONING SCALE (PF)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	952	38.8	100.0
95.00	408	16.7	61.2
94.54	2	0.1	44.6
92.86	1	0.0	44.5
90.00	236	9.6	44.5
88.89	8	0.3	34.8
87.50	1	0.0	34.5
85.72	1	0.0	34.5
85.00	163	6.7	34.4
83.34	4	0.2	27.8
81.25	1	0.0	27.6
80.00	107	4.4	27.6
77.78	4	0.2	23.2
75.00	72	3.0	23.1
72.23	2	0.1	20.1
71.49	1	0.0	20.0
70.00	62	2.5	20.0
68.75	1	0.0	17.5
66.67	2	0.1	17.4
65.00	59	2.4	17.4
64.26	1	0.0	15.0
62.50	1	0.0	15.0
61.12	3	0.1	14.9
60.00	37	1.5	14.8
58.34	1	0.0	13.3
57.14	1	0.0	13.3
56.25	1	0.0	13.2
55.56	1	0.0	13.2

continued on next page

TABLE A.1 CONTINUED

Transformed Score	Frequency	Percent	Cumulative Percent
55.00	31	1.3	13.2
50.00	44	1.8	11.9
45.00	30	1.2	10.1
44.45	2	0.1	8.9
41.67	1	0.0	8.8
40.00	32	1.3	8.8
38.89	2	0.1	7.4
37.50	1	0.0	7.4
35.00	31	1.3	7.3
33.34	1	0.0	6.0
31.25	1	0.0	6.0
30.00	25	1.1	5.9
27.78	1	0.0	4.9
25.00	25	1.0	4.9
22.23	1	0.0	3.8
20.00	18	0.8	3.8
16.67	1	0.0	3.0
15.00	20	0.8	3.0
12.50	1	0.0	2.2
10.00	16	0.6	2.2
5.56	1	0.0	1.5
5.00	16	0.7	1.5
0.00	21	0.8	0.8
	2454		

Frequency Missing = 20

TABLE A.2 ROLE-PHYSICAL SCALE (RP)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	1737	70.9	100.0
75.00	200	8.1	29.1
66.70	1	0.0	21.0
50.00	127	5.2	21.0
33.30	1	0.1	15.8
25.00	131	5.4	15.7
0.00	253	10.3	10.3
	2450		

Frequency Missing = 24

TABLE A.3 BODILY PAIN SCALE (BP)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	787	31.9	100.0
90.00	3	0.1	68.1
88.00	3	0.1	68.0
84.00	462	18.7	67.9
80.00	26	1.1	49.2
74.00	202	8.2	48.2
72.00	177	7.2	40.0
70.00	2	0.1	32.8
67.50	2	0.1	32.8
64.00	21	0.8	32.7
62.00	228	9.2	31.8
61.00	47	1.9	22.6
60.00	4	0.2	20.7
54.00	4	0.2	20.5
52.00	57	2.3	20.4
51.00	102	4.1	18.0
50.00	1	0.1	13.9
45.00	1	0.1	13.9
44.00	1	0.0	13.8
42.00	29	1.2	13.8
41.00	121	4.9	12.6
40.00	3	0.1	7.7
32.00	20	0.8	7.6
31.00	47	1.9	6.8
30.00	3	0.1	4.9
24.00	1	0.0	4.7
22.50	2	0.1	4.7
22.00	67	2.7	4.6
21.00	6	0.3	1.9
20.00	1	0.1	1.6
12.00	17	0.7	1.6
10.00	7	0.3	0.8
0.00	24	0.6	0.6
	2468		

Frequency Missing = 6

TABLE A.4 GENERAL HEALTH SCALE (GH)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	181	7.4	100.0
97.00	82	3.3	92.6
95.00	76	3.1	89.3
92.00	87	3.5	86.2
90.00	77	3.1	82.6
87.00	173	7.1	79.5
85.00	66	2.7	72.4
82.00	234	9.5	69.8
81.25	3	0.1	60.3
80.00	48	2.0	60.1
77.50	2	0.1	58.1
77.00	223	9.1	58.0
75.00	33	1.4	48.9
72.00	228	9.3	47.6
71.25	2	0.1	38.3
70.00	15	0.6	38.2
68.75	1	0.0	37.6
67.00	177	7.2	37.6
65.00	30	1.2	30.3
62.50	1	0.1	29.1
62.00	148	6.0	29.1
61.67	1	0.0	23.0
60.00	24	1.0	23.0
58.75	5	0.2	22.0
58.33	1	0.0	21.8
57.00	88	3.6	21.8
56.25	1	0.0	18.2
55.00	24	1.0	18.2
52.00	70	2.8	17.2
50.00	20	0.8	14.3
47.00	30	1.2	13.5
45.00	41	1.7	12.3
43.75	3	0.1	10.6
42.00	23	0.9	10.5
40.00	33	1.3	9.5
37.50	1	0.1	8.2

continued on next page

TABLE A.4 CONTINUED

Transformed Score	Frequency	Percent	Cumulative Percent
37.00	20	0.8	8.1
35.00	34	1.4	7.3
33.33	1	0.0	5.9
32.00	6	0.2	5.9
31.25	1	0.0	5.7
30.00	35	1.5	5.6
27.00	2	0.1	4.2
25.00	36	1.5	4.1
22.00	3	0.1	2.6
20.00	25	1.0	2.5
18.75	1	0.0	1.5
16.67	1	0.0	1.5
15.00	11	0.4	1.4
12.50	1	0.0	1.0
10.00	19	0.8	1.0
6.25	1	0.0	0.2
5.00	4	0.2	0.2
	2453		

Frequency Missing = 21

TABLE A.5 VITALITY SCALE (VT)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	37	1.5	100.0
95.00	33	1.3	98.5
93.33	1	0.0	97.2
90.00	86	3.5	97.2
86.67	4	0.2	93.7
85.00	134	5.5	93.5
80.00	309	12.6	88.0
75.00	231	9.4	75.5
73.33	1	0.0	66.1
70.00	279	11.4	66.0
65.00	170	6.9	54.7
60.00	206	8.4	47.8
55.00	193	7.9	39.4
53.33	2	0.1	31.5
50.00	205	8.3	31.5
46.67	1	0.0	23.1
45.00	101	4.1	23.1
40.00	113	4.6	19.0
35.00	75	3.1	14.4
33.33	2	0.1	11.3
30.00	66	2.7	11.2
25.00	59	2.4	8.5
20.00	53	2.2	6.1
15.00	31	1.3	3.9
10.00	34	1.4	2.7
5.00	18	0.8	1.3
0.00	13	0.5	0.5
	2457		

Frequency Missing = 17

TABLE A.6 SOCIAL FUNCTIONING SCALE (SF)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	1293	52.3	100.0
87.50	307	12.4	47.7
75.00	327	13.2	35.3
62.50	153	6.2	2.0
50.00	216	8.7	15.8
37.50	86	3.5	7.1
25.00	50	2.0	3.6
12.50	24	.9	1.6
0.00	16	0.6	0.6
	2472		

Frequency Missing = 7

TABLE A.7 ROLE-EMOTIONAL SCALE (RE)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	1735	71.0	100.0
66.67	275	11.3	29.0
50.00	5	0.2	17.7
33.33	194	7.9	17.5
0.00	235	9.6	9.6
	2444		

Frequency Missing = 30

TABLE A.8 MENTAL HEALTH SCALE (MH)

Transformed Score	Frequency	Percent	Cumulative Percent
100.00	97	3.9	100.0
96.00	114	4.6	96.1
92.00	281	11.4	91.4
90.00	1	0.0	80.0
88.00	268	10.9	80.0
86.67	1	0.0	69.1
85.00	2	0.1	69.0
84.00	282	11.5	69.0
80.00	266	10.8	57.5
76.00	188	7.6	46.7
75.00	1	0.0	39.1
72.00	169	6.9	39.0
70.00	2	0.1	32.1
68.00	145	5.9	32.1
66.67	1	0.0	26.2
65.00	2	0.1	26.2
64.00	105	4.3	26.1
60.00	120	4.9	21.8
56.00	85	3.5	16.9
55.00	2	0.1	13.5
52.00	64	2.6	13.4
50.00	2	0.1	10.7
48.00	54	2.2	10.7
45.00	1	0.0	8.5
44.00	47	1.9	8.4
40.00	34	1.4	6.5
36.00	36	1.5	5.1
33.33	1	0.0	3.6
32.00	23	0.9	3.6
30.00	1	0.0	2.7
28.00	19	0.8	2.7
25.00	1	0.0	1.9
24.00	14	0.6	1.8
20.00	17	0.7	1.3
16.00	5	0.2	0.6
12.00	3	0.1	0.4
8.00	2	0.1	0.2
4.00	2	0.1	0.1
0.00	1	0.0	0.0
	2459		

Frequency Missing = 15

APPENDIX B: COPIES OF SF-36 FORMS

STANDARD SF-36, BOOKLET FORM – PAGE ONE OF FIVE

THE MOS 36-ITEM SHORT-FORM HEALTH SURVEY (SF-36)

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.

Answer every question by marking the answer as indicated. If you are unsure about how to answer a question, please give the best answer you can.

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

(circle one)

Excellent 1

Very good 2

Good 3

Fair 4

Poor 5

2. Compared to one year ago, how would you rate your health in general now?

(circle one)

Much better now than one year ago 1

Somewhat better now than one year ago 2

About the same as one year ago 3

Somewhat worse now than one year ago 4

Much worse now than one year ago 5

STANDARD SF-36, BOOKLET FORM – PAGE TWO OF FIVE

3. The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

(circle one number on each line)

ACTIVITIES	Yes, Limited A Lot	Yes, Limited A Little	No, Not Limited At All
a. **Vigorous activities**, such as running, lifting heavy objects, participating in strenuous sports	1	2	3
b. **Moderate activities**, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
c. Lifting or carrying groceries	1	2	3
d. Climbing **several** flights of stairs	1	2	3
e. Climbing **one** flight of stairs	1	2	3
f. Bending, kneeling, or stooping	1	2	3
g. Walking **more than a mile**	1	2	3
h. Walking **several blocks**	1	2	3
i. Walking **one block**	1	2	3
j. Bathing or dressing yourself	1	2	3

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

(circle one number on each line)

	YES	NO
a. Cut down on the **amount of time** you spent on work or other activities	1	2
b. **Accomplished less** than you would like	1	2
c. Were limited in the **kind** of work or other activities	1	2
d. Had **difficulty** performing the work or other activities (for example, it took extra effort)	1	2

STANDARD SF-36, BOOKLET FORM – PAGE THREE OF FIVE

5. During this past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

(circle one number on each line)

	YES	NO
a. Cut down the amount of time you spent on work or other activities	1	2
b. Accomplished less than you would like	1	2
c. Didn't do work or other activities as carefully as usual	1	2

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

(circle one)

Not at all 1

Slightly 2

Moderately 3

Quite a bit 4

Extremely 5

7. How much bodily pain have you had during the past 4 weeks?

(circle one)

None 1

Very mild 2

Mild 3

Moderate 4

Severe 5

Very severe 6

STANDARD SF-36, BOOKLET FORM – PAGE FOUR OF FIVE

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

(circle one)

Not at all 1

A little bit 2

Moderately 3

Quite a bit 4

Extremely 5

9. 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 -

(circle one number on each line)

	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
a. Did you feel full of pep?	1	2	3	4	5	6
b. Have you been a very nervous person?	1	2	3	4	5	6
c. Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5	6
d. Have you felt calm and peaceful?	1	2	3	4	5	6
e. Did you have a lot of energy?	1	2	3	4	5	6
f. Have you felt downhearted and blue?	1	2	3	4	5	6
g. Did you feel worn out?	1	2	3	4	5	6
h. Have you been a happy person?	1	2	3	4	5	6
i. Did you feel tired?	1	2	3	4	5	6

STANDARD SF-36, BOOKLET FORM – PAGE FIVE OF FIVE

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

(circle one)

All of the time 1

Most of the time 2

Some of the time 3

A little of the time 4

None of the time 5

11. How TRUE or FALSE is each of the following statements for you?

(circle one number on each line)

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
a. I seem to get sick a little easier than other people	1	2	3	4	5
b. I am as healthy as anybody I know	1	2	3	4	5
c. I expect my health to get worse	1	2	3	4	5
d. My health is excellent	1	2	3	4	5

STANDARD SF-36, RT-2000 SCANNING FORM – FRONT

SF-36™ HEALTH STATUS SURVEY

SIDE 1

MARKING INSTRUCTIONS

- Use a No. 2 Pencil Only.
- Make dark heavy marks that fill the oval completely.
- Erase unwanted marks cleanly.
- Make no stray marks on this survey form.

RIGHT MARK: ⬭ ● ⬭ IMPROPER MARKS: ✓ ✗ ◯

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.

Answer every question by marking the appropriate oval. If you are unsure about how to answer a question, please give the best answer you can.

Before beginning this questionnaire. . .
Please pencil in your ID number in the squares to the right
And then darken in the appropriate oval below each number.

If you don't know what ID number to use, ask the person who gave you this questionnaire.

Now begin with the questions below.

-3. START HERE - Leave out spaces, dashes, etc.

ID NUMBER

1. In general, would you say your health is:
   - 1. ① Excellent
   - ② Very good
   - ③ Good
   - ④ Fair
   - ⑤ Poor

2. **Compared to one year ago,** how would you rate your health in general **now?**
   - 2. ① Much better now than 1 year ago
   - ② Somewhat better now than 1 year ago
   - ③ About the same as 1 year ago
   - ④ Somewhat worse now than 1 year ago
   - ⑤ Much worse now than 1 year ago

3. The following items are about activities you might do during a typical day. **Does your health now limit you in these activities? If so, how much?** (Mark one oval on each line.)

	3. Yes, Limited A Lot	Yes, Limited A Little	No, Not Limited At All
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	a. ①	②	③
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	b. ①	②	③
c. Lifting or carrying groceries	c. ①	②	③
d. Climbing several flights of stairs	d. ①	②	③
e. Climbing one flight of stairs	e. ①	②	③
f. Bending, kneeling, or stooping	f. ①	②	③
g. Walking more than a mile	g. ①	②	③
h. Walking several blocks	h. ①	②	③
i. Walking one block	i. ①	②	③
j. Bathing or dressing yourself	j. ①	②	③

4. During the **past 4 weeks,** have you had any of the following problems with your work or other regular daily activities **as a result of your physical health?** (Mark one oval on each line.)

	4. Yes	No
a. Cut down the amount of time you spent on work or other activities	a. ①	②
b. Accomplished less than you would like	b. ①	②
c. Were limited in the kind of work or other activities	c. ①	②
d. Had difficulty performing the work or other activities (for example, it took extra effort)	d. ①	②

PLEASE TURN CARD OVER TO COMPLETE QUESTIONNAIRE

This is a reduced copy of the scannable SF-36 Standard Form available from Response Technologies, Inc.

STANDARD SF-36, RT-2000 SCANNING FORM – BACK

This is Side 2 of this Questionnaire.
Make sure you complete the OTHER side first.

SIDE 2

5. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? (Mark one oval on each line.)

		Yes	No
a. Cut down the amount of time you spent on work or other activities	5.	a. ◯	◯
b. Accomplished less than you would like		b. ◯	◯
c. Didn't do work or other activities as carefully as usual		c. ◯	◯

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups? (Mark one oval.)

6. ◯ Not at all ◯ Quite a bit
◯ Slightly ◯ Extremely
◯ Moderately

7. How much bodily pain have you had during the past 4 weeks? (Mark one oval.)

7. ◯ None ◯ Moderate
◯ Very mild ◯ Severe
◯ Mild ◯ Very severe

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)? (Mark one oval.)

8. ◯ Not at all ◯ Quite a bit
◯ A little bit ◯ Extremely
◯ Moderately

9. 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 . . . (Mark one oval on each line.)

	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
a. Did you feel full of pep?	a. ◯	◯	◯	◯	◯	◯
b. Have you been a very nervous person?	b. ◯	◯	◯	◯	◯	◯
c. Have you felt so down in the dumps that nothing could cheer you up?	c. ◯	◯	◯	◯	◯	◯
d. Have you felt calm and peaceful?	d. ◯	◯	◯	◯	◯	◯
e. Did you have a lot of energy?	e. ◯	◯	◯	◯	◯	◯
f. Have you felt downhearted and blue?	f. ◯	◯	◯	◯	◯	◯
g. Did you feel worn out?	g. ◯	◯	◯	◯	◯	◯
h. Have you been a happy person?	h. ◯	◯	◯	◯	◯	◯
i. Did you feel tired?	i. ◯	◯	◯	◯	◯	◯

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)? (Mark one oval.)

10. ◯ All of the time ◯ A little of the time
◯ Most of the time ◯ None of the time
◯ Some of the time

11. How true or false is each of the following statements for you? (Mark one oval on each line.)

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
a. I seem to get sick a little easier than other people.	a. ◯	◯	◯	◯	◯
b. I am as healthy as anybody I know.	b. ◯	◯	◯	◯	◯
c. I expect my health to get worse.	c. ◯	◯	◯	◯	◯
d. My health is excellent.	d. ◯	◯	◯	◯	◯

12a. Which are you?

12. a. ◯ Male ◯ Female

b. How old were you on your last birthday?

b. ◯ Less than 35 ◯ 65-74
◯ 35-44 ◯ 75-84
◯ 45-54 ◯ 85 and older
◯ 55-64

13. Have you ever filled out this form before?

13. ◯ Yes
◯ No
◯ Don't remember

Thank you for your time.

14. DO NOT MARK HERE ◯ ◯ ◯ ◯ ◯
NOT LAST CARD ◯

This product form of the SF-36 Health Survey is licensed for use with the RT-2000 Response Terminal made by Response Technologies, Inc. - West Greenwich, RI 02817

This is a reduced copy of the scannable SF-36 Standard Form available from Response Technologies, Inc.

ACUTE SF-36, ST-2000 SCANNING FORM – FRONT

PAGE 1

SF-36™ HEALTH STATUS SURVEY (Acute)

MARKING INSTRUCTIONS

- Use a No. 2 Pencil ONLY
- Make only heavy marks that fill the oval completely.
- Erase unwanted marks cleanly.
- Make no stray marks on this entire sheet.

PROPER MARK: ● WRONG MARKS: ✓ ✗ ○

INSTRUCTIONS: This survey asks for your views about your health now and during the past week. The information will help keep track of how you feel and how well you are able to do your usual activities.

Answer every question by marking the appropriate oval. If you are unsure about how to answer a question, please give the best answer you can.

Before beginning the questionnaire...
Please print in your ID number in the squares to the right and then darken in the appropriate oval below each number.

If you don't know what ID number to use, ask the person who gave you this questionnaire.

Now begin with the questions below.

START HERE – Leave no spaces, dashes, etc.

ID NUMBER

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9

1. In general, would you say your health is:
   - 1. Excellent
   - Very good
   - Good
   - Fair
   - Poor

2. Compared to one week ago, how would you rate your health in general now?
   - 2. Much better now than 1 week ago
   - Somewhat better now than 1 week ago
   - About the same
   - Somewhat worse now than 1 week ago
   - Much worse now than 1 week ago

3. The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? (Mark one oval on each line)

	Yes, Limited A Lot	Yes, Limited A Little	No, Not Limited At All
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	a.		
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	b.		
c. Lifting or carrying groceries	c.		
d. Climbing several flights of stairs	d.		
e. Climbing one flight of stairs	e.		
f. Bending, kneeling, or stooping	f.		
g. Walking more than a mile	g.		
h. Walking several blocks	h.		
i. Walking one block	i.		
j. Bathing or dressing yourself	j.		

4. During the past week, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? (Mark one oval on each line)

	Yes	No
a. Cut down the amount of time you spent on work or other activities	a.	
b. Accomplished less than you would like	b.	
c. Were limited in the kind of work or other activities	c.	
d. Had difficulty performing the work or other activities (for example, it took extra effort)	d.	

PLEASE TURN CARD OVER TO COMPLETE QUESTIONNAIRE

This is a reduced copy of the scannable SF-36 Acute Form available from Response Technologies, Inc.

ACUTE SF-36, RT-2000 SCANNING FORM – BACK

This is Side 2 of this Questionnaire.
Make sure you complete the OTHER side first.

Side 2

5. During the **past week**, have you had any of the following problems with your work or other regular daily activities as a result of **any emotional problems** (such as feeling depressed or anxious)? (Mark one oval on each line.)

	Yes	No
a. Cut down the amount of time you spent on work or other activities	a. ①	②
b. Accomplished less than you would like	b. ①	②
c. Didn't do work or other activities as carefully as usual	c. ①	②

6. During the **past week**, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups? (Mark one oval.)
   6. ① Not at all ② Quite a bit ③ Slightly ④ Extremely ⑤ Moderately

7. How much **bodily** pain have you had during the **past week**? (Mark one oval.)
   7. ① None ② Very mild ③ Mild ④ Moderate ⑤ Severe ⑥ Very severe

8. During the **past week**, how much did **pain** interfere with your normal work (including both work outside the home and housework)? (Mark one oval.)
   8. ① Not at all ② A little bit ③ Moderately ④ Quite a bit ⑤ Extremely

9. These questions are about how you feel and how things have been with you **during the past week**. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past week. . . (Mark one oval on each line.)

	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
a. Did you feel full of pep?	a. ①	②	③	④	⑤	⑥
b. Have you been a very nervous person?	b. ①	②	③	④	⑤	⑥
c. Have you felt so down in the dumps nothing could cheer you up?	c. ①	②	③	④	⑤	⑥
d. Have you felt calm and peaceful?	d. ①	②	③	④	⑤	⑥
e. Did you have a lot of energy?	e. ①	②	③	④	⑤	⑥
f. Have you felt downhearted and blue?	f. ①	②	③	④	⑤	⑥
g. Did you feel worn out?	g. ①	②	③	④	⑤	⑥
h. Have you been a happy person?	h. ①	②	③	④	⑤	⑥
i. Did you feel tired?	i. ①	②	③	④	⑤	⑥

10. During the **past week**, how much of the time has your **physical health or emotional problems** interfered with your social activities (like visiting with friends, relatives, etc.)? (Mark one oval.)
    10. ① All of the time ② Most of the time ③ Some of the time ④ A little of the time ⑤ None of the time

11. Please choose the answer that best describes how **true** or **false** each of the following statements is for you. (Mark one oval on each line.)

	Definitely True	Mostly True	Not Sure	Mostly False	Definitely False
a. I seem to get sick a little easier than other people	a. ①	②	③	④	⑤
b. I am as healthy as anybody I know	b. ①	②	③	④	⑤
c. I expect my health to get worse	c. ①	②	③	④	⑤
d. My health is excellent	d. ①	②	③	④	⑤

12a. Which are you?
12. ① Male ② Female

b. How old were you on your last birthday?
b. ① Less than 35 ② 35-44 ③ 45-54 ④ 55-64 ⑤ 65-74 ⑥ 75-84 ⑦ 85 and older

13. Have you ever filled out this form before?
13. ① Yes ② No ③ Don't remember

Thank you for your time.

14. **DO NOT MARK HERE** ① ② ③ ④ ⑤
NOT LAST CARD ①

This printed form of the SF-36 Health Status Survey is for use on the RT-2000 Response Terminal made by Response Technologies, Inc. - East Greenwich, RI 02818

This is a reduced copy of the scannable SF-36 Acute Form available from Response Technologies, Inc.

STANDARD SF-36, NSABP SCANNING FORM — PAGE ONE OF FOUR

14782 ○○○○○○○○○○■■■○○■■○■■■■■■☒

PERSONAL HEALTH*

***INSTRUCTIONS:** This survey asks for your views about your health.*

Answer every question by marking the answer as indicated. If you are unsure about how to answer a question, please give the best answer you can.

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

- ○ Excellent
- ○ Very good
- ○ Good
- ○ Fair
- ○ Poor

2. *Compared to six months ago*, how would you rate your health in general *now*?

- ○ Much better now than 6 months ago
- ○ Somewhat better now than 6 months ago
- ○ About the same as 6 months ago
- ○ Somewhat worse now than 6 months ago
- ○ Much worse now than 6 months ago

3. The following items are about activities you might do during a typical day. Does *your health now limit you* in these activities? If so, how much?

Activities	Yes, Limited A Lot	Yes, Limited A Little	No, Not Limited At All
a. ***Vigorous activities***, such as running, lifting heavy objects, participating in strenuous sports	○	○	○
b. ***Moderate activities***, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	○	○	○
c. Lifting or carrying groceries	○	○	○
d. Climbing ***several*** flights of stairs	○	○	○
e. Climbing ***one*** flight of stairs	○	○	○
f. Bending, kneeling, or stooping	○	○	○
g. Walking ***more than a mile***	○	○	○
h. Walking ***several blocks***	○	○	○
i. Walking ***one block***	○	○	○
j. Bathing or dressing yourself	○	○	○

■ ■ Twa Bins

This is a reduced copy of the scannable SF-36 Standard Form being used in the NIH-sponsored National Breast Cancer Prevention Trial.

STANDARD SF-36, NSABP SCANNING FORM – PAGE TWO OF FOUR

14782

4. **During the *past 4 weeks*, have you had any of the following problems with your work or other regular daily activities *as a result of your physical health*?**

	YES	NO
a. Cut down the *amount of time* you spent on work or other activities	○	○
b. *Accomplished less* than you would like	○	○
c. Were limited in the *kind* of work or other activities	○	○
d. Had *difficulty* performing the work or other activities (for example, it took extra effort)	○	○

5. **During the *past 4 weeks*, have you had any of the following problems with your work or other regular daily activities *as a result of any emotional problems* (such as feeling depressed or anxious)?**

	YES	NO
a. Cut down the *amount of time* you spent on work or other activities	○	○
b. *Accomplished less* than you would like	○	○
c. Didn't do work or other activities as *carefully* as usual	○	○

Page Five

This is a reduced copy of the scannable SF-36 Standard Form being used in the NIH-sponsored National Breast Cancer Prevention Trial.

STANDARD SF-36, NSABP SCANNING FORM - PAGE THREE OF FOUR

6. During the *past 4 weeks*, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

- ○ Not at all
- ○ Slightly
- ○ Moderately
- ○ Quite a bit
- ○ Extremely

7. How much *bodily* pain have you had during the *past 4 weeks*?

- ○ None
- ○ Very mild
- ○ Mild
- ○ Moderate
- ○ Severe
- ○ Very severe

8. During the *past 4 weeks*, how much did *pain* interfere with your normal work (including both work outside the home and housework)?

- ○ Not at all
- ○ A little bit
- ○ Moderately
- ○ Quite a bit
- ○ Extremely

7x5x3x

This is a reduced copy of the scannable SF-36 Standard Form being used in the NIH-sponsored National Breast Cancer Prevention Trial.

STANDARD SF-36, NSABP SCANNING FORM – PAGE FOUR OF FOUR

9. These questions are about how you feel and how things have been with you during the *past 4 weeks*. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the *past 4 weeks*:

	All of the Time	Most of the Time	A Good Bit of the Time	Some of the Time	A Little of the Time	None of the Time
a. Did you feel full of pep?	○	○	○	○	○	○
b. Have you been a very nervous person?	○	○	○	○	○	○
c. Have you felt so down in the dumps that nothing could cheer you up?	○	○	○	○	○	○
d. Have you felt calm and peaceful?	○	○	○	○	○	○
e. Did you have a lot of energy?	○	○	○	○	○	○
f. Have you felt downhearted and blue?	○	○	○	○	○	○
g. Did you feel worn out?	○	○	○	○	○	○
h. Have you been a happy person?	○	○	○	○	○	○
i. Did you feel tired?	○	○	○	○	○	○

10. During the *past 4 weeks*, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

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

11. How TRUE or FALSE is *each* of the following statements for you?

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
a. I seem to get sick a little easier than other people	○	○	○	○	○
b. I am as healthy as anybody I know	○	○	○	○	○
c. I expect my health to get worse	○	○	○	○	○
d. My health is excellent	○	○	○	○	○

PAGE SEVEN

This is a reduced copy of the scannable SF-36 Standard Form being used in the NIH-sponsored National Breast Cancer Prevention Trial.

STANDARD SF-36, FAX FORM – PAGE ONE OF THREE

RTI

SF-36 Health Status Survey

PAGE 1 OF 3

Please write in your ID number in the boxes below - start in the left box. Write your numbers like this example 0 1 2 3 4 5 6 7 8 9

Shade circles like this: ●
Not like this: ✓ ✗

INSTRUCTIONS: This survey asks for your views about your health. Answer every question by marking the appropriate oval. If you are unsure about how to answer a question, please give the best answer you can.

1. In general, would you say your health is: ○ Excellent ○ Very Good ○ Good ○ Fair ○ Poor

2. Compared to one year ago, how would you rate your health in general now?
   - ○ Much better now than 1 yr ago
   - ○ Somewhat better than 1 yr ago
   - ○ About the same as 1 yr ago
   - ○ Somewhat worse than 1 yr ago
   - ○ Much worse than 1 yr ago

3. The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

Activities	Yes, Limited A Lot	Yes, Limited A Little	No, Not Limited At All
a. Vigorous activities such as running, lifting heavy objects, participating in strenuous sports	○	○	○
b. Moderate activities such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	○	○	○
c. Lifting or carrying groceries	○	○	○
d. Climbing several flights of stairs	○	○	○
e. Climbing one flight of stairs	○	○	○
f. Bending, kneeling or stooping	○	○	○
g. Walking more than one mile	○	○	○
h. Walking several blocks	○	○	○
i. Walking one block	○	○	○
j. Bathing or dressing	○	○	○

This is a reduced copy of the fillable SF-36 Standard Form available from Response Technologies, Inc.

STANDARD SF-36, FAX FORM – PAGE TWO OF THREE

37773

PAGE 2 OF 3

Shade circles like this: ●
Not like this: ☒ ✓

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your PHYSICAL HEALTH?

a. Cut down the amount of time you spent on work or other activities	○ Yes	○ No
b. Accomplished less than you would like	○ Yes	○ No
c. Were limited in the kind of work or other activities	○ Yes	○ No
d. Had difficulty performing the work or other activities (for example, it took extra effort)	○ Yes	○ No

5. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any EMOTIONAL PROBLEMS (such as feeling depressed or anxious)?

a. Cut down the amount of time you spent on work or other activities	○ Yes	○ No
b. Accomplished less than you would like	○ Yes	○ No
c. Didn't do work or other activities as carefully as usual	○ Yes	○ No

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

○ Not at all ○ Slightly ○ Moderately ○ Quite a bit ○ Extremely

7. How much bodily pain have you had during the past 4 weeks?

○ None ○ Very mild ○ Mild ○ Moderate ○ Severe ○ Very severe

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

○ Not at all ○ A little bit ○ Moderately ○ Quite a bit ○ Extremely

This is a reduced copy of the fillable SF-36 Standard Form available from Response Technologies, Inc.

STANDARD SF-36, FAX FORM – PAGE THREE OF THREE

37778

PAGE 3 OF 3

9. These questions are about how you feel and how things have been with you during the past 4 weeks. For each question, please give the one answer that comes closest to the way you have been feeling. How much of the time during the past 4 weeks:

	All of the Time	Most of the Time	A Good Bit of the Time	Some of the Time	A Little of the Time	None of the Time
a. Did you feel full of pep?	○	○	○	○	○	○
b. Have you been a very nervous person?	○	○	○	○	○	○
c. Have you felt so down in the dumps that nothing could cheer you up?	○	○	○	○	○	○
d. Have you felt calm and peaceful?	○	○	○	○	○	○
e. Did you have a lot of energy?	○	○	○	○	○	○
f. Have you felt downhearted and blue?	○	○	○	○	○	○
g. Did you feel worn out?	○	○	○	○	○	○
h. Have you been a happy person?	○	○	○	○	○	○
i. Did you feel tired?	○	○	○	○	○	○

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

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

11. Please choose the answer that best describes how TRUE or FALSE each of the following statements is for you.

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
a. I seem to get sick a little easier than other people ...	○	○	○	○	○
b. I am as healthy as anybody I know	○	○	○	○	○
c. I expect my health to get worse	○	○	○	○	○
d. My health is excellent	○	○	○	○	○

12a. Which are you? ○ Male ○ Female

b. How old were you on your last birthday?
○ Less than 35 ○ 65-74
○ 35-44 ○ 75-84
○ 45-54 ○ 85 and older
○ 55-64

13. Have you ever filled out this form before?
○ Yes ○ No ○ Don't remember

Do not mark here ○A ○B ○C ○D ○E

This is a reduced copy of the fillable SF-36 Standard Form available from Response Technologies, Inc.

DEVELOPMENTAL (PRE-PUBLICATION) SF-36, BOOKLET FORM – PAGE ONE OF SIX

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.

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 left margin.

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

(circle one number)

Excellent 1

Very good 2

Good 3

Fair 4

Poor 5

2. Compared to one year ago, how would you rate your health in general now?

(circle one number)

Much better now than one year ago 1

Somewhat better now than one year ago 2

About the same 3

Somewhat worse now than one year ago 4

Much worse now than one year ago 5

DEVELOPMENTAL (PRE-PUBLICATION) SF-36, BOOKLET FORM – PAGE TWO OF SIX

HEALTH AND DAILY ACTIVITIES

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much? (Circle 1, 2, or 3 on each line.)

	Yes, Limited A Lot	Yes, Limited A Little	No, Not Limited At All
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	1	2	3
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
c. Lifting or carrying groceries	1	2	3
d. Climbing several flights of stairs	1	2	3
e. Climbing one flight of stairs	1	2	3
f. Bending, kneeling, or stooping	1	2	3
g. Walking more than a mile	1	2	3
h. Walking several blocks	1	2	3
i. Walking one block	1	2	3
j. Bathing and dressing yourself	1	2	3

DEVELOPMENTAL (PRE-PUBLICATION) SF-36; BOOKLET FORM – PAGE THREE OF SIX

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health? (Please answer YES or NO for each question by circling 1 or 2 on each line.)

	YES	NO
a. Cut down on the amount of time you spent on work or other activities	1	2
b. Accomplished less than you would like	1	2
c. Were limited in the kind of work or other activities	1	2
d. Had difficulty performing the work or other activities (for example, it took extra effort)	1	2

5. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)? (Please answer YES or NO for each question by circling 1 or 2 on each line.)

	YES	NO
a. Cut down the amount of time you spent on work or other activities	1	2
b. Accomplished less than you would like	1	2
c. Didn't do work or other activities as carefully as usual	1	2

DEVELOPMENTAL (PRE-PUBLICATION) SF-36, BOOKLET FORM – PAGE FOUR OF SIX

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?

(circle one number)

Not at all 1

Slightly 2

Moderately 3

Quite a bit 4

Extremely 5

PAIN

7. How much bodily pain have you had during the past 4 weeks?

(circle one number)

None 1

Very mild 2

Mild 3

Moderate 4

Severe 5

Very severe 5

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

(circle one number)

Not at all 1

A little bit 2

Moderately 3

Quite a bit 4

Extremely 5

DEVELOPMENTAL (PRE-PUBLICATION) SF-36, BOOKLET FORM. - PAGE FIVE OF SIX

YOUR FEELINGS

9. These questions are about how you feel and how things have been with you during the past month. 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 past month ...

(circle one number on each line)

	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
a. did you feel full of pep?	1	2	3	4	5	6
b. have you been a very nervous person?	1	2	3	4	5	6
c. have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5	6
d. have you felt calm and peaceful?	1	2	3	4	5	6
e. did you have a lot of energy?	1	2	3	4	5	6
f. have you felt downhearted and blue?	1	2	3	4	5	6
g. did you feel worn out?	1	2	3	4	5	6
h. have you been a happy person?	1	2	3	4	5	6
i. did you feel tired?	1	2	3	4	5	6
j. has your health limited your social activities (like visiting with friends or close relatives)?	1	2	3	4	5	6

DEVELOPMENTAL (PRE-PUBLICATION) SF-36, BOOKLET FORM – PAGE SIX OF SIX

HEALTH IN GENERAL

10. Please choose the answer that best describes how true or false each of the following statements is for you.

(circle one number on each line)

	Definitely True	Mostly True	Not Sure	Mostly False	Definitely False
a. I seem to get sick a little easier than other people	1	2	3	4	5
b. I am as healthy as anybody I know	1	2	3	4	5
c. I expect my health to get worse	1	2	3	4	5
d. My health is excellent	1	2	3	4	5

U.K. STANDARD SF-36, BOOKLET FORM – PAGE ONE OF FIVE

THE MOS 36-ITEM SHORT-FORM HEALTH SURVEY (SF-36)

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.

Answer every question by marking the answer as indicated. If you are unsure about how to answer a question, please give the best answer you can.

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

(circle one)

Excellent 1

Very good 2

Good 3

Fair 4

Poor 5

2. Compared to one year ago, how would you rate your health in general now?

(circle one)

Much better now than one year ago 1

Somewhat better now than one year ago 2

About the same as one year ago 3

Somewhat worse now than one year ago 4

Much worse now than one year ago 5

U.K. STANDARD SF-36, BOOKLET FORM – PAGE TWO OF FIVE

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

(circle one number on each line)

ACTIVITIES	Yes, Limited A Lot	Yes, Limited A Little	No, Not Limited At All
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	1	2	3
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf	1	2	3
c. Lifting or carrying groceries	1	2	3
d. Climbing several flights of stairs	1	2	3
e. Climbing one flight of stairs	1	2	3
f. Bending, kneeling, or stooping	1	2	3
g. Walking more than a mile	1	2	3
h. Walking half a mile	1	2	3
i. Walking one hundred yards	1	2	3
j. Bathing or dressing yourself	1	2	3

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

(circle one number on each line)

	YES	NO
a. Cut down on the amount of time you spent on work or other activities	1	2
b. Accomplished less than you would like	1	2
c. Were limited in the kind of work or other activities	1	2
d. Had difficulty performing the work or other activities (for example, it took extra effort)	1	2

U.K. STANDARD SF-36, BOOKLET FORM – PAGE THREE OF FIVE

6. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

(circle one number on each line)

	YES	NO
a. Cut down on the amount of time you spent on work or other activities	1	2
b. Accomplished less than you would like	1	2
c. Didn't do work or other activities as carefully as usual	1	2

5. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbours, or groups?

(circle one)

Not at all 1

Slightly 2

Moderately 3

Quite a bit 4

Extremely 5

7. How much bodily pain have you had during the past 4 weeks?

(circle one)

None 1

Very mild 2

Mild 3

Moderate 4

Severe 5

Very severe 6

U.K. STANDARD SF-36, BOOKLET FORM – PAGE FOUR OF FIVE

8. During the past 4 weeks, how much did pain interfere with your normal work (including both work outside the home and housework)?

(circle one)

Not at all 1

A little bit 2

Moderately 3

Quite a bit 4

Extremely 5

9. 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 -

(circle one number on each line)

	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
a. Did you feel full of life?	1	2	3	4	5	6
b. Have you been a very nervous person?	1	2	3	4	5	6
c. Have you felt so down in the dumps that nothing could cheer you up?	1	2	3	4	5	6
d. Have you felt calm and peaceful?	1	2	3	4	5	6
e. Did you have a lot of energy?	1	2	3	4	5	6
f. Have you felt downhearted and low?	1	2	3	4	5	6
g. Did you feel worn-out?	1	2	3	4	5	6
h. Have you been a happy person?	1	2	3	4	5	6
i. Did you feel tired?	1	2	3	4	5	6

U.K. STANDARD SF-36, BOOKLET FORM – PAGE FIVE OF FIVE

10. During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities (like visiting with friends, relatives, etc.)?

(circle one)

All of the time 1
Most of the time 2
Some of the time 3
A little of the time 4
None of the time 5

11. How TRUE or FALSE is each of the following statements for you?

(circle one number on each line)

	Definitely True	Mostly True	Don't Know	Mostly False	Definitely False
a. I seem to get ill more easily than other people	1	2	3	4	5
b. I am as healthy as anybody I know	1	2	3	4	5
c. I expect my health to get worse	1	2	3	4	5
d. My health is excellent	1	2	3	4	5

U.K. DEVELOPMENTAL SF-36, BOOKLET FORM – PAGE ONE OF FIVE

For office use

HEALTH STATUS QUESTIONNAIRE (SF-36)

THE FOLLOWING QUESTIONS ASK FOR YOUR VIEWS ABOUT YOUR HEALTH, HOW YOU FEEL AND HOW WELL YOU ARE ABLE TO DO YOUR USUAL ACTIVITIES. IF YOU ARE UNSURE ABOUT HOW TO ANSWER ANY QUESTION, PLEASE GIVE THE BEST ANSWER YOU CAN AND MAKE ANY COMMENTS IN THE SPACE AVAILABLE AFTER QUESTION 10

Please tick one

1. In general would you say your health is:

- Excellent ○
- Very good ○
- Good ○
- Fair ○
- Poor ○

2. Compared to one year ago, how would you rate your health in general now?

- 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 ○

U.K. DEVELOPMENTAL SF-36, BOOKLET FORM – PAGE TWO OF FIVE

For office use

HEALTH AND DAILY ACTIVITIES

3. The following questions are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

Please tick one cross on each line

	Yes, limited a lot	Yes, limited a little	No, not limited at all	
a. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports	○	○	○	☐
b. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf	○	○	○	☐
c. Lifting or carrying groceries	○	○	○	☐
d. Climbing several flights of stairs	○	○	○	☐
e. Climbing one flight of stairs	○	○	○	☐
f. Bending, kneeling or stooping	○	○	○	☐
g. Walking more than a mile	○	○	○	☐
h. Walking half a mile	○	○	○	☐
i. Walking 100 yards	○	○	○	☐
j. Bathing and dressing yourself	○	○	○	☐

4. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of your physical health?

Answer Yes or No to each question

	YES	NO	
a. Cut down on the amount of time you spent on work or other activities	○	○	☐
b. Accomplished less than you would like	○	○	☐
c. Were limited in the kind of work or other activities	○	○	☐
d. Had difficulty performing the work or other activities (e.g. it took extra effort)	○	○	☐

U.S. DEVELOPMENTAL SF-36, BOOKLET FORM – PAGE THREE OF FIVE

For office use

5. During the past 4 weeks, have you had any of the following problems with your work or other regular daily activities as a result of any emotional problems (such as feeling depressed or anxious)?

Answer Yes or No to each question

	YES	NO	
a. Cut down on the amount of time you spent on work or other activities	○	○	☐
b. Accomplished less than you would like	○	○	☐
c. Didn't do work or other activities as carefully as usual	○	○	☐

6. During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors or groups?

Please fill one

Not at all	○	☐
Slightly	○	☐
Moderately	○	☐
Quite a bit	○	☐
Extremely	○	☐

7. How much bodily pain have you had during the past 4 weeks?

None	○	☐
Very mild	○	☐
Mild	○	☐
Moderate	○	☐
Severe	○	☐
Very severe	○	☐

8. During the past 4 weeks, how much did pain interfere with your normal work (including work both outside the home and housework)?

Not at all	○	☐
A little bit	○	☐
Moderately	○	☐
Quite a bit	○	☐
Extremely	○	☐

U.K. DEVELOPMENTAL SF-36, BOOKLET FORM – PAGE FOUR OF FIVE

For office use

YOUR FEELINGS

9. These questions are about how you feel and how things have been with you during the past month. *(For each question, please indicate the one answer that comes closest to the way you have been feeling)*

Please tick one circle on each line

How much time during the past month:	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	For office use
a. Did you feel full of life?	○	○	○	○	○	○	☐
b. Have you been a very nervous person?	○	○	○	○	○	○	☐
c. Have you felt so down in the dumps that nothing could cheer you up?	○	○	○	○	○	○	☐
d. Have you felt calm and peaceful?	○	○	○	○	○	○	☐
e. Did you have a lot of energy?	○	○	○	○	○	○	☐
f. Have you felt downhearted and low?	○	○	○	○	○	○	☐
g. Did you feel worn out?	○	○	○	○	○	○	☐
h. Have you been a happy person?	○	○	○	○	○	○	☐
i. Did you feel tired?	○	○	○	○	○	○	☐
j. Has your health limited your social activities (like visiting friends or close relatives)?	○	○	○	○	○	○	☐

U.K. DEVELOPMENTAL SF-36, BOOKLET FORM — PAGE FIVE OF FIVE

For office use

HEALTH IN GENERAL

10. Please choose the answer that best describes how true or false each of the following statements is for you.

Please tick one circle on each line

	Definitely true	Mostly true	Not sure	Mostly false	Definitely false
a. I seem to get ill more easily than other people	○	○	○	○	○
b. I am as healthy as anybody I know	○	○	○	○	○
c. I expect my health to get worse	○	○	○	○	○
d. My health is excellent	○	○	○	○	○

Comments

Thank you very much for your assistance

Please return the completed booklet in the envelope provided

NO STAMP IS REQUIRED.

Medical Care Research Unit, Department of Public Health Medicine,
University of Sheffield Medical School, Beech Hill Road, Sheffield, S10 2RX

SAMPLE ITEMS FROM THE SF-36 HEALTH SURVEY: U.S. ENGLISH AND MEXICAN-AMERICAN (SPANISH) VERSIONS

Scale	Item	U.S. English[a]	Mexican-American[b]
Physical Functioning (PF)	3i	Walking one block	Caminar una cuadra
Role Functioning (RP, RE)	4b, 5b	Accomplished less than you would like	Ha logrado menos de lo que le hubiera gustado
Bodily Pain (BP)	7	How much bodily pain have you had during the past 4 weeks?	¿Cuánto dolor del cuerpo ha tenido usted durante las últimas 4 semanas?
General Health (GH)	1	In general, would you say your health is: Excellent, Very Good, Good, Fair, Poor	En términos generales, ¿diría que su salud es: Excelente, Muy Buena, Buena, Pasable, Mala
Vitality (VT)	9i	Did you feel tired?	¿Se ha sentido cansado?
Social Functioning (SF)	6	During the past 4 weeks, to what extent has your physical health or emotional problems interfered with your normal social activities with family, friends, neighbors, or groups?	Durante las últimas 4 semanas, ¿en qué medida su salud física o sus problemas emocionales han dificultado sus actividades sociales normales con la familia, amigos, vecinos o grupos?
Mental Health (MH)	9f	Have you felt downhearted and blue?	¿Se ha sentido desanimado y triste?
Reported Health Transition (HT)	2	Compared to one year ago, how would you rate your health in general now?	Comparando su salud con hace un año, ¿cómo la calificaría en general ahora?

[a]

[b]

APPENDIX C: SCRIPT FOR PERSONAL INTERVIEW SF-36 ADMINISTRATION

The script included in this appendix is recommended for interviewer administrations of the SF-36 items. It can be administered both by telephone and in-person. Standard SF-36 instructions should precede this script (see Chapter 4 for details). Interviewers also should follow standard procedures for repeating questions and response choices as required by the respondent.

These first questions are about your health now and your current daily activities. Please try to answer every question as accurately as you can.

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

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

Q2 Compared to 1 year ago, how would you rate your health in general now? Would you say it is...

1. *much better now than one year ago*
2. *somewhat better now than one year ago*
3. *about the same as one year ago*
4. *somewhat worse now than one year ago*
5. *much worse now than one year ago*

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

Q3 First, vigorous activities, such as running, lifting heavy objects, participating in strenuous sports. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q4 ...moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q5 ...lifting or carrying groceries. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q6. ...climbing several flights of stairs. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q7 ...climbing one flight of stairs. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q8. ...bending, kneeling, or stooping. Does your health now limit you a lot, limit you a little, or not limit you at all?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q9 ...walking more than a mile. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q10 ...walking several blocks. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q11 ...walking one block. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

Q12 ...bathing or dressing yourself. Does your health now limit you a lot, limit you a little, or not limit you at all?

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

Is that because of your health?

1. *Yes, limited a lot*
2. *Yes, limited a little*
3. *No, not limited at all*

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

Q13 During the past 4 weeks, have you had to cut down the amount of time you spent on work or other regular daily activities as a result of your physical health?

1. *Yes*
2. *No*

Q14 During the past 4 weeks, have you accomplished less than you would like as a result of your physical health?

1. *Yes*
2. *No*

Q15 During the past 4 weeks, were you limited in the kind of work or other regular daily activities you do as a result of your physical health?

1. *Yes*
2. *No*

Q16 During the past 4 weeks, have you had difficulty performing work or other regular daily activities as a result of your physical health, for example, it took extra effort?

1. *Yes*
2. *No*

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

Q17 During the past 4 weeks, have you cut down the amount of time you spent on work or regular daily activities as a result of any emotional problems, such as feeling depressed or anxious?

1. *Yes*
2. *No*

Q18 During the past 4 weeks, have you accomplished less than you would like as a result of any emotional problems, such as feeling depressed or anxious?

1. *Yes*
2. *No*

Q19 During the past 4 weeks, did you not do work or other regular daily activities as carefully as usual as a result of any emotional problems, such as feeling depressed or anxious?

1. *Yes*
2. *No*

Q20 During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your social activities like visiting with friends or relatives? Has it interfered...

1. *not at all*
2. *slightly*
3. *moderately*
4. *quite a bit*
5. *or extremely*

Q21 During the past 4 weeks, how much did pain interfere with your normal work, including both work outside the home and housework? Did it interfere...

1. *not at all*
2. *a little bit*
3. *moderately*
4. *quite a bit*
5. *or extremely*

Q22 How much bodily pain have you had during the past 4 weeks? Have you had...

1. *none*
2. *very mild*
3. *mild*
4. *moderate*
5. *severe*
6. *or very severe*

Q23 During the past 4 weeks, how much of the time has your physical health or emotional problems interfered with your physical activities like visiting wish friends or relatives? Has it interfered...

1. all of the time

2. most of the time

3. some of the time

4. a little of the time

5. or none of the time

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

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

Q24 how much of the time during the past 4 weeks...did you feel full of pep? Read categories:

1. all of the time

2. most of the time

3. a good bit of the time

4. some of the time

5. a little of the time

6. none of the time

Q25 How much of the time during the past 4 weeks...have you have been a very nervous parson? Read categories:

1. all of the time

2. most of the time

3. a good bit of the time

4. some of the time

5. a little of the time

6. none of the time

Q26 How much of the time during the past 4 weeks...have you felt so down in the dumps that nothing could cheer you up? Read categories only if necessary.

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

Q27 How much of the time during the past 4 weeks...have you felt calm and peaceful? Read categories only if necessary.

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

Q28 How much of the time during the past 4 weeks...did you have a lot of energy? Read categories only if necessary.

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

Q29 How much of the time during the past 4 weeks...have you felt downhearted and blue? Read categories only if necessary.

1. *all of the time*
2. *most of the time*
3. *a good bit of the time*

4. *some of the time*
5. *a little of the time*
6. *none of the time*

Q30 How much of the time during the past 4 weeks...did you feel worn out? Read categories only if necessary.

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

Q31 How much of the time during the past 4 weeks...have you been a happy person? Read categories only if necessary.

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

Q32 How much of the time during the past 4 weeks...did you feel tired? Read categories only if necessary.

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

These next questions are about your health and health-related matters.

Now I'm going to read a list of statements. After each one, please tell me if it is definitely true, mostly true, mostly false, or definitely false. If you don't know, just tell me.

Q33 I seem to get sick a little earlier than other people. Would you say that's... Read categories:

1. *definitely true*
2. *mostly true*
3. *don't know*
4. *mostly false*
5. *definitely false*

Q34 I am as healthy as anybody I know. Would you say that's... Read categories.

1. *definitely true*
2. *mostly true*
3. *don't know*
4. *mostly false*
5. *definitely false*

Q35 I expect my health to get worse. Would you say that's... Read categories.

1. *definitely true*
2. *mostly true*
3. *don't know*
4. *mostly false*
5. *definitely false*

Q36 My health is excellent. Would you say that's... Read categories.

1. *definitely true*
2. *mostly true*
3. *don't know*
4. *mostly false*
5. *definitely false*

SF-36™ USER MAILING LIST REGISTRATION FORM

Users of the *SF-36 Health Survey Manual and Interpretation Guide* on our mailing list are sent updates without charge as they become available. If you would like to be on this mailing list, please fill out and mail or FAX this form.

Contact person

Title

Organization

Address

Telephone

Facsimile

We welcome your comments, including suggestions for improvement:

Please return form to:

SF-36 User Manual Mailing List, The Health Institute, NEMC – Box 345, 750 Washington Street, Boston, MA 02111, or FAX to: 617-350-8077.

REFERENCES

Aaronson, N.K., Acquadro, C., Alonso, J., Apolone, G., Bucquet, D., Bullinger, M., Bungay, K., Fukuhara, S., Gandek, B., Keller, S., Razavi, R., Sanson-Fisher, M., Sullivan, S., Wood-Dauphinee, S., Wagner, A., & Ware, J.E. (1992). International quality of life assessment (IQOLA) project. *Quality of Life Research, 1*, 349-351.

American College of Physicians. (1988). Comprehensive functional assessment for elderly patients. *Annals of Internal Medicine, 109*, 70-72.

American Psychological Association. (1985). *Standards for education and psychological testing*. Washington, DC: American Psychological Association.

Armor, D.J. (1974). Theta reliability and factor scaling. In H.L. Costner (Ed.), *Sociological Methodology 1973-1974* (pp. 17-50). San Francisco, CA: Jossey-Bass, Inc.

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.

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., Barsky, A.J., & Weinstein, M.C. (1991). Performance of a five-item mental health screening test. *Medical Care, 29*, 169-176.

Birdman, A.B., Keane, D., & Lurie, N. (1990). Measuring health changes among severely ill patients: The floor phenomenon. *Medical Care, 28*, 1142-1151.

Bombardier, C., Ware, J.E., 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.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.

Brook, R.H., Fink, A., Kosecoff, J., Linn, L.S., Watson, W.E., Davies, A.R., Clark, V.A., Kamberg, C., & DelBlanco, 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., Davies-Avery, A., Stewart, A.L., Donald, C.A., Rogers, W.H., Williams, K.N., & Johnston, S.A. (1979). *Conceptualization and measurement of health for adults in the Health Insurance Study: Volume VIII: Overview*. Santa Monica, CA: The RAND Corporation (publication no. R-1987/8-HEW).

Brook, R.H., Ware, J.E., Rogers, W.H., Keeler, E.B., Davies, A.R., Donald, C.A., Goldberg, G.A., Lohr, K.N., Masthay, P.C., & Newhouse, J.P. (1983). Does free care improve adults' health? Results from a randomized controlled trial. *New England Journal of Medicine, 309*, 1426-1434.

Bungay, K.M., & Ware, J.E. (1993). Measuring and monitoring health-related quality of life. *Current Concepts*. Kalamazoo, MI: The Upjohn Company.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cantril, H. (1965). *The pattern of human concerns*. New Brunswick, NJ: Rutgers University Press.

Cassileth, B.R., Lusk, E.J., Strouse, T.B., Miller, D.S., Brown, L.L., Cross, P.A., & Tenaglia, A.N. (1984). Psychosocial status in chronic illness: A comparative analysis of six diagnostic groups. *New England Journal of Medicine, 311*, 506-511.

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: MTP Press Limited.

Chobanian, A.V. (1986). Antihypertensive therapy in evolution. *New England Journal of Medicine, 314*, 1701-1702.

Cleary, P.D., Epstein, A.M., Oster, G., Morrissey, G.S., Sraison, W.B., Debussey, S., Plachetka, J., & Zimmerman, M. (1991). Health-related quality of life among patients undergoing percutaneous transluminal coronary angioplasty. *Medical Care, 29*, 939-950.

Cluff, L.E. (1961). Chronic disease, function and the quality of care. *Journal of Chronic Diseases, 14*, 299-304.

Coates, A., Gebski, V., Starr, M., Bishop, J.F., Jeal, P.N., Woods, R.L., Snyder, R., Tattersall, M.H., Byrne, M., Harvey, M., Gill, G., Simpson, J., Drummond, R., Browne, J., Van Couteo, 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. (1914). The product of a hospital. *Surgery, Gynecology and Obstetrics, 18*, 491-496.

Cohen, J. (1988). *Statistical power for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-Experimentation: Design & analysis issues for field settings*. Chicago, IL: Rand McNally College Publishing Company.

Cronbach, L.J., & Warrington, W.G. (1951). Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika, 16*, 167-188.

Croog, S.H., Levine, S., Testa, M.A., Brown, B., Bulpitt, C.J., Jenkins, D., Klerman, G.L., & Williams, G.H. (1986). The effects of antihypertensive therapy on the quality of life. *New England Journal of Medicine, 314*, 1657-1664.

Davies, A.R., Sherbourne, C.D., Peterson, J.R., & Ware, J.E. (1988). *Scoring manual: Adult health status and patient satisfaction measures used in RAND's Health Insurance Experiment*. Santa Monica, CA: The RAND Corporation (publication no. N-2190-HHS).

Davies, A.R., & Ware, J.E. (1981). *Measuring health perceptions in the Health Insurance Experiment*. Santa Monica, CA: The RAND Corporation (publication no. R-2711-HHS).

DiCocco, L., & Apple, D. (1958). Health needs and opinions of older adults. *Public Health Reports, 73*, 479-487.

Donald, C.A., & Ware, J.E. (1984). The measurement of social support. In J.R. Greenley (Ed.), *Research in community and mental health*. Greenwich, CT: JAI Press, 325-370.

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, NY: Le Jacq Publishing Company.

Eisen, M., Donald, C.A., Ware, J.E., & Brook, R.H. (1980). *Conceptualization and measurement of health for children in the Health Insurance Study.* Santa Monica, CA: The RAND Corporation (publication no. R-2313-HEW).

Elinson, J., & Mattson, M.E. (1984). Assessing the quality of life in clinical trials of cardiovascular therapies: Introduction to the panel presentations. 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. 143–145). New York, NY: Le Jacq Publishing Company.

Ellwood, P.M. (1988). Outcomes management: A technology of patient experience [Shattuck Lecture]. *New England Journal of Medicine, 318,* 1549–1556.

Fowler, F.J., Wennberg, J.E., Timothy, R.P., Barry, M.J., Mulley, A.G., & Henley, D. (1988). Symptom status and quality of life following prostatectomy. *Journal of the American Medical Association, 259,* 3018–3022.

Fryback, D.G., Dasbach, E.J., Klein, R., Klein, B.E., Dorn, N., Peterson, K., & Martin, P.A. (1993). The Beaver Dam health outcomes study: Initial catalog of health state quality factors. *Medical Decision Making, 13,* 89–102.

Gandek, B. (1992). International Quality of Life Assessment (IQOLA) project. *The Quality of Life Newsletter, 5,* 10.

Garratt, A.M., Ruta, D.A., Abdalla, M.I., Buckingham, J.K., & Russell, I.T. (1993). The SF-36 Health Profile: An outcome measure suitable for routine use within the NHS? *British Medical Journal, 306,* 1440–1444.

Geigle, R., & Jones, S.B. (1990). Outcomes measurement: A report from the front. *Inquiry, 27,* 7–13.

Gelberg, L., & Linn, L.S. (1989). Psychological distress among homeless adults. *Journal of Nervous and Mental Disease, 177,* 291–295.

Guttman, L.A. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.

Haley, S.M., McHorney, C.A., & Ware, J.E. (in review). Evaluation of the MOS SF-36 Physical Functioning scale (PF-10) using Rasch item response theory methodology: Unidimensionality and reproducibility.

Hays, R.D., & Stewart, A.L. (1990). The structure of self-reported health in chronic disease patients. *Journal of Consulting and Clinical Psychology, 2,* 22–30.

Helmstadter, G.C. (1964). *Principles of psychological measurement.* New York, NY: Appleton-Century-Crofts, Inc.

Howard, K.I., & Forehand, G.G. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. *Educational and Psychological Measurement, 22,* 731–735.

Hunt, S.M., McKenna, S., McEwen, J., Williams, J., & Papp, E. (1981). The Nottingham Health Profile: Subjective health status and medical consultations. *Social Science in Medicine, 15A,* 221–229.

Jenkinson, C., Coulter, A., & Wright, L. (1993). The Short Form 36 (SF-36) Health Survey Questionnaire: Normative data for adults of working age. *British Medical Journal, 306,* 1437–1440.

Jette, A.M., Davies, A.R., Cleary, P.D., Calkins, D.R., Rubenstein, L.V., Fink, A., Kosecoff, J., Young, R.T., Brook, R.H., & DelBanco, T.L. (1986). The Functional Status Questionnaire: Reliability and validity when used in primary care. *Journal of General Internal Medicine, 1*, 143-149.

Kantz, M.E., Harris, W.J., Levitsky, K., Ware, J.E., & Davies, A.R. (1992). Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. *Medical Care, 30*(Suppl.), 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 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.). (1967). The Portugal conference: Measuring quality of life and functional status in clinical and epidemiological research. *Journal of Chronic Diseases, 40*(Special issue).

Katz, S., Downs, T.D., Cash, H.R., & Grotz, R.C. (1970). Progress in development of the index of ADL. *Gerontologist, 10*, 20-30.

Katz, S., Ford, A.B., Moskowitz, R.W., Jacobson, 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.

Kravitz, R.L., Greenfield, S., Rogers, W.H., Manning, W.G., Zubkoff, M., Nelson, E., Tarlov, A.R., & Ware, J.E. (1992). Differences in the mix of patients among medical specialties and systems of care: Results from the Medical Outcomes Study. *Journal of the American Medical Association, 267*, 1617-1623.

Kutner, P.S., Davies, A.R., Mayer, 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*(Suppl.), MS136-MS149.

Lancaster, T.R., Singer, D.E., Sheehan, M.A., Oertel, L.B., Manwaring, S.W., Hughes, R.A., & Kistler, J.P. (1991). The impact of long-term warfarin therapy on quality of life: Evidence from a randomized trial. *Archives of Internal Medicine, 151*, 1944-1949.

Lindley, D., Butler, J.B.V., & Waller, E.T. (1992). Using health status measures in the hospital setting: From acute care to 'outcomes management'. *Medical Care, 30*(Suppl.), MS57-MS73.

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.

Lohr, K.N. (1989). Advances in health status assessment: Overview of the conference. *Medical Care, 27*, S1-S11.

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*(Suppl.), MS1-MS14.

Lohr, K.N., & Ware, J.E. (Eds.). (1987). Proceeding of the advances in health assessment conference. *Journal of Chronic Disease, 40*(Suppl. 1).

Louis Harris and Associates, Inc. (1984). *A report and on HMO: 1980-1984: Summary Report*. Menlo Park, CA: The Henry J. Kaiser Family Foundation.

Lurie, N., Ward, N.B., Shapiro, K.F., & Brook, R.H. (1984). Termination from Medi-Cal: Does it affect health? *New England Journal of Medicine, 311*, 480-484.

Macmillan, A.M. (1957). The health opinion survey: Technique 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. (1982). The status of health in demand estimation: Beyond excellent, good, fair, and poor. In V.R. Fuchs (Ed.), *Economic aspects of health*. Chicago, IL: University of Chicago Press. (Also RAND publication no. R-2696-HHS).

McDermott, W. (1981). Absence of indicators of the influence of its physicians on 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* (pp. 269-321). New York: Oxford University Press.

McHorney, C.A., Kosinski, M., & Ware, J.E. (in review). Comparisons of the costs and quality of norms collected by mail versus telephone interview: Results from a national survey. *Medical Care*.

McHorney, C.A., & Ware, J.E. (forthcoming). Construction and validation of an alternate form general mental health scale for the MOS SF-36 Health Survey.

McHorney, C.A., Ware, J.E., Lu, J.F.R., & Sherbourne, C.D. (in press). 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*.

McHorney, C.A., Ware, J.E., & 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., Rogers, W.H., Raczek, A., & Lu, J.F.R. (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.), MS253-MS265.

Meyerboom-De Jong, 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*. New York: Springer-Verlag.

Montgomery, E.A., & Paranjpe, A.V. (1985). *A report and on HMOs 1980-1984*. Menlo Park, CA: The Henry J. Kaiser Family Foundation.

Moseley, F., Ware, J.E., & Levine, S. (1989). Finale panel: Comments on the conference on advances in health status assessment. *Medical Care, 27*(Suppl.), S293-298.

National Center for Health Statistics. (1991). *Health, United States*. U.S. Department of Health and Human Services.

Nelson, E.C., & Berwick, D.M. (1989). The measurement of health status in clinical practice. *Medical Care, 27*, S77-S90.

Nelson, E.C., Conger, B., Douglass, R., Gephart, D., Kirk, J., Page, R., Clark, A., Johnson, K., Stone, K., Wasson, J., & Zubkoff, M. (1983). Functional health status level of primary care patients. *Journal of the American Medical Association, 249*, 3331-3338.

Nelson, E.C., Landgraf, J.M., Hays, R.D., Wasson, J.H., & Kirk, J.W. (1990a). The functional status of patients, how can it be measured in physicians' offices? *Medical Care, 28,* 1111-1126.

Nelson, E.C., Landgraf, J.M., Hays, R.D., Kirk, J.W., Wasson, J.H., Keller, A., & Zubkoff, M. (1990b). 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., Wasson, J., Kirk, J., Keller, A., Clark, D., Dietrich, A., Stewart, A., & Zubkoff, M. (1987). Assessment of function in routine clinical practice: Description of the COOP chart method and preliminary findings. *Journal of Chronic Disease, 40*(Suppl. 1), 55S-63S.

Nerenz, D.R., Repasky, D.P., Whitehouse, F.W., & Kahkonen, D.M. (1992). Ongoing assessment of health status in patients with diabetes mellitus. *Medical Care, 30*(Suppl.), MS112-MS124.

Newhouse, J.P., Manning, W.G., Morris, C.N., Orr, L.L., Duan, N., Keeler, E.B., Leibowitz, A., Marquis, K.H., Marquis, M.S., Phelps, C.E., & Brook, R.H. (1981). Some interim results from a controlled trial of cost sharing in health insurance. *New England Journal of Medicine, 305,* 1501-1507.

Nunnally, J.C. (1978). *Psychometric theory, 2nd edition.* New York, NY: McGraw-Hill.

Parkerson, G.R., Broadhead, W.E., & Tse, C.K.J. (1990). The Duke Health Profile: A 17-item measure of health and dysfunction. *Medical Care, 28,* 1056-1072.

Parkerson, G.R., Broadhead, W.E., & Tse, C.K.J. (1991). Comparison of the Duke Health Profile and the MOS Short-Form in healthy young adults. *Medical Care, 29,* 679-683.

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,* 229-234.

Patrick, D.L., & Deyo, R.A. (1989). Generic and disease-specific measures in assessing health status and quality of life. *Medical Care, 27*(Suppl.), S217-S232.

Phillips, R.C., & Lansky, D.J. (1992). Outcomes management in heart valve replacement surgery: Early experience. *Journal of Heart Valve Disease, 1,* 42-50.

Read, J.L., Quinn, R.J., & Hoefer, M.A. (1987). Measuring overall health: An evaluation of three important approaches. *Journal of Chronic Disease, 40*(Suppl. 1), 7S-22S.

Relman, A.S. (1988). Assessment and accountability: The third revolution in medical care. *New England Journal of Medicine, 319,* 1220-1222.

Reuben, D.B., & Siu, A.L. (1990). An objective measure of physical function of elderly outpatients: The Physical Performance Test. *Journal of the American Geriatrics Society, 38,* 1105-1112.

Roper, W.L., Winkenwerder, W., Hackbarth, G.M., & Krakauer, H. (1988). Effectiveness in health care: An initiative to evaluate and improve medical practice. *New England Journal of Medicine, 319,* 1197-1202.

Schroeder, S.A. (1987). Outcome assessment 70 years later: Are we ready? *New England Journal of Medicine, 316,* 160-162.

Selltiz, C., Wrightsman, L.S., & Cook, S.W. (1976). *Research methods in social relations* (p. 183). New York: Holt Rinehart & Winston.

Shapiro, M.F., Ware, J.E., & Sherbourne, C.D. (1986). Effects of cost-sharing on seeking care for serious and chronic symptoms: Results of a randomized clinical trial. *Annals of Internal Medicine, 104,* 246-251.

Sherbourne, C.D., Stewart, A.L., & Wells, K.B. (1992). Role functioning measures. In A.L. Stewart & J.E. Ware (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 205-219). Durham, NC: Duke University Press.

Silver, G.A. (1990). Paul Anthony Lembcke, M.D., M.P.H.: A pioneer in medical care evaluation. *American Journal of Public Health, 80*, 342-348.

Siu, A.L., Reuben, D.B., & Hays, R.D. (1990). Hierarchical measures of physical functioning in ambulatory geriatrics. *Journal of the American Geriatric Society, 38*, 1113-1119.

Smith, G.R., 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.

Spiegel, J.S., Lenke, R., Spiegel, T.M., Paulos, H.E., Kane, R.L., Ward, N.B., & Wise, J.E. (1988). What are we measuring? An examination of self-reported functional status measures. *Arthritis and Rheumatism, 31*, 721-728.

Stoddard, D.M. (1989). Application of health status assessment measures to policy research. *Medical Care, 27* (3 Suppl.), S12-S26.

Stewart, A.L., Greenfield, S., Hays, R.D., Wells, K.B., Rogers, W.H., Berry, S.D., McGlynn, E.A., & Ware, J.E. (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. (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. (1992). Methods of validating MOS health measures. In A.L. Stewart & J.E. Ware (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 (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 86-101). Durham, NC: Duke University Press.

Stewart, A.L., Sherbourne, C.D., Hays, R.D., Wells, K.B., Nelson, E.C., Kamberg, C.J., Rogers, W.H., Berry, S.H., & Ware, J.E. (1992). Summary and Discussion of MOS Measures. In A.L. Stewart & J.E. Ware (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 345-372). Durham, NC: Duke University Press.

Stewart, A.L., & Ware, J.E. (Eds.). (1992). *Measuring functioning and well-being: The Medical Outcomes Study approach*. Durham, NC: Duke University Press.

Stewart, A.L., Ware, J.E., & 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., & Brook, R.H. (1982a). *Construction and scoring of aggregate functional status indexes. Volume I.* Santa Monica, CA: The RAND Corporation, (publication no. R-2551-HHS).

Stewart, A.L., Ware, J.E., & Brook, R.H. (1982b). *Construction and scoring of aggregate functional status measures. Volume II: Appendices.* Santa Monica, CA: The RAND Corporation, (publication no. N-1706-1-HHS).

Stewart, A.L., Ware, J.E., 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.* Santa Monica, CA: The RAND Corporation (publication no. R-1987/2-HEW).

Tarlov, A.R. (1983). The increasing supply of physicians, the changing structure of the health-services system, and the future practice of medicine [Shattuck Lecture]. *New England Journal of Medicine, 309*, 1235-1244.

Tarlov, A.R., Ware, J.E., 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*, 923-930.

Thalji, L., Haggerty, C.C., Rubin, R., Berekmans, T.R., & Barker, B.L. (1991). *1990 National Survey of Functional Health Status: Final report.* Chicago, IL: National Opinion Research Center.

The Health Institute (THI), International Resource Center (IRC) for Health Care Assessment. (1991). *How to score the MOS 36-Item Short-Form Health Survey (SF-36).* MOS Trust: Boston, MA.

The Health Institute (THI), International Resource Center (IRC) for Health Care Assessment. (1992). *Scoring exercise for the MOS SF-36 Health Survey.* MOS Trust: Boston, MA.

Thurstone, L.L., & Chave, E.J. (1929). *The measurement of attitude.* Chicago, IL: University of Chicago Press.

Tyler, T.A., & Fiske, D.W. (1968). Homogeneity indices and test length. *Educational and Psychological Measurement, 28*, 767-777.

Vikler, R.R., Ware, J.E., 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.

Veit, C.T., & Ware, J.E. (1983). The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology, 51*, 730-742.

Vickrey, B.G., Hays, R.D., Graber, J., Rausch, R., Engel, J., & Brook, R.H. (1992). A health-related quality of life instrument for patients evaluated for epilepsy surgery. *Medical Care, 30*, 299-319.

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.

Ware, J.E. (1976). Scales for measuring general health perceptions. *Health Services Research, 11*, 396-415.

Ware, J.E. (1984a). Methodological considerations in the selection of health status assessment procedures. 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. 87-111). New York, NY: Le Jacq Publishing Company.

Ware, J.E. (1984b). The General Health Rating 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. 184-188). New York, NY: Le Jacq Publishing Company.

Ware, J.E. (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. (1987). Standards for validating health measures: Definition and content. *Journal of Chronic Diseases, 40*, 473-480.

Ware, J.E. (1988). *How to score the revised MOS Short Form health scales.* Boston, MA: The Health Institute, New England Medical Center Hospitals.

Ware, J.E. (1994). Measuring patient function and well-being: Some lessons from the Medical Outcomes Study. In R.A. Heirgolf & 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. (1996b). Outcomes study foresees greater patient input. *QA Review, 5.*

Ware, J.E. (1996c). The use of health status and quality of life measures in outcomes and effectiveness research. Discussion paper prepared for: *National Agenda Setting Conference on Outcomes and Effectiveness Research, Agency for Health Care Policy and Research, Department of Health and Human Services.* April 14-16.

Ware, J.E. (1992). Measures for a new era of health assessment. In A.L. Stewart & J.E. Ware (Eds.), *Measuring functioning and well-being* (pp. 1-11). Durham, NC: Duke University Press.

Ware, J.E. (1993). Measuring patients' views: The optimum outcome measure. *British Medical Journal, 306,* 1429-1430.

Ware, J.E., Brook, R.H., Davies-Avery, A., Williams, K.N., Stewart, A.L., Rogers, W.H., Donald, C.A., & Johnston, S.A. (1980). *Conceptualization and measurement of health for adults in the Health Insurance Study: Volume I: Model of health and methodology.* Santa Monica, CA: The RAND Corporation (publication no. R-1987/1-HEW).

Ware, J.E., 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. (Also RAND publication no. N-1642-HHS).

Ware, J.E., Brook, R.H., Rogers, W.H., Keeler, E.B., Davies, A.R., Sherbourne, C.D., Goldberg, G.A., Camp, P., & Newhouse, J.P. (1986). Comparison of health outcomes at a health maintenance organization with those of fee-for-service care. *Lancet, 1,* 1017-1022.

Ware, J.E., 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.* Santa Monica, CA: The RAND Corporation (publication no. R-2987/6-HEW).

Ware, J.E., Davies-Avery, A., & Donald, C. (1978). *Conceptualization and measurement of health for adults in the Health Insurance Study: Volume V: General health perception.* Santa Monica, CA: The RAND Corporation (publication no. R-1987/5-HEW).

Ware, J.E., 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.* Santa Monica, CA: The RAND Corporation (publication no. R-1987/3-HEW).

Ware, J.E., & Karnos, A. (1976). *Development and validation of scales to measure perceived health and patient role propensity: Volume II: Final report.* Springfield, VA: National Technical Information Service (publication no. 288-331).

Ware, J.E., 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., 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 (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 291-308). Durham, NC: Duke University Press.

Ware, J.E., & 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., 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 (Eds.), *Measuring functioning and well-being: The Medical Outcomes Study approach* (pp. 277-290). Durham, NC: Duke University Press.

Ware, J.E., Snyder, M.K., McClure, R.E., & Jarrett, J.M. (1972). *The measurement of health concepts.* Technical Report No. HCP-72-5, Publication No. PB-293-508/AS. Springfield, VA: National Technical Information Service.

Weinberger, M., Samsa, G.P., Hanlon, J.T., Schmader, K.E., Doyle, M.E., Cowper, P.A., Uttech, K.M., Cohen, H.J., & Feussner, J.R. (1991). An evaluation of a brief health status measure in elderly veterans. *Journal of the American Geriatric Society, 39*, 691-694.

Weinstein, M.C., Berwick, D.M., Goldman, R.A., Murphy, J.M., & Barsky, A. (1989). A comparison of three psychiatric screening tests using receiver operating characteristics (ROC) analysis. *Medical Care, 27*, 593-607.

Wells, K.B., Burnam, M.A., Rogers, W.H., & Camp, P. (1992). The course of depression in adult outpatients: Results from the Medical Outcomes Study. *Archives of General Psychiatry, 49*, 788-794.

Wells, K.B., Hays, R.D., Burnam, M.A., Rogers, W.H., Greenfield, S., & Ware, J.E. (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.* New York: Le Jacq Publishing Company.

Wexler, H.P., & Radosevich, D.M. (1992, May). *Health status questionnaire (SF-36) technical report.* InterStudy.

Williams, A.H., Ware, J.E., & Donald, C.A. (1981). A model of mental health, life events, and social supports applicable to general populations. *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 (ANCOVM2).* Grand Forks: Computer Center, University of North Dakota.

Winslow, R. (1992, July 7). Questionnaire probes patients' quality of life. *The Wall Street Journal*, pp. B1, B4.

Wu, A.W., Rubin, H.R., Matthews, W.C., Ware, J.E., Brysk, L.T., Hardy, W.D., Bozzette, S.A., Spector, S.A., & Richman, D.D. (1991). A health status questionnaire using 30 items from the Medical Outcomes Study: Preliminary validation in persons with HIV infection. *Medical Care, 29*, 786-798.

ANNOTATED BIBLIOGRAPHY

This bibliography includes abstracts of recently published journal articles that used the SF-36, SF-20, and MHI-5 measures. It is arranged alphabetically by first author. The descriptions of each article are largely taken from the author's abstract of the article, but in some instances, additional information was added to describe the use of the SF-36 or MHI-5 measure within the study. The SF-20 and SF-36 both include the MHI-5 scale and have several other items in common. For studies that used the SF-20 measure, only statistics related to the MHI-5 scale are discussed in this manual.

Berwick DM, Murphy JM, Goldman PA, Ware JE, Jr, Barsky AJ, Weinstein MC. Performance of a five-item mental health screening test. *Medical Care 29*(2):169-76, 1991.

The authors compared the screening accuracy of a short, five-item version of the Mental Health Inventory (MHI-5) with that of the 18-item MHI, the 30-item version of the General Health Questionnaire (GHQ-30), and a 28-item Somatic Symptom Inventory (SSI-28). Subjects were newly enrolled members of a health maintenance organization (HMO), and the criterion diagnoses were those found through use of the Diagnostic Interview Schedule (DIS) in a stratified sample of respondents to an initial, mailed GHQ. To compare questionnaires, the authors used receiver operating characteristic analysis, comparing areas under curves through the method of Hanley and McNeil. The MHI-5 was as good as the MHI-18 and the GHQ-30, and better than the SSI-28, for detecting most significant DIS disorders, including major depression, affective disorders generally, and anxiety disorders. Areas under curve for the MHI-5 ranged from 0.739 (for anxiety disorders) to 0.892 (for major depression). Single items from the MHI also performed well. In this population, short screening questionnaires, and even single items, may detect the majority of people with DIS disorders while incurring acceptably low false-positive rates. Perhaps such extremely short questionnaires could more commonly reach use in actual practice than the longer versions have so far, permitting earlier assessment and more appropriate treatment of psychiatrically troubled patients in primary care settings.

Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients. The floor phenomenon. *Medical Care 28*(12):1142-52, 1990.

The interest in measuring health status with survey instruments has not been matched with an analysis of their performance characteristics in the field. The Medical Outcome Study Short Form (MOS-20) was used to assess health outcomes among patients who were hospitalized in one of two public hospitals. MOS-20 and a series

of transition questions was added, which asked about changes in health, to patients admitted in the previous year. 414 completed surveys were received from 480 patients at baseline and follow-up data on 90% of these patients six months later. Baseline MOS-20 scores for study patients were significantly lower, corresponding to worse health, than previously reported outpatient and general population cohorts. While the direction of change on serial applications of the MOS-20 paralleled the patients' perception of change reported on transition questions, many patients who reported their health had become worse also recorded the lowest possible score on the MOS-20 at baseline. These *low* baseline MOS-20 scores prohibited the recognition of larger declines in function during the follow-up period. This floor in the response range creates an instrument bias against documenting a decline in health among severely ill patients, the group in which it may be most important to detect such a change.

Bjorner JB, Harper R, Jones NM, O'Cathain A, Thomas KJ, Usherwood T, Westlake L. Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *British Medical Journal 305*(6846):160-4, 1992.

Objectives — To test the acceptability, validity, and reliability of the Short Form 36 Health Survey questionnaire (SF-36) and to compare it with the Nottingham Health Profile. Design — Postal survey using a questionnaire booklet together with a letter from the general practitioner. Non-respondents received two reminders at two week intervals. The SF-36 questionnaire was retested on a subsample of respondents two weeks after the first mailing. Setting — Two general practices in Sheffield. Patients — 1980 patients aged 16-74 years randomly selected from the two practice lists. Main Outcome Measures — Scores for each health dimension on the SF-36 questionnaire and the Nottingham Health Profile. Response to questions on recent use of health services and sociodemographic characteristics. [This study used the U.K. version of the SF-36.] Results — The response rate for the SF-36 questionnaire was high (83%) and the rate of completion for each dimension was over 95%. Considerable evidence was found for the reliability of the SF-36 (Cronbach's alpha greater than 0.85, reliability coefficient greater than 0.75 for all dimensions except social functioning) and for construct validity in terms of distinguishing between groups with expected health differences. The SF-36 was able to detect low levels of ill health in patients who had scored 0 (good health) on the Nottingham Health Profile. Conclusions — The SF-36 is a promising new instrument for measuring health perception in a general population. It is easy to use, acceptable to patients, and fulfills stringent criteria of reliability and validity. Its use in other contexts and with different disease groups requires further research.

Cleary PD, Epstein AM, Oster G, Morrissey GS, Stason WB, Dabbussey S, Plachetka J, Zimmerman M. Health-related quality of life among patients undergoing percutaneous transluminal coronary angioplasty [published erratum appears in *Medical Care. 30*(1):76, 1992.]. *Medical Care. 29*(10):939-50, 1991.

A randomized clinical trial was recently conducted to investigate whether a new antiplatelet agent could prevent restenosis in patients who had undergone percutaneous transluminal coronary artery angioplasty (PTCA). Approximately 1,200 patients were enrolled at 13 separate clinical sites. To assess the impact of this intervention on health-related quality of life, a patient questionnaire for telephone administration was developed. This questionnaire focused attention on several specific dimensions likely to be important in this patient population: physical well-being, perceived health, emotional well-being, home management, work, recreation, and social and sexual functioning. This paper describes the instrument used in this trial and reports on its psychometric properties based on completed interviews with approximately 500 patients at study entry and 1 month after PTCA. [The patient questionnaire included the MHI-5 and General Health measures from the SF-36.]

Fryback DG, Dasbach EJ, Klein, R, Klein B, Dorn N, Peterson K, Martin AM, The Beaver Dam Health Outcomes Study. Initial catalog of health state quality factors, *Medical Decision Making, 13*:89–102, 1993.

The Beaver Dam Health Outcomes Study (BDHOS) is an ongoing cohort study of health status and health-related quality of life for a random sample of adults (age range at interview was 45 to 89 years; mean = 64.1 with s.d. = 10.8) in a community population. In a face-to-face interview lasting approximately an hour, each participant responds to several batteries of questions. Included are a history of chronic medical conditions, current medications, and past surgeries; the SF-36 (a general health-status questionnaire); the Quality of Well-Being Index; self-rated health status on a 5 point scale from "Excellent" to "Poor", and evaluation of current health using the method of time trade-offs. This paper presents results from 1356 interviews on these four principal measures. Mean scores are reported by sex, by age, and for persons reporting being affected by various medical conditions. Data from the BDHOS will provide researchers and policy makers a reference collection of vital statistics for health-related quality of life. Additionally, the data provide a way to compare results from studies that utilize different indices from among the four principal measures of the BDHOS.

Garratt AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF-36 Health Profile: An outcome measures suitable for routine use within the NHS? *British Medical Journal, 306*:1440–1444, 1993.

Objective — To assess the validity, reliability and acceptability of the SF-36 health profile as a measure of patient outcome in a broad sample of patients suffering from four common clinical conditions. Design — Posted questionnaire followed up, if necessary, by two reminders at two week intervals. Setting — Clinics and four training practices in the North East of Scotland. Patients — More than 1700 patients aged 16 to 86 years with one of four conditions: low back pain, menorrhagia, suspected peptic ulcer, and varicose veins; and a comparison sample of 900 members of the general population. Main outcome measures — The eight scales within the SF-36

health profile. [This study used the U.K. version of the SF-36.] Results — The response rate exceeded 75% in a patient population. The SF-36 satisfied rigorous psychometric criteria for validity and internal consistency. Clinical validity was demonstrated by the distinctive profiles generated for each condition, each of which differed from the general population in a predictable manner. Furthermore, SF-36 scores were lower in referred patients than in those not referred, and were closely related to general practitioners' perceptions of severity. Conclusions — These results provide support for the SF-36 as a potential measure of patient outcome within the NHS. The SF-36 appears acceptable to patients, internally consistent and a valid measure of the health status of a wide range of patients. Before it can be used in the new health service, however, its sensitivity to changes in health status over time must also be tested.

Gelberg L., Linn LS. Psychological distress among homeless adults. *Journal of Nervous & Mental Disease. 177*(5):291-5, 1989.

Recent studies have reported a high prevalence of mental illness among the homeless. As part of a community-based survey of 529 homeless adults, the authors developed and tested a model to increase understanding of the factors related to their psychological distress. Using a previously validated and reliable scale of perceived psychological distress, homeless adults were found more likely to report psychological distress than the general population (30% vs. 4%). Distress levels were not associated with most demographic or homeless characteristics or general appearance. However, distress was related to unemployment, greater cigarette and alcohol use, worse physical health, fewer social supports, and perceived barriers to obtaining needed medical care. Since mental, physical, and social health are strongly related among homeless adults, alleviating distress among them may be most effectively done by implementing a broad-based health services package coupled with employment programs provided in an accessible service delivery setting. [One measure of mental health used in this study was the MHI-5.]

Jenkinson C., Coulter A., Wright L. The SF-36 Health Survey Questionnaire: Normative Data from a Large Random Sample of Working Age Adults. *British Medical Journal. 306*:1436-1440, 1993.

Objectives — To gain population norms for the Short Form 36 Health Survey questionnaire (SF-36) in a large community sample, and to explore the instruments internal consistency and validity. Design — Postal survey using a questionnaire booklet, containing the SF-36 and a number of other items concerned with lifestyles and illness. A letter outlining the purpose of the study was included. Setting — The sample was drawn from Family Health Services Authority (FHSA) computerized registers for Berkshire, Buckinghamshire, Northamptonshire and Oxfordshire. Sample — The questionnaire booklet was sent to 13,042 randomly selected subjects between the ages of 17-65. This paper is based upon the 9,332 (72%) responses gained. Outcome measures — Scores for the eight dimensions of the SF-36. [This

study used the U.K. version of the SF-36.] Results — The response rate for the questionnaire booklet was 72%. Internal consistency of domains was found to be high. Normative data for the SF-36 is reported, by age and sex, and social class. Conclusion — The SF-36 is a potentially valuable tool in medical research. The normative data provided in this paper may further facilitate its validation and use.

Kantz ME, Harris WJ, Levitsky K, Waite JE Jr, Davies AR. Methods for assessing condition-specific and generic functional status outcomes after total knee replacement. *Medical Care. 30*(Suppl.):MS240-52, 1992.

Many assume that, relative to generic measures, condition-specific health measures are both more sensitive to the condition's severity and more specific because they are less affected by other conditions. The authors analyzed sensitivity and specificity of the generic SF-36, condition-specific scales based on the SF-36, and condition-specific measures based on the Knee Society's Clinical Rating System in a study of osteoarthritis patients following knee replacement. As hypothesized, knee-specific role function and pain measures were more specific than generic measures among patients with other comorbid conditions, and less so among patients with only knee problems. Physical function scales of both types were equally specific. Clinical indicators based on x-ray and range of motion were only weakly related to all measures of function.

Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. Comparative measurement sensitivity of short and longer health status instruments. *Medical Care. 30*(10):917-25, 1992.

Short measures of health status are used increasingly in health services research, yet their sensitivities to clinical change have not been compared with longer, established instruments. In this study, 5 health status measures were administered preoperatively and 3 months postoperatively to 54 patients undergoing total hip arthroplasty. These instruments included the Sickness Impact Profile (SIP) — an established, long measure — and 4 short forms: the SF-36, Functional Status Questionnaire, shortened Arthritis Impact Measurement Scales, and Modified Health Assessment Questionnaire. Scores for physical, psychological, and global dimensions were constructed by aggregating subscales. Sensitivity to change, or responsiveness, was expressed with the standardized response mean (SRM), calculated as the mean change in score divided by the standard deviation of the change in score. The sampling distribution of the SRM was estimated with a jackknife procedure. Preoperative scores were moderately to highly correlated across instruments. The physical and global dimension SRMs of the brief health status measures ranged from 0.85 to 1.27 and were as large as or larger than the corresponding SIP SRMs. The SIP had the highest SRM on the psychological dimension. None of the instruments was significantly more sensitive than the others at the critical value ($P = 0.005$) adjusted for multiple comparisons. The brief health status measures were equally or more responsive than the SIP after total hip arthroplasty in the

physical and global dimensions. Much larger samples are required to demonstrate statistically significant differences in SRMs among instruments.

Kurtin PS, Davies AR, Meyer KB, DeGiacomo JM, Kaufman ME. Patient-based health status measures in outpatient dialysis. Early experiences in developing an outcomes assessment program. *Medical Care, 30*(Suppl.):MS136-49, 1992.

This paper describes the initial development of a patient-based outcomes assessment program in an outpatient dialysis unit. This project presented four logistical and practical issues that are discussed in this paper: patient acceptance of quarterly administrations of a generic health status survey (the SF-36); timing of administration during dialysis session; respondent burden; and staff burden. Also discussed are three issues related to the clinical use of these assessments: medical record status of SF-36 data; use in clinical decision-making; and clinicians' responses to aggregate data from patient-based health status assessments. The investigation reported presents strong evidence of patient acceptance of the SF-36. Data collection problems reflected the nature of a busy dialysis unit, and most have been corrected. Considering functional status, the role functioning of dialysis patients is most adversely affected; among well-being measures, patients are most compromised by pain and lack of energy. Clinicians' reviews of these results point to the need for normative data, information about severity of primary and comorbid diseases, and knowledge of relationships between SF-36 scores and physiologic parameters to make clinical use of generic health outcome assessments.

Lancaster TR, Singer DE, Sheehan MA, Oertel LB, Maraventano SW, Hughes RA, Kistler JP. The impact of long-term warfarin therapy on quality of life. Evidence from a randomized trial. Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators. *Archives of Internal Medicine, 151*(10):1944-9, 1991.

To determine the effect of long-term warfarin sodium therapy on quality of life, 333 patients participating in a randomized, controlled trial of warfarin for the prevention of stroke in nonrheumatic atrial fibrillation were surveyed. No significant differences between warfarin-treated and control patients were found on well-validated measures of functional status, well-being, and health perceptions. For example, the summary score for health perceptions was 68.8 in the warfarin-treated vs. 66.6 in the control group (scale of 0 to 100; 95% confidence intervals for the difference, -1.6 to 6.0). In contrast, patients taking warfarin who had a bleeding episode had a significant decrease in health perceptions (-11.9; 95% confidence interval, -4.1 to -19.6). Warfarin therapy is not usually associated with a significant decrease in perceived health, unless a bleeding episode has occurred. Negative effects of warfarin treatment on health perceptions may be balanced by confidence in its protective effects. [The SF-36 was used as the measure of functional status, role functioning, and well-being.]

Lansky D, Butler JB, Waller FT. Using health status measures in the hospital setting: from acute care to 'outcomes management'. *Medical Care* 30(Suppl.):MS57–73, 1992.

In recent years, employers, insurers, and government purchasers have paid increased attention to the measurement of patient outcomes and health status. Such interest is stimulated less by policy or quality concerns than by the need to reduce health care expenditures. Any expected benefits which might accrue from health status measurement will require active participation by community hospitals and their affiliated physicians. St. Vincent Hospital and Medical Center in Portland, Oregon has begun hospital-wide use of outcomes measurement systems. This study presents case studies of outcomes measurement for low back pain and total hip replacement, summarizes the hospital's objectives in implementing such measures, and identifies several strategies for successful adoption of health status measures in community practice. [The SF-36 was used as a measure of generic health status in low back pain and total hip replacement patients.]

McHorney CA, Kosinski M, Ware JE. Comparisons of the costs and quality of norms for the SF-36 survey by mail versus telephone interview: Results from a national survey. *Medical Care*. In review.

Many health status surveys have been designed for mail, telephone, or in-person administration. However, with rare exception, investigators have not studied the effect survey mode of administration has on how respondents assess their health and other important parameters, such as response rates, non-response bias, and data quality, that can affect the generalizability of results. Using a national sampling frame of non-institutionalized adults from the General Social Survey, adults were randomly assigned to a mail survey (80%) or a computer-assisted telephone survey (20%). The surveys were designed to provide national norms for the SF-36 Health Survey. Total data collection costs per case for the telephone survey ($47.86) was 77% higher than that for the mail survey ($27.07). A significantly higher response rate was achieved among respondents randomly assigned to the mail (79.2%) than telephone survey (68.9%). Non-response bias was evident in both modes, but, with the exception of age, was not different between modes. The rates of missing responses was higher for mail than telephone respondents. Health ratings based on the SF-36 survey are provided for use in interpreting individual and group scores. Results are discussed in light of the trade-offs involved in choosing a survey methodology for health status assessment applications.

McHorney CA, Ware JE. Construction and Validation of an Alternate Form General Mental Health Scale for the MOS SF-36 Health Survey. Forthcoming.

Alternate-form general mental health scales are useful for clinical trials or health services research requiring repeated administrations of patient reports of mental health over a short interval of time. They can also be used to estimate score relia-

bility using the alternate-form methodology. Data from the Medical Outcomes Study were used to evaluate seven alternate forms of the SF-36 general mental health scale (MHI-5). Well-established psychometric criteria were used to select the best alternate form and to estimate the reliability of the MHI-5 using the alternate-form methodology. Two of the seven scales best satisfied a crucial design requirement of item balancing and were subject to further psychometric evaluation. One scale best satisfied an array of criteria pertaining to distributional characteristics of items and scales (including equivalence of means and standard deviations across diverse groups), estimates of reliability, homogeneity, and internal consistency, and comparative tests of empirical validity. Using the alternate-form methodology of estimating reliability, results suggest that the internal-consistency method underestimates the reliability of the MHI-5 by 3%. The methodology presented here should prove useful to others interested in constructing alternate forms, and the alternate form recommended here should prove useful across many health status assessment applications.

McHorney CA, Ware JE, Lu JFR, Sherbourne CD. The MOS SF-36 Short-Form Health Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions, and Reliability Across Diverse Patient Groups. *Medical Care*. In press.

The widespread use of standardized health surveys is predicated on the largely untested assumption that scales constructed from those surveys will satisfy minimum psychometric requirements across diverse population groups. Data from the Medical Outcomes Study (MOS) were used to evaluate data completeness and quality, test scaling assumptions, and estimate internal-consistency reliability for the eight scales constructed from the MOS SF-36 Health Survey. Analyses were conducted among 3,445 patients and were replicated across 24 subgroups differing in sociodemographic characteristics, diagnosis, and disease severity. For each scale, item-completion rates were high across all groups (88% to 95%), but tended to be somewhat lower among the elderly, those with less than a high school education, and those in poverty. On average, surveys were complete enough to compute scale scores for over 96% of the sample. Across patient groups, all scales passed tests for item-internal consistency (97% passed) and item-discriminant validity (92% passed). Reliability coefficients ranged from a low of .65 to a high of .94 across scales (median = .85) and varied somewhat across patient subgroups. Floor effects were negligible except for the two role disability scales. Noteworthy ceiling effects were observed for both role disability scales and the social functioning scale. These findings support the use of the SF-36 survey across the diverse populations studied and call attention to population groups in which use of standardized health status measures may be more or less problematic.

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

Cross-sectional data from the Medical Outcomes Study (MOS) were analyzed to test the validity of the MOS 36-Item Short-Form Health Survey (SF-36) scales as measures of physical and mental health constructs. Results from traditional psychometric and clinical tests of validity were compared. Principal components analysis was used to test for hypothesized physical and mental health dimensions. For purposes of clinical tests of validity, clinical criteria defined mutually exclusive adult patient groups differing in severity of medical and psychiatric conditions. Scales shown in the components analysis to primarily measure physical health (physical functioning and role limitations-physical) best distinguished groups differing in severity of chronic medical condition and had the most pure physical health interpretation. Scales shown to primarily measure mental health (mental health and role limitations-emotional) best distinguished groups differing in the presence and severity of psychiatric disorders and had the most pure mental health interpretation. The social functioning, vitality, and general health perceptions scales measured both physical and mental health components, and, thus, had the most complex interpretation. These results are useful in establishing guidelines for the interpretation of each scale and in documenting the size of differences between clinical groups that should be considered very large.

McHorney CA, Ware JE, Rogers W, Raczek AE, Lu JF. 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.): MS253-65, 1992.

This study estimated the validity and relative precision (RP) of four methods (MOS long- and short-form scales [SF-20 and SF-36], global items, and COOP Poster Charts) in measuring six general health concepts. The authors also tested whether and how precisely each method discriminated relatively well adult patients (N=638) from those with only severe chronic medical (N=168) and only psychiatric conditions (N=163), as clinically defined. For comparisons between the well group and both medical and psychiatric groups, RP estimates favored long-form over short-form, multi-item scales, and favored multi-item scales over single-item global measures and poster charts. In relation to long form, short-form multi-item scales achieved a median RP of .93; RP estimates for global items and poster charts were .81 and .67, respectively. Variations in RP across methods and concepts were linked to differences in the coarseness of measurement scales, reliability, and content (including the effects of chart illustrations). These variations in RP have implications for the interpretation of scores, the statistical power of comparisons between clinical groups, and the size of confidence intervals around individual patient scores.

Nerenz DR, Repasky DP, Whitehouse FW, Kahkonen DM. Ongoing assessment of health status in patients with diabetes mellitus. *Medical Care. 30*(Suppl.):MS112-24, 1992.

In 1990, the Division of Endocrinology and Metabolism of Henry Ford Hospital established an Outcomes Management data base for patients with Type I and Type II diabetes. A first cohort of 117 patients completed a baseline and 6-month follow-up assessment; a second cohort of 116 patients completed the baseline assessment. Assessment at each time point includes: the Short Form-36 Questions (SF-36) health status instrument; a set of clinical variables known as the Diabetes TyPE scale Form 2.2 abstracted from the medical record; and the physicians' ratings of patient's health status along the major dimensions of the SF-36. Success with both face-to-face and mailed administration of the SF-36 has been good, with response rates of over 85% using both methods. Comparison of patient and physician ratings of patient health status indicated a significant discrepancy on ratings of general health status, with physicians' ratings higher than those of patients themselves. "Tight" glycemic control (as measured by glycosylated hemoglobin) was associated with somewhat lower ratings on the various SF-36 dimensions for all patients in the first cohort and for Type I patients in the second cohort. However, this effect did not seem to be attributable to those features of a complex regimen used to achieve tight control, but rather reflected a complex combination of age, education level, and number of daily injections associated with achieving good control.

Parkerson GR Jr, Broadhead WE, Tse CK. Comparison of the Duke Health Profile and the MOS Short-form in healthy young adults. *Medical Care. 29(7):*679-83, 1991.

Two of the brief health measures available today are the Duke Health Profile (DUKE) and the Medical Outcomes Study Short-form (SF-20). Because these two are similar in length and conceptualization, it is of interest to compare them in same population. The current study accomplished this in a young and healthy group of medical students (N=101). Six of the ten DUKE measures (i.e., physical health, mental health, social health, perceived health, pain, and disability) were selected as appropriate for direct comparison with the six SF-20 measures. This study supports the convergent and discriminant validity of all but the physical and social measures of both instruments. However, because of the marked differences in content of these two DUKE and SF-20 measures, the finding of low score correlations is more likely a reflection of dissimilar content in measures with similar titles than of the invalidity of either measure. This study shows similarities and differences in the two instruments with regard to their sensitivity for detection of health variations in a basically well population. The DUKE appears to have greater sensitivity than the SF-20 for the physical and social measures, while the SF-20 has greater sensitivity for health perception and pain. Both the DUKE and SF-20 detect differences well for mental health, but neither is very sensitive for disability or role function. Both the DUKE and the SF-20 have sufficient reliability, validity, utility, and conceptual strength to warrant further testing in the quest for health outcome measures suitable for widespread clinical use.

Phillips RC, Lansky DJ. Outcomes Management in Heart Valve Replacement Surgery: Early Experience. *Journal of Heart Valve Disease*. 1:42-50, 1992.

Recent efforts at reducing health care expenditures and practice variations have focused attention on treatment appropriateness, patient preference and quality of life as important elements of treatment evaluation. These characteristics of medical treatment can be assessed by administration of a structured, valid health status assessment questionnaire before and at a fixed intervals following treatment, and standardized scores may be compared to normal scores derived from healthy untreated populations. One hundred valve replacement patients completed a standardized questionnaire, the SF-36, preoperatively and at 1- and 6-months postoperatively. Preoperatively, valve replacement patients are substantially impaired in their physical capacity, their ability to function in important roles, and their vitality. At one month, they report significant decrements in social and role functioning and in pain which is attributed to the trauma of surgery and recovery. At six months, mean scores approximate those for sex- and age-matched normals except for persistent moderate impairment in role functioning. Post-operative functional impairment is greatest for those undergoing mitral valve replacement, particularly when coronary artery grafts are performed concurrently, and for patients receiving bioprosthetic valves. Structured health status assessment provides a useful adjunct to other methods of assessing clinical status and evaluating treatment outcomes.

Read JL, Quinn RJ, Hoefer MA. Measuring Overall Health: An Evaluation of Three Important Approaches. *Journal of Chronic Disease*. 40(Suppl.):7S-21S, 1987.

There is growing recognition that meaningful measures of health-related quality of life must be used to evaluate health care interventions. Examined the practicality and validity of three promising measures of overall health: the General Health Rating Index (GHRI), the Quality of Well-Being Scale (QWB), and the Sickness Impact Profile (SIP). Practicality was assessed in terms of interviewer training required, administration time, and respondent burden. Content validity, convergent construct validity, and tests of discriminant validity were also evaluated. Although differing in theory and application, each instrument performed according to the claims of the developers and could provide useful, valid data on overall health. The GHRI may be preferred where brief, self-administered forms are required; the QWB has advantages when health assessments are used to calculate cost-effectiveness, and the SIP is a versatile, easy to understand measure dealing with a wide range of specific dysfunctions. It is worth the required effort to include well-studied measures such as these in any trial intended to provide definitive information on the effectiveness of health care interventions. [The GHRI is the long-form of the SF-36 General Health scale, and the MHI-5 scale was used as a criterion variable in validating the three methods.]

Siu AL, Reuben DB, Hays RD. Hierarchical measures of physical function in ambulatory geriatrics. *Journal of the American Geriatrics Society*. 38(10):1113-9, 1990.

Brief and uncomplicated methods for obtaining information on functional status would facilitate the assessment of older patients. Evaluated the potential usefulness, reliability, and validity of four hierarchical measures of physical function in 123 elderly subjects seen in four ambulatory geriatrics settings. Although the vast majority (83.2%) of subjects were fully independent on the Katz Activities of Daily Living Scale, a broader scope of functional difficulty was reported on the Spector-Katz, five-item OARS, and Rosow-Breslau scales. The three scales all had either borderline or more acceptable coefficients of scalability (0.57-0.77); the hierarchical order of items was not observed in 5.3% to 13.6% of subjects. Combining items from these established measures resulted in two new scales with acceptable scalability and construct validity; however, some errors in item order persisted. Although their ease of administration is clearly advantageous, clinicians using short hierarchical scales to assess functional status of older patients should be aware of their limitations. [This study included the MHI-5 and General Health scale from the SF-36 in order to assess the validity of the new measures.]

Stewart AL, Greenfield S, Hays RD, Wells K, Rogers WH, Berry SD, McGlynn EA, Ware JE Jr. Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study [published erratum appears in *JAMA 262*(18):2542, 1989.]. *Journal of the American Medical Association. 262*(7):907-13, 1989.

Enhancing daily functioning and well-being is an increasingly advocated goal in the treatment of patients with chronic conditions. The functioning and well-being of 9,385 adults were evaluated at the time of office visits to 362 physicians in three U.S. cities, using brief surveys completed by both patients and physicians. For eight of nine common chronic medical conditions, patients with the condition showed markedly worse physical, role, and social functioning; mental health; health perceptions; and/or bodily pain (as measured by the SF-20) compared with patients with no chronic conditions. Each condition had a unique profile among the various health components. Hypertension had the least overall impact; heart disease and patient-reported gastrointestinal disorders had the greatest impact. Patients with multiple conditions showed greater decrements in functioning and well-being than those with only one condition. Substantial variations in functioning and well-being within each chronic condition group remain to be explained.

Vickrey BG, Hays RD, Graber J, Rausch R, Engel J Jr, Brook RH. A health-related quality of life instrument for patients evaluated for epilepsy surgery. *Medical Care, 30*(4):299-319, 1992.

The goals of surgery in treating intractable epilepsy are to eliminate seizures and improve quality of life. This report describes the development of the Epilepsy Surgery Inventory (ESI)-55, a 55-item measure of health-related quality of life for epilepsy patients. The ESI-55 includes the following scales (number of items in parentheses): health perceptions (9), energy/fatigue (4), overall quality of life (2), social func-

tion (2), emotional well-being (5), cognitive function (5), physical function (10), pain (2), and three separate scales of role limitations due to emotional, physical, or memory problems (3 items each). Also included is one change in health item. The ESI-55 was completed by 89% of 224 adults who had undergone a protocol evaluation for epilepsy surgery since 1974. Alpha internal consistency reliability coefficients ranged from 0.76 to 0.88 except for social function (alpha = 0.68). Multitrait scaling analyses supported item discrimination across scales. Factor analysis confirmed previously identified mental and physical health factors, and yielded a third factor defined by cognitive function and role limitations scales. Construct validity was supported by correlations of the ESI-55 with a mood profile instrument. Analysis of ESI-55 scale scores by seizure classification showed that the 44 patients who were seizure-free following surgery scored higher than did 55 patients who continued to have seizures (P less than 0.05 for all comparisons); 43 patients having seizures without loss of consciousness scored in between. Results of this study indicate that the ESI-55 is reliable, valid, and sensitive to differences in seizure status. [The SF-36 was included as part of the ESI-55 measure.]

Wachtel T, Piette J, Mohr V, Stein M, Fleishman J, Carpenter C. Quality of Life in Persons with Human Immunodeficiency Virus Infection: Measurement by the Medical Outcomes Study Instrument. *Annals of Internal Medicine. 116*:129-137, 1992.

Objective — To assess the reliability and validity of the Medical Outcomes Study (MOS) Short Form Health Survey (SF-20) as an indicator for quality of life in patients infected with the human immunodeficiency virus (HIV). Design — Patient interview survey. Setting — The AIDS Health Services Program in seven sites: Newark and Jersey City, New Jersey; Nassau County, New York; Atlanta, Georgia; Dallas, Texas; Fort Lauderdale and Miami, Florida; New Orleans, Louisiana, and Seattle, Washington. Patients — Patients (520) with HIV infection receiving health services at one of the above sites. Measurements — All components of the SF-20 were included in the interview. Minor modifications were made to adapt the survey to the particular circumstances of the study. Measured sociodemographic characteristics included age, sex, race, intravenous drug use, and education. Symptoms were assessed by closed-ended questions concerning memory, seizure, weakness or numbness, fever, chills, diaphoresis, dyspnea, diarrhea, and weight loss. Information on the frequency of symptoms was also collected. History of Pneumocystis carinii pneumonia and Kaposi sarcoma was noted. Main Results — The sociodemographic characteristics resemble those of patients with the acquired immunodeficiency syndrome (AIDS) reported to the Centers for Disease Control (CDC): mean age, 36; men, 89%; nonwhite, 31%; intravenous drug use, 34%. Neurologic symptoms (memory trouble, seizures, weakness or numbness) occurred in 71% of patients; constitutional symptoms (fever, chills, night sweats, weight loss) in 69%; dyspnea in 50%, and diarrhea in 47%. Although older age, female sex, nonwhite race, and intravenous drug use were associated with lower MOS scores in several areas, the

strongest single or adjusted indicator of lower MOS scores was the presence of symptoms. Finally, patients with HIV infection had significantly lower scores than did previously reported patients with other chronic medical conditions ($F < 0.001$). Conclusions — The SF-20 is a reliable measure of quality of life for patients with HIV infection. These patients tend to have low scores, suggesting validity of the survey. The SF-20 is extremely sensitive to the effect of symptoms, which suggests that it might be useful as a quality-of-life indicator for AIDS clinical drug trials.

Ware JE Jr. Standards for validating health measures: Definition and content. *Journal of Chronic Diseases. 40(6)*:473-480, 1987

Adherence to standards for judging the content validity of health measures and for labeling them is needed for the field of health assessment to proceed in an orderly fashion. This paper discusses the dimensionality of health and the range of health states that can be measured within each dimension. These two attributes of published definitions of health are used to derive minimum standards for judging the validity of health measures in terms of their content. Five generic health concepts are defined: physical health, mental health, social functioning, role functioning, and general health perceptions. Items from widely used health measures are presented to clarify distinctions among these concepts and the different health states they encompass. It is recommended that labels be assigned to health measures in a manner consistent with their content and other evidence of validity. [These standards of content were the foundation of the selection of content for the SF-36.]

Ware JE Jr Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care. 30*(6):473-83, 1992.

A 36-item short-form (SF-36) was constructed to survey health status in the Medical Outcomes Study. The SF-36 was designed for use in clinical practice and research, health policy evaluations, and general population surveys. The SF-36 includes one multi-item scale that assesses eight health concepts: (1) limitations in physical activities because of health problems; (2) limitations in social activities because of physical or emotional problems; (3) limitations in usual role activities because of physical health problems; (4) bodily pain; (5) general mental health (psychological distress and well-being); (6) limitations in usual role activities because of emotional problems; (7) vitality (energy and fatigue); and (8) general health perceptions. The survey was constructed for self-administration by persons 14 years of age and older, and for administration by a trained interviewer in person or by telephone. The history of the development of the SF-36, the origin of specific items, and the logic underlying their selection are summarized. The content and features of the SF-36 are compared with the 20-item Medical Outcomes Study short-form (SF-20).

Weinberger M. Samsa GP. Hanlon JT. Schmader K. Doyle ME. Cowper PA. Uttech KM. Cohen HJ. Feussner JR. An evaluation of a brief health status measure in elderly veterans. *Journal of the American Geriatrics Society. 39*(7):691-4, 1991.

Objective — To examine the feasibility of a brief 36-item health status measure (the SF-36) in elderly male veterans, by comparing it with the 136-item Sickness Impact Profile. Design — Cross-sectional study in which all subjects completed both measures in a random order. Setting — Durham VAMC General Medicine and Geriatrics Clinics. Patients — Convenience sample of 25 male veterans aged 65 and older (mean age = 73.5 years; 68% white; 68% currently married; mean annual income = $7,000). Main Outcome Measures — Two well-validated health status measures, the Sickness Impact Profile and the SF-36. Results — The SF-36 took less time to administer than the Sickness Impact Profile in both the Geriatrics Clinic (mean: 15 vs. 33 minutes) and General Medicine Clinic (mean: 14 vs. 21 minutes). Although SIP scores consistently displayed a more optimistic picture of respondents' health compared with the SF-36, the two instruments were highly correlated: overall functioning ($r = 0.73$), physical functioning ($r = 0.76$), and social functioning ($r = 0.67$). Conclusions — These two measures provide a similar ranking of elderly male veterans' health status. The significantly shorter administration time of the SF-36 is an attractive feature for researchers and clinicians interested in assessing health status.

Wells KB, Burnam MA, Rogers W, Hays R, Camp P. The course of depression in adult outpatients. Results from the Medical Outcomes Study. *Archives of General Psychiatry*. 49(10):788-94, 1992.

Objective — To compare the course of depression during a 2-year period in adult outpatients (N=626) with current major depression, dysthymia and either both current disorders ("double depression") or depressive symptoms with no current depressive disorder. Methods — Depressed patients visiting 523 clinicians (mental health specialists and general medical providers) were identified using a two-stage screening procedure including the Diagnostic Interview Schedule. The course of depression was assessed in 2 follow-up years with a structured telephone interview based on the format of the Diagnostic Interview Schedule. Results — Baseline severity of depressive symptoms was greatest in patients with double depression, but initial functional status was poor in those with dysthymia with or without concurrent major depression. Patients with dysthymia had the worst outcomes, those with current major depression alone had intermediate outcomes, and those with subthreshold depressive symptoms had the best outcomes. Even the latter group, however, had a high incidence (25%) of major depressive episode over 2 years. Initial depression severity and level of functional status accounted for more explained variance in outcomes than did type of depressive disorder. Conclusions — The findings emphasize the poor prognosis associated with dysthymia even in the absence of major depression; the prognostic significance of subthreshold depressive symptoms; and the clinical significance of assessing level of severity of symptoms as well as functional status and well-being, regardless of type of depressive disorder. [The SF-36 was shown to predict the course of depression during the 2-year follow-up period.]

Wu AW, Rubin HR, Mathews WC, Ware JE Jr, Brysk LT, Hardy WD, Bozzette SA, Spector SA, Richman DD. A health status questionnaire using 30 items from the Medical Outcomes Study: Preliminary validation in persons with early HIV infection. *Medical Care*. 29(8):786-98, 1991.

Many current health status instruments either are too long to use in many acquired immune deficiency syndrome (AIDS) clinical trials or omit important concepts. In this study, human immunodeficiency virus (HIV)-relevant items developed for the Medical Outcomes Study (MOS) from subscales for cognitive function, energy/fatigue, health distress, and a single quality of life item were added to a portion of the MOS Short-Form General Health Survey (SF-20). The resulting 30-item questionnaire reliably and distinctly measured ten aspects of health and took less than 5 minutes to complete. To test its validity, this modified measure was used to compare the health of 73 subjects with asymptomatic HIV infection and 44 with early AIDS-related complex (ARC). Compared with ARC subjects, asymptomatic individuals reported superior overall health, less pain, and better physical function, role function, cognitive function, and quality of life (rank-sum, P less than 0.02). Asymptomatic subjects' scores were higher on most subscales than the age-adjusted scores of MOS outpatients with hypertension, diabetes, recent myocardial infarction, or depression; ARC patients scored lowest to hypertensive patients. This instrument, containing a subset of the MOS measures of health-related quality of life, may be a useful outcome measure for AIDS clinical trials. [The questionnaire used in this study has two scales identical to the SF-36, the MH-5 and the Vitality scales, along with several identical individual items.]

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: a temporary state or condition.*

Affect: emotional or feeling state.*

Alpha (Coefficient): Cronbach's Alpha, an estimate of internal-consistency reliability based on the average inter-item correlation and number of items.*

Alternative forms: administration of two versions of a test that have been shown to be equivalent to elicit information about the same characteristic or variables.†

Alternate-form reliability: estimate of reliability based on the correlation between two forms constructed to be equivalent (i.e., equal mean, variance, and content) measures of the same concept.

Anxiety/depression: feelings of anxiety, nervousness, tenseness, depressions, moodiness, 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. Distinct from well-being which pertains to subjective, internal states that cannot be directly observed.†

Bodily pain: the intensity, duration, and frequency of bodily pain and limitations in usual activities due to pain, such as headaches or backaches.†

Ceiling effect: the percentage of respondents who score at the highest possible scale level (see also Floor effect).

Chronic: a state or condition that is persistent or long lasting, usually more than 3 months.*

Clinical trial: a study, usually a randomized group experiment, usually designed to evaluate treatment; referred to as a "controlled trial" if a comparison with another treatment or placebo is involved.

Closed-ended questions: a question that contains specific response options (e.g., yes or no).*

Coarse: 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.

Component: part of a larger concept or construct. For example, anxiety is a component of psychological distress.*

Concurrent validity: a form of 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 describes 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

---

* Definitions followed by this symbol are reprinted from the glossary published in Stewart & Ware (1992).

† Definitions followed by this symbol are edited from definitions published in Bergner & Ware (1993).

lower limit, associated with a particular probability (e.g., the 95% confidence interval around a mean, which 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; also 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: 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 (such as in a multitrait-multimethod matrix of correlations with reliability coefficients in the diagonal).*

Corrected 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 (multi-trait 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. Also called a Pearson product-moment correlation.

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.*

Descriptive statistics: indicators that characterize score distributions for a particular sample such as the mean, standard deviation, range, skewness, and percent 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; for example, arthritis or diabetes.†

Domain: any one of the 12 dimensions of health first defined by Campbell: community, education, family life, friendships, health, housing, marriage, nation, neighborhood, self, standard of living, work.†

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: analysis of data yields evidence that two measures do not have the same interpretation.*

External validity: representativeness or generalizability of results.*

Face-to-face administration: in person

administration of a questionnaire by an interviewer as opposed to over the telephone (see Telephone administration).

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.*†

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. 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; extent to which the structural relationship among measures corresponds to their underlying theoretical framework.*

Floor effect: the percentage of respondents scoring at the lowest possible scale level (see also Ceiling effect).

Frequency distribution: the number of respondents who score at each level of a scale.*

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, resistance to illness.†

General population: refers to 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; a category of health measures that are valued by 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.*

Health: The World Health Organization definition: A state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.

Health assessment: a standardized procedure used to quantify an individual's health.†

Health burden: the total impediment to physical, mental, and social functioning and well-being in the personal evaluation of health.

Health dimension: theoretical component of health such as physical or mental.*

Health framework: 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: randomized experiment conducted by The RAND Corporation in 1974 to 1981.*

Health outlook: expectation for health in the future, for example, as measured by the Health Outlook scale in the Health Perceptions questionnaire (Ware, 1976).

Health-related quality of life: personal health status. It usually refers to aspects of our lives that are dominated or significantly influenced

by our mental or physical well-being.†

HIE: see Health Insurance Experiment.

Index: an aggregation of two or more distinct health measures into an overall summary measure.*

Internal consistency: the extent with 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 (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).*

Interval scale: a scale in which the distances between all levels along the scale have known numerical values.*

Item: a single question or statement and its standardized set of responses.*

Item analysis: evaluation of the psychometric attributes of an item such as its descriptive statistics, correlation with the scale, convergence, and discrimination for purposes of combining into scales.*

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).*

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 elicits statistically useful and consistent information from individuals. Synonymous with questionnaire, tool, survey, or instrument.*†

Measurement error: random error occurring in the measurement of an attribute; portion of observed score that is not true score.*

Median: the midpoint of a particular score distribution; marking the 50th percentile.*

Medical Outcomes Study: A study launched in 1988 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 of patient-based assessment methods for assessing health outcomes.

Mental health: a person's emotional, cognitive, and intellectual status.†

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).*

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 allow for broad comparisons and interpretations of unlike populations.†

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.†

Pain: see Bodily pain.

PAQ: 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 were selected to become the MOS panel sample.*

Personal evaluations: a respondent's own rating of his/her health, such as that based on the widely used rating of health in terms of "excellent" to "poor" (see SF-36 General Health scale) and the Health Perceptions Questionnaire (Ware, 1976).

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: 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 measure is capable of detecting small differences.*

Predictive validity: a form of construct validity in which the hypothesis being tested is whether the measure can forecast the probability of another event (e.g., use of services) in future years.*

Psychological distress: frequency and intensity of negative psychological states including anxiety, depression, 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 or 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 by emotional distress.*

Quality of life: an evaluation of all aspects of our lives, including, for example, 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. It includes standardized questions and response choices. Synonyms are measure, test, tool, survey, or instrument.†

Range: the full gamut of levels for a given variable or domain, for example, from well-being to deathly ill.†

Rating: data obtained from a respondent that are subjective, including an evaluative component. Ratings are based on 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.*

Recall period: the interval of time the respondent is instructed to consider in reporting or rating a given health phenomenon (e.g., depression "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 give an objective account of an occurrence, not influenced by emotional or personal prejudice.†

Respondent: person answering questions or completing a survey.*

Respondent burden: the amount of time and effort required of those completing questionnaires.*

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 (respective of content (e.g., the tendency to present oneself in 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 social responsibility, such as working at a job, housework, schoolwork, child care, community activities, and volunteer work.†

Scale: an item or aggregation of one or more items (questions) to elicit information concerning a variable or domain; or may be used to refer to a graded series of tests. Combined in such a way 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 will be grouped together to create a scale. Scales may then be grouped together to provide an index or indices.*†

Scale level: a point on the scale that defines a particular rank order or quantity of the concept being measured (e.g., the 21 levels of the Physical Functioning scale).

Scale score: the result of the aggregation and manipulation of the responses to the individual items in a scale.

Scoring rules: numbers assigned to item responses and if applicable, the formula for their aggregation in a scale or index.*

Self-administration: 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 (Ware & Sherbourne, 1992) or the 17-item Duke Health Profile (Parkerson et al., 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, sexual functioning. Often separated into two areas: (1) whether and with what frequency social contacts are occurring, and (2) 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 is supposed to be obtaining 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: 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: determines the confidence interval around an individual score; equals the standard deviation times the square root of one minus the score reliability.*

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.*

Telephone administration: interviewer administration of a questionnaire over the telephone, as opposed to in person (see 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. In the MOS, the following tracer conditions were defined: hypertension, diabetes, heart disease (myocardial infarction, congestive heart failure), and depression.*

Validity: the extent to which a measure measures what it is supposed to and does not measure what it is not supposed to (See Concurrent validity, Construct validity, Content validity, Criterion validity, Face validity, Predictive 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.*

INDEX

N

O

P

Q

R

W