The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs
Author(s): Colleen A. McHorney, John E. Ware, Jr., Anastasia E. Raczek
Source: *Medical Care*, Vol. 31, No. 3 (Mar., 1993), pp. 247-263
Published by: Lippincott Williams & Wilkins
Stable URL: http://www.jstor.org/stable/3765819
Accessed: 02/06/2011 02:12

MEDICAL CARE
Volume 31, Number 3, pp 247–263

The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs

COLLEEN A. MCHORNEY, PHD, JOHN E. WARE, JR., PHD, AND ANASTASIA E. RACZEK, AB

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. Key words: health status assessment; health-related quality of life; construct validity; MOS SF-36 health survey. (Med Care 1993; 31:247–263)

A major goal of the Medical Outcomes Study (MOS) was to advance the state-of-the-art of methods used for routine monitoring of patient outcomes in medical practice and clinical research.[1] The value of a standardized and more practical short-form questionnaire for measuring general health concepts was demonstrated by the 20-item MOS Short-Form General Health Survey (SF-20).[2,3] That form has been used in com-

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

This research was supported by the Henry J. Kaiser Family Foundation, Menlo Park, CA (Grant No. 85-6515); the Functional Outcomes Program of the Henry J. Kaiser Family Foundation of the Health Institute, New England Medical Center, Boston MA (Grant No. 91-013); and the National Institute on Aging, Bethesda, MD (Grant No. AG07508). Data come from the Medical Outcomes Study (MOS), which is also supported by the Robert Wood Johnson Foundation, The Pew Charitable Trusts, The Agency for Health Care Policy and Research, and the National Institute of Mental Health.

Address correspondence to: Colleen A. McHorney, PhD, The Health Institute, Box 345, New England Medical Center, 750 Washington Street, Boston, MA 02111.

This article was originally submitted for consideration on 11-19-92. It was accepted for publication on 10-21-92 following the completion of all necessary reviews and/or revisions.

parisons of patients with both medical and psychiatric conditions[4–10] and in comparisons with general populations.[3,11,12]

The MOS 36-Item Short-Form Health Survey (SF-36) was constructed to broaden the health concepts measured and improve measurement precision for each concept over that achieved by the SF-20. Noteworthy improvements include the addition of items tapping vitality, better representation of the domain of general health perceptions, distinguishing between physical and mental causes of role limitations, and increased measurement precision for physical, role, social, and bodily pain scales.[13] The eight SF-36 measures constitute the core set of generic health outcomes assessed in the longitudinal component of the MOS.

A continuous aspect of evaluating both the SF-20 and SF-36 surveys has been accumulating evidence for validity—the fidelity with which a scale measures what it purports to measure.[14] Validity is the basis of the interpretability and meaningfulness of scores.[15] One traditional psychometric approach to validation is through components or factor analysis, which gauges the congruence between the hypothesized constructs of interest and scales constructed to measure those attributes. However, traditional psychometric tests often do not explicitly address other key validity issues, such as the relevance of scores to the intended use of a measure and the "quality of inferences"[16] derived from specific applications. A more unified approach to validity emphasizes both kinds of tests: 1) psychometric tests, which are the foundation of scale construction and scoring; and 2) applied tests of relevance and usefulness that approximate a particular use of the measure.[15]

Construct validation, the accumulation of evidence of validity in relation to theoretical constructs, requires three steps:[17] 1) specifying the domain of variables, i.e., preparing a blueprint for constructs; 2) establishing the internal structure of observed variables; and 3) verifying theoretical relationships between scale scores and external criteria. In this article, we focus on the second and third aspects of construct validity. The conceptual blueprint and rationale underlying item selection for the eight SF-36 health concepts has been previously reported.[13]

Briefly, the SF-36 survey was constructed to achieve two well-accepted standards of comprehensiveness: 1) representation of multidimensional health concepts; and 2) measurement of the full range of health states, including levels of well-being and personal evaluations of health. Accordingly, the SF-36 measures the health concepts most frequently included in widely used health surveys (physical, role, and social functioning, mental health, and general health perceptions) as well as two additional concepts strongly supported by empirical work (bodily pain and vitality). To achieve depth of measurement for each health concept, i.e., measurement precision, short-form multi-item scales were constructed from a subset of items shown to best reproduce a full-length and well-validated scale.

The full-length measures of general health status that preceded the SF-36 were constructed to capture two major dimensions of health—physical and mental—and these dimensions have been empirically confirmed in both general and patient populations.[18,19] We replicated this important psychometric test of construct validity for the SF-36 measures. We also went beyond psychometric tests and evaluated whether similar patterns of results are observed when the scales are examined in relation to clinical criteria of physical and mental health status. Finally, we compared results from psychometric and clinical criteria to determine the extent to which conclusions about the convergent and discriminant validity of each scale are replicated across criteria. Because interest in using general health scales in clinical research and medical practice is growing rapidly, information about validity in relation to clinical criteria is crucial to document the size of small and large differences

and to advance understanding of how these differences should be interpreted.

Methods

Sample and Data Collection

The data for this analysis came from MOS forms completed by patients and physicians and from health examinations administered in 1986–1987. Details on study objectives and design, including selection of sites and recruitment of clinicians and patients, has been extensively reported[1,4,5,10,20,21] and is briefly summarized here. The MOS was conducted in three cities (Boston, Mass; Chicago, Ill; and Los Angeles, Calif) selected from three of four census regions. In each city, one large health maintenance organization (HMO), numerous multispecialty groups, and representative solo practices were studied. From these systems of care, physicians board certified or board eligible in family practice, internal medicine, cardiology, endocrinology, and psychiatry were identified along with clinical psychologists, clinical social workers, and other mental health providers. Solo and small-group clinicians were identified from master files of the American Medical Association, American Academy of Family Physicians, and American Psychological Association. Multispecialty group clinicians were identified from the Medical Group Management Association membership directory, and HMO clinicians were identified by upper-level management.

The process of enrolling clinicians differed by system of care. Of eligible clinicians practicing in HMOs or large multispecialty groups, 225 (79%) agreed to participate in the MOS.[10] Solo and small-group clinicians were selected by a multistage sampling process. This process yielded 298 solo or small-group practitioners (58% of those eligible and who agreed to be contacted).[10] Physician participants were similar to nonparticipants regarding clinical training and sociodemographic and practice characteristics; participants tended to be more involved in direct patient care than nonparticipants.[1]

Study participants were English-speaking adults (18 years of age and older) who had an office visit with an enrolled clinician during 9-day screening periods in February to November, 1986. Patients seen during this period were asked to complete a brief, standardized, self-report questionnaire that gathered information about chronic disease, depressive symptoms, sociodemographic characteristics, and general health status. Complete questionnaires were obtained from 74% of eligible patients treated in group practices and from 65% of patients treated in solo or small-group practices (N = 22,462). For 96% of these patients, their clinicians also completed a brief, standardized questionnaire that elicited information on diagnosis, disease severity, and visit content.

Data from the physician-completed questionnaires were used to identify patients with the four MOS medical tracer conditions (hypertension, diabetes, congestive heart failure (CHF), and recent myocardial infarction(MI)).[1,4,10] Patients with these conditions were identified on the basis of a standardized physician report form. A two-stage process, involving a depressive symptom scale[22] included in the patient-completed questionnaire and the National Institute of Mental Health's Diagnostic Interview Schedule (DIS), was used to identify patients with depression and to stage their severity.[5,21]

Patients with matched patient and physician questionnaires who were determined to have one of the medical tracer conditions or current depressive symptoms were subsequently contacted for a telephone interview. This interview was designed to: 1) determine the presence of psychiatric disorder among those with current symptoms by using the depression section of the DIS;[5,21] and 2) to enroll patients who met DIS criteria for psychiatric disorders and patients who met original diagnostic criteria for the medical tracers. Of those eligible for enrollment and

who were successfully contacted by telephone, 73% (N = 5,341) completed the interview and 91% (N = 4,824) of interviewed patients agreed to enroll in the study. Upon enrollment, patients were invited to the MOS Health Examination and were sent the baseline Patient Assessment Questionnaire. The health examination (standardized medical history and clinical examination) was independently conducted by specially trained MOS medical staff. Health examinations were completed on 2,583 patients and 3,445 patients returned the baseline questionnaire.

The MOS patient sample used for the psychometric analyses included all enrolled patients who completed the 245-item baseline questionnaire, which included the 36 items which were later used to construct the SF-36 survey (N = 3,445). Because disease-specific information from the health examination was used to stage severity for clinical tests of validity reported here, we limited that sample to a subset of enrolled patients who completed *both* the baseline questionnaire and health examination within a 1-month period (N = 1,014). We required the completion of the baseline questionnaire and the health examination to be within a 1-month period so that the clinical criteria and the health scales they are compared with were measured in close proximity. The sample analyzed here for clinical tests of validity is similar to that used in previously reported comparisons of the relative precision of single-item and MOS short- and long-form general health status measures.[23] In this article, we add a fourth group of patients—those who have *both* chronic medical and psychiatric conditions. We also add clinical tests of validity using the *severity* of psychiatric disorders as additional clinical criteria.

Tests of Validity Using Psychometric Criteria

Previous studies investigating the dimensionality of self-reports of health have confirmed distinct physical and mental health components.[18,19,24–27] To test for these dimensions of health within the SF-36, we extracted principal components from the correlations among its eight scales.[28] Correlations between the scales and the first unrotated component test for the large general health factor hypothesized to be common to all eight scales. The pattern of correlations between the eight scales and the two rotated components test the validity of each scale in relation to hypothesized physical and mental health dimensions.

Tests of Validity Using Clinical Criteria

We also assessed the validity of each scale by comparing patient groups differing in physical and/or mental health status and severity. Using clinical criteria, four mutually exclusive groups were formed: Group 1, minor (uncomplicated) chronic medical conditions only (N = 638); Group 2, serious (complicated) chronic medical conditions only (N = 168); Group 3, psychiatric conditions only (N = 163); and Group 4, both serious medical and psychiatric conditions (N = 45). The first three groups are identical to those studied elsewhere.[23] We document here more thoroughly the clinical criteria used to define each group.

To distinguish patients differing in severity of chronic medical condition, we used disease-specific severity scales constructed from the standardized medical history interview.[29,30] Patients classified as having a serious chronic medical condition (Groups 2 and 4) included the following: 1) CHF patients reporting edema, orthopnea, or dyspnea on exertion (5% of CHF patients); 2) MI survivors with noteworthy and recurring angina symptoms and/or severe CHF symptomology (2% of MI patients); and 3) hypertension patients with reports of severe CHF symptomology and/or history of a stroke (2% of hypertension patients). Twelve percent of diabetic patients were classified as severe because of the presence of at least one of the following complications: history of an MI; weekly angina; se-

vere autonomic neuropathy; moderately severe peripheral neuropathy *and* lack of blood sugar control or severe vision problems or moderately severe autonomic neuropathy; or recurring angina monthly *and* lack of blood sugar control or severe vision problems or severe peripheral neuropathy or moderately severe autonomic neuropathy.

We defined psychiatric conditions using well-established psychiatric diagnostic criteria, as reported in detail elsewhere.[5,21,22] Briefly, patients were determined to have current depressive symptoms based on responses to an eight-item depression symptom scale[22] administered during the screening visit. The subsequent DIS telephone interview (described earlier) was used to classify them as having current unipolar affective disorder (major depression or dysthymia) *or* serious depressive symptoms in the absence of a disorder. Patients with either depressive disorders or current symptoms were included in Groups 3 and 4. To test validity in relation to severity of psychiatric condition, we disaggregated Group 3 and compared patients with current unipolar affective disorder to those with serious depressive symptoms in the absence of a disorder.

Hypotheses

The first panel of Table 1 presents hypotheses regarding the factor content of each SF-36 scale along with results of tests of those hypotheses, which are presented below. We define a strong association as a correlation greater than 0.70, moderate to substantial as a correlation of 0.30 to 0.70, and weak as a correlation less than 0.30. These are equivalent, in variance terms, to shared variances of > 50%, 10% to 50%, and < 10%.

On the basis of previous research,[18,19,26] we expected SF-36 scales measuring physical functioning, role limitations due to physical health problems, and bodily pain 1) to be most highly correlated with an empirically derived physical health component; 2) to be most valid in distinguishing groups differing in severity of chronic medical condition; 3) to show little or no association with the mental health component; and 4) to perform less well than the mental health scales in distinguishing groups differing in the presence

TABLE 1. Hypothesized Associations Between SF-36 Scales and Results From Psychometric Tests

	Hypothesized Association		Rotated Principal Components			Relative Validity[b]	
	Physical	Mental	Physical[a]	Mental[a]	h^2	Physical	Mental
Physical functioning	+	−	0.88	0.04	0.78	1.00	0.00
Role—physical	+	−	0.78	0.30	0.70	0.79	0.11
Bodily pain	+	−	0.77	0.24	0.65	0.77	0.07
Mental health	−	+	0.12	0.90	0.82	0.02	1.00
Role—emotional	−	+	0.19	0.81	0.69	0.05	0.81
Social functioning	*	+	0.44	0.71	0.70	0.25	0.62
Vitality	*	*	0.59	0.57	0.67	0.45	0.40
General health perceptions	*	*	0.68	0.32	0.56	0.60	0.13

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

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

[b] Computed by the ratio of the common-factor variance of each scale relative to the scale with the greatest common-factor variance. The common-factor variance of each scale is the square of each scale-component correlation.

+ Strong Association ($r \geq 0.70$)

* Moderate to Substantial Association ($0.30 < r < 0.70$)

− Weak Association ($r \leq 0.30$)

and severity of psychiatric disorders. Similarly, on the basis of previous research,[18,19,26] we expected SF-36 measures of general mental health and role limitations due to emotional problems 1) to be most highly correlated with the mental health component; 2) to be most valid in distinguishing groups differing in the presence and severity of psychiatric disorders; 3) to show little or no association with the physical health component; and 4) to perform less well than the physical health scales in distinguishing patients differing in severity of chronic medical condition.

On the basis of their content, we expected some scales to measure both physical and mental health factors and, thus, to be valid for purposes of comparing groups differing in both physical and mental health status as clinically defined. First, on the basis of previous research,[18,19,26] we expected the vitality and general health perceptions scales to be moderately correlated with both physical and mental health components and to distinguish groups differing in both physical and mental health status. Second, although we expected the social functioning scale to be highly correlated with the mental health component,[18] we also expected a moderate correlation with the physical health component. Because the social functioning items confound physical and mental health by design, that scale should be sensitive to the burden of both physical and mental health as clinically defined.

Methods of Analysis

The general methodology used for assessing the relative validity (RV) of the eight SF-36 scales as measures of physical and mental health constructs has its roots in the concept of statistical efficiency.[31,32] Briefly, a measure is more efficient, relative to another, if it yields the right information with greater accuracy (less error). Liang et al.[33] applied the concept of statistical efficiency in health status assessment by comparing the relative efficiency of five health status instruments (using the ratio of squared *t*-statistics) in detecting change in functioning over time. We improved on this methodology in tests of the relative precision of short- and long-form health status scales by holding sample size constant within comparisons, holding groups constant across comparisons, and defining clinical groups to differ in clearly interpretable ways.[23]

We extend here the methodology to test the RV of the eight SF-36 scales as indicators of two unobservable health *constructs*. For clinical tests of validity, we used unadjusted general linear models to estimate mean differences between pairs of clinical groups for each of the eight scales. The resulting *F*-statistic for each scale defines the ratio of between-groups (systematic) variance relative to within-group (error) variance. The greater the *F*-ratio, the greater the amount of information (systematic variance) a scale provides about the criterion relative to error variance. Sample size was held constant across scales to standardize comparisons. By analyzing identical samples across scales for each clinical contrast, the relative size of *F*-ratios reflects the relevance of the scales to a particular criterion. We estimated RV for the eight scales for each clinical-group contrast by computing the ratio of pair-wise *F*-statistics (*F* for each comparison scale divided by *F* for the most valid scale). The resulting RV estimates indicate in proportional terms how much less valid each scale is as a measure of physical or mental health status, relative to the most valid scale.

We used principal components analysis to test the hypothesized dimensionality of the SF-36 scales. Because we hypothesized two dimensions to underlie the structure of the eight scales, we extracted two principal components. The size of the first unrotated component and the pattern of correlations between it and the eight scales gauge the extent to which the scales contribute to a common general health dimension. To facilitate interpretation, we rotated the compo-

nents to orthogonal simple structure using the varimax method. To interpret the components, we examined the pattern of correlations across the eight scales. To evaluate the validity of each of the eight scales, we compared their correlations with the hypothesized component(s) (convergent validity) versus the other component (discriminant validity).

To evaluate the factorial validity of each scale as a measure of each component, we first squared each factor loading (scale-component correlation) to estimate the proportion of variance shared with that component (common-factor variance). We defined the scale sharing the most variance with each component as the most valid measure of that component. For each component, we then estimated RV for each scale by dividing the variance shared with the component by that estimate for the most valid scale. These ratios indicate in proportional terms how much less valid each scale is relative to the most valid scale. The higher the RV of a scale, the more precisely or efficiently it measures the underlying construct of interest as defined by the most valid scale.

Results

Validation of Clinical Groups Compared

As Table 2 shows, clinical criteria produced the desired mutually exclusive groups differing in the severity of medical and psy-

TABLE 2. Characteristics of Patients in Four Clinical Groups

Patient Characteristics	Minor Medical Conditions[a] (N = 638) 1	Serious Medical Conditions[b] (N = 168) 2	Psychiatric Condition Only[c] (N = 163) 3	Psychiatric and Serious Medical Conditions[d] (N = 45) 4
Sociodemographics				
Mean age (SD)	57.4 (12.8)	61.0 (12.4)	41.8 (12.6)	54.4 (12.5)
% female	47.0	49.7	73.0	68.9
Medical and psychiatric conditions				
% Complicated advanced coronary artery disease	0.0	35.1	0.0	17.8
% Complicated hypertension	0.0	20.8	0.0	28.9
% Complicated diabetes	0.0	61.3	0.0	62.2
% Current depressive symptoms	0.0	0.0	100.0	100.0
% Current depressive disorder	0.0	0.0	63.8	22.2
Health status				
% Self-rated health fair or poor	17.4	43.8	21.6	74.4
% Any bed days last 3 months	8.7	15.8	35.2	46.7
Provider specialty				
Medical subspecialist	13.9	25.6	3.7	24.4
Mental health professional	0.0	0.0	42.3	0.0
Utilization of health care services				
% Clinician visit within past two weeks	29.1	40.0	45.3	51.3
% Hospitalized past 12 months	12.3	25.2	20.8	47.4
% Ever utilized mental health services	21.2	22.4	82.5	56.8

[a] Minor medical: patients with uncomplicated chronic medical conditions.
[b] Serious medical: patients with advanced or complicated chronic medical conditions.
[c] Psychiatric only: patients with either current depressive symptoms or disorder but no chronic medical condition.
[d] Psychiatric and serious medical: patients with either current depressive symptoms or disorder and a serious chronic medical condition.

chiatric conditions. None of the patients assigned to Groups 1 and 3 had complicated medical conditions, whereas patients in Groups 2 and 4 all had complicated medical conditions. As intended, none of the patients in Groups 1 and 2 had current depressive symptoms or disorders. All patients in Groups 3 and 4 had current depressive symptoms, and 64% of Group 3 and 22% of Group 4 patients had a current depressive disorder.

Demographic differences among the groups correspond well with epidemiologic trends in the United States[34] (Group 3 patients were the youngest and disproportionately female and Group 2 patients were the oldest). The substantial differences in personal ratings of health, proportion reporting any bed days in the last 3 months, and utilization of health services across groups provide further evidence of the desired distinctions between the groups in health status as clinically defined. For example, 74% of patients with both serious medical and psychiatric conditions reported their health as fair or poor, compared with 44% of serious medical patients and 22% or less of patients with solely psychiatric or minor medical conditions. Report of any bed days in the last 3 months was also greatest among patients with both serious medical and psychiatric conditions. Patients with minor medical conditions were the least likely to have recently used health care services, while patients with psychiatric conditions were the most likely to have ever consulted a mental health professional. In summary, these data provide prima facie evidence that the intended differences in the presence and severity of medical and psychiatric conditions were achieved across the comparison groups.

Psychometric Validity

The components analysis confirmed the substantial general health dimension hypothesized to be common to all eight scales. The first principal component accounted for 55% of the total measured variance and correlated highly with all eight scales (range = 0.67 for role-emotional to 0.82 for vitality, median = 0.74). Extraction of the second component increased the percentage of total variance explained from 55% to 70%. Communalities (h^2 in Table 1) indicate the extent of overlap in terms of common variance between each measure and the two extracted factors. The percentage of total variance in each scale accounted for by the two-factor solution ranged from 0.56 to 0.82 across scales, indicating that the two factors accounted for the majority of the reliable variance in each scale.

The middle panel of Table 1 presents correlations between the SF-36 scales and the two rotated components. Rotation of these components confirmed the hypothesized physical and mental dimensions of health. As hypothesized for a physical health component, the physical functioning, role-physical, and bodily pain scales correlated most highly with the first rotated component, while the mental health and role-emotional scales correlated weakly. As hypothesized for a mental health component, the order of correlations with the eight scales was nearly reversed for the second component. Specifically, the mental health, role-emotional, and social functioning scales correlated most highly with the second component, while physical functioning, bodily pain, and role-physical scales correlated weakly. Based on these patterns of correlations, we interpreted the first and second components as "physical" and "mental" health dimensions, respectively.

The third panel of Table 1 presents estimates of the RV of the eight scales as measures of physical and mental health components. Because the physical functioning and mental health scales had the highest correlations, respectively, with the physical and mental health components, they served as the standards for estimating RV. As hypothesized, the role-physical and bodily pain

scales showed strong associations, in terms of shared common-factor variance, with the physical health component (RV = 79% and 77%, respectively). The mental health component was best measured by the mental health scale, followed by the role-emotional and social functioning scales (RV = 0.81 and 0.62, respectively). The three scales hypothesized to measure more than one health dimension (social functioning, vitality, and general health perceptions) showed moderate to strong associations with both components. However, for these three scales, there was substantial variation in observed RV estimates: the general health perceptions scale was clearly more strongly associated with the physical than mental component; the social functioning scale was more highly associated with the mental than physical component; and the vitality scale showed nearly equal associations with both components.

Clinical Validity

Tables 3, 4, and 5 present results from tests of validity based on comparisons among the four clinical groups. These comparisons test the validity of the scales in detecting decrements in health status associated with chronic medical and/or psychiatric conditions. Table 3 presents means and standard errors for each group across the eight SF-36 scales.

Table 4 presents pair-wise mean differences, pair-wise *F*-statistics, and estimates of RV for group comparisons involving minor medical patients. Patients with serious medical conditions scored significantly lower on all eight scales compared to patients with minor medical conditions (Group 2 vs. 1). However, as indicated by the wide range of observed RV estimates, all scales were not equally valid in this clinical-group comparison. As hypothesized, the physical functioning scale was most valid in detecting differences between patients with minor versus serious medical conditions. The general health perceptions scale nearly equaled that standard (RV = 0.99), followed by the role-physical and vitality scales (RV = 0.71 and 0.67, respectively). As hypothesized, the best mental health scales (mental health and

TABLE 3. Means (and Standard Errors) for Groups Differing in Medical and Psychiatric Conditions

	Comparison Groups			
	Group 1	Group 2	Group 3	Group 4
Scale	Minor Medical N = 576	Serious Medical N = 144	Psychiatric Only N = 153	Psychiatric & Serious Medical N = 43
Physical functioning	80.53	57.35	80.62	46.37
	(0.89)	(2.34)	(1.64)	(4.24)
Role-physical	70.27	43.92	55.56	23.84
	(1.48)	(3.31)	(3.18)	(4.63)
Bodily pain	76.06	65.10	63.30	50.23
	(0.91)	(2.06)	(1.91)	(3.52)
Mental health	82.49	77.59	52.75	56.90
	(0.59)	(1.32)	(1.63)	(3.08)
Role-emotional	84.26	76.16	40.74	52.71
	(1.27)	(3.11)	(3.20)	(5.89)
Social functioning	91.62	80.03	64.54	65.12
	(0.62)	(2.03)	(2.06)	(3.44)
Vitality	62.02	47.79	45.32	37.05
	(0.82)	(1.82)	(1.65)	(3.11)
General health perceptions	67.02	49.13	57.91	39.93
	(0.74)	(1.80)	(1.75)	(2.30)

TABLE 4. Summary of Clinical Validity Tests Involving Minor Medical Patients

Scale	Group 2 vs. 1 Serious Medical vs. Minor Medical			Group 3 vs. 1 Psychiatric vs. Minor Medical			Group 4 vs. 1 Both Serious Medical and Psychiatric vs. Minor Medical		
	Mean Difference	F	Relative Validity	Mean Difference	F	Relative Validity	Mean Difference	F	Relative Validity
Physical functioning	−23.18[a]	85.9	1.00	0.09	0.0	0.00	−34.16[a]	62.2	0.66
Role-physical	−26.35[a]	60.6	0.71	−14.71[a]	19.9	0.07	−46.43[a]	69.9	0.74
Bodily pain	−10.96[a]	23.6	0.27	−12.76[a]	39.9	0.14	−25.83[a]	55.6	0.59
Mental health	−4.90[a]	13.3	0.15	−29.74[a]	294.7	1.00	−25.59[a]	66.7	0.71
Role-emotional	−8.10[b]	5.8	0.07	−43.52[a]	159.9	0.54	−31.55[a]	27.4	0.29
Social functioning	−11.59[a]	29.9	0.35	−27.08[a]	158.6	0.54	−26.50[a]	57.4	0.61
Vitality	−14.23[a]	57.8	0.67	−16.70[a]	86.0	0.29	−24.97[a]	64.4	0.68
General health perceptions	−17.89[a]	84.7	0.99	−9.11[a]	22.9	0.08	−27.09[a]	94.4	1.00

[a] $P < 0.001$.
[b] $P < 0.01$.

role-emotional) performed most poorly in this test. The bodily pain scale performed less well than hypothesized (RV = 0.27).

As hypothesized, for clinical comparisons involving the presence or absence of a psychiatric condition (Group 3 vs. 1), the mental health scale proved to be the most valid, followed by the role-emotional and social functioning scales (RV = 0.54 each). Also as hypothesized, the physical functioning scale did not distinguish between groups differing only in psychiatric condition (RV = 0.00), and the role-physical and bodily pain scales were less valid measures for this group contrast. The general health perceptions scale also yielded poor validity relative to the standard in this test (RV = 0.08).

Patients with both serious medical and psychiatric conditions scored significantly lower than minor medical patients in all eight scales (Group 4 vs. 1). The general health perceptions scale was most valid in detecting the combined effects of medical and psychiatric conditions. The other scales performed similarly in this test (RV range = 0.59–0.74), with the exception of the role-emotional scale (RV = 0.29).

Table 5 extends tests of validity to groups of patients with serious medical and psychiatric conditions. The mental health scale was most valid in detecting the incremental burden of a psychiatric condition among patients with serious medical conditions (Group 4 vs. 2). The other seven scales were well below that standard (RV range = 0.13 to 0.34, median = 0.32). The physical functioning scale was most valid in detecting the incremental burden of a serious medical condition among patients with a psychiatric condition (Group 4 vs. 3), followed by the general health perceptions and role-physical scales (RV = 0.68 and 0.56, respectively). The remaining five scales performed relatively poorly in this test (RV range = 0.00 to 0.18).

The mental health scale was most valid in distinguishing serious medical from psychiatric patients (Group 3 vs. 2). Although the physical functioning and role-emotional scales had similar RV estimates (RV = 0.47 and 0.45, respectively), their group mean differences were in opposite directions, as would be expected. Specifically, patients with psychiatric conditions had better physical functioning but worse role-emotional functioning than patients with serious medical conditions. Scales measuring social functioning, general health perceptions, and role-physical showed significant differences between the groups but were far less valid.

TABLE 5. Summary of Clinical Validity Tests Involving Chronically Ill Patients

Scale	Group 4 vs. 2 Psychiatric Incremental: Psychiatric Among Serious Medical			Group 4 vs. 3 Medical Incremental: Serious Medical Among Psychiatric			Group 3 vs. 2 Psychiatric vs. Serious Medical		
	Mean Difference	F	Relative Validity	Mean Difference	F	Relative Validity	Mean Difference	F	Relative Validity
Physical functioning	−10.98[c]	5.1	0.13	−34.25[a]	56.8	1.00	23.27[a]	66.4	0.47
Role-physical	−20.08[a]	12.5	0.33	−31.72[a]	31.9	0.56	11.64[c]	6.4	0.05
Bodily pain	−14.87[a]	12.3	0.32	−13.07[b]	10.4	0.18	−1.80	0.4	0.00
Mental health	−20.69[a]	38.1	1.00	4.15	1.4	0.02	−24.84[a]	140.2	1.00
Role-emotional	−23.45[a]	12.9	0.34	11.97	3.1	0.05	−35.42[a]	62.7	0.45
Social functioning	−14.91[a]	12.9	0.34	0.58	0.0	0.00	−15.49[a]	28.7	0.20
Vitality	−10.74[b]	8.3	0.22	−8.27[c]	5.5	0.10	−2.47	1.0	0.01
General health perceptions	−9.20[b]	9.9	0.26	−17.98[a]	38.6	0.68	8.78[a]	12.3	0.09

[a] $P < 0.001$.
[b] $P < 0.01$.
[c] $P < 0.05$.

The vitality and bodily pain scales did not distinguish these two groups.

Table 6 presents results for tests of validity in relation to the *severity* of psychiatric disorder for patients within Group 3—symptomatic depression versus more severe clinical depression. As hypothesized, the mental health scale was most valid in detecting these differences, followed by scales measuring role-emotional (RV = 0.43), social functioning (RV = 0.32), and vitality (RV = 0.31). The best physical health measures (physical functioning, role-physical, bodily pain, and general health perceptions) all had RV estimates close to 0.

Summary of Results

Table 7 presents hypotheses for each scale and summarizes RV estimates obtained

TABLE 6. Summary of Clinical Validity Results for Groups Differing in Severity of Psychiatric Condition

Scale	Symptomatic Depression N = 56	Clinical Depression N = 97	Mean Difference	F	Relative Validity
Physical functioning	81.20 (2.92)	80.28 (1.97)	−0.92	0.07	0.00
Role-physical	62.95 (5.21)	51.29 (3.97)	−11.66	3.16	0.06
Bodily pain	64.71 (3.25)	62.48 (2.37)	−2.23	0.31	0.01
Mental health	65.19 (2.00)	45.56 (1.96)	−19.63[a]	49.03[a]	1.00
Role-emotional	58.93 (5.50)	30.24 (3.53)	−28.69[a]	21.10[a]	0.43
Social functioning	74.78 (3.10)	58.63 (2.53)	−16.15[a]	15.64[a]	0.32
Vitality	53.39 (2.47)	40.65 (2.05)	−12.74[a]	15.06[a]	0.31
General health perceptions	59.95 (2.78)	56.74 (2.25)	−3.21	0.78	0.02

[a] $P < 0.001$.

TABLE 7. Summary of Results for Psychometric and Clinical Tests of Relative Validity

Scale	Physical Health				Mental Health				
	Hypotheses	Psychometric	Medical Severity	Medical Incremental: Serious Medical Among Psychiatric	Hypotheses	Psychometric	Psychiatric Disorder	Psychiatric Severity	Psychiatric Incremental: Psychiatric Among Serious Medical
Physical functioning	+	1.00	1.00	1.00	−	0.00	0.00	0.00	0.13
Role-physical	+	0.79	0.71	0.56	−	0.11	0.07	0.06	0.33
Bodily pain	+	0.77	0.27	0.18	−	0.07	0.14	0.01	0.32
Mental health	−	0.02	0.15	0.02	+	1.00	1.00	1.00	1.00
Role-emotional	−	0.05	0.07	0.05	+	0.81	0.54	0.43	0.34
Social functioning	*	0.25	0.35	0.00	+	0.62	0.54	0.32	0.34
Vitality	*	0.45	0.67	0.10	*	0.40	0.29	0.31	0.22
General health perceptions	*	0.60	0.99	0.68	*	0.13	0.08	0.02	0.26

+, Strong Association (RV ≥ 0.50).
*, Moderate to Substantial Association (0.10 < RV < 0.50).
−, Weak Association (RV ≤ .10).

from psychometric tests and five clinical tests. These clinical tests were judged to be most useful because they tested convergent and discriminant validity in relation to unconfounded differences in physical or mental health as clinically defined. This table can be interpreted both row-wise and column-wise. Each column summarizes results across scales for a particular validity criterion. Table entries for a given column (criterion) are RV estimates, which indicate how much less valid a scale is relative to the best scale. These results serve as guidelines for hypothesizing which scale/concept is most relevant to each criterion. Summaries of results by row indicate whether the interpretation of each scale is pure or complex; that is, whether observed differences are largely due to one health component or likely due to both components. These results serve as guidelines for interpreting each scale.

As summarized in Table 7, the scales identified in the components analysis to best represent the physical and mental health dimensions— physical functioning and mental health—were most valid, respectively, in clinical tests involving detection of the burden of severe medical versus psychiatric conditions. Further, the mental health scale best distinguished between patients within the psychiatric group who differed only in the severity of their disorder. These findings support the convergent validity of the physical functioning and mental health scales. Consistent with results from psychometric tests, the physical functioning scale was least valid in tests involving the presence and severity of psychiatric conditions and the mental health scale was least valid in the medical severity test. These findings support the discriminant validity of these two scales.

The incremental burden tests summarized in Table 7 provide further evidence for the convergent and discriminant validity of these two scales. The physical functioning scale was most valid, and the mental health scale least valid, in detecting the incremental burden of serious medical conditions among

those with a psychiatric condition. The mental health scale was most valid, and physical functioning least valid, in detecting the incremental burden of psychiatric conditions among those with serious medical conditions.

The role-physical and role-emotional scales showed strong convergent and discriminant validity in relation to role disabilities associated with medical versus psychiatric disorders. In both psychometric and clinical tests, each role functioning scale was strongly related to one component (physical or mental) and unrelated to the other component. For both physical and mental health dimensions, the social functioning scale showed moderate to strong convergent validity across psychometric and clinical tests but fairly poor discriminant validity. As hypothesized, the vitality scale showed good convergent validity for physical and mental health effects in both psychometric and clinical tests, but it has poor discriminant validity.

Two exceptions in expected results from psychometric and clinical validity tests are apparent in Table 7. First, the bodily pain scale showed strong convergent validity in the physical-health factorial test as hypothesized, but poor convergent validity in both medical-severity clinical tests. Second, we hypothesized moderate convergent validity for the general health perceptions scale in relation to both physical and mental components of health. However, for both psychometric and clinical criteria, it performed relatively better than hypothesized in physical health tests and relatively worse in mental health tests.

Discussion

Well-accepted definitional standards[35–39] and empirical work to date[18,19,24–27] have identified physical and mental components of health status. The SF-36 survey was constructed to provide a comprehensive assessment of each of these dimensions.[13] We used psychometric and clinical standards to assess the validity of each SF-36 scale as a measure of the physical or mental dimension of health status. Overall, results from the psychometric and clinical tests of validity agreed with one another and converged with study hypotheses. Thus, there is a good basis for establishing guidelines for the interpretation of score differences for each scale as a measure of physical and/or mental health effects and also specifying the size of differences in each scale score that should be considered large.

Our results indicate that the physical functioning and mental health scales are relatively pure and, therefore, their interpretation is unequivocal. These two scales, respectively, measure the physical and mental dimensions of health and are most sensitive, respectively, to the clinical manifestations of medical and psychiatric conditions. Therefore, when observed differences are found on these scales, interpretation attributed to physical *or* mental causes can be made with a high degree of confidence. Unambiguous interpretations of these scores were generalizable both within and across various combinations of the medical and psychiatric conditions studied here. This information is important because little is known about the validity of health status measures in patients with both medical and psychiatric conditions.[27]

However, a comprehensive assessment of health requires representation of more than physical and mental functioning as defined by these two scales. To be comprehensive, an assessment should provide information on limitations in engaging in normative roles as a result of health problems. To capture aspects of disability, role and social functioning scales were included in the SF-36 survey. Observed differences on the role-physical scale can be interpreted as role disability associated largely, but not entirely, with physical health effects. Interpretation of scores may be complicated somewhat when psychiatric conditions are present (see

incremental test of physical health). Differences in role-emotional scores can be interpreted with confidence as role disability associated with mental health problems. By design, the social functioning scale confounds physical and mental health attributions. Accordingly, while the social functioning scale appears most sensitive to social disability associated with mental health problems, it is moderately sensitive to the burden of physical health problems as well. Interpretation of social functioning scores is, therefore, complex and observed differences can not be confidently attributed to *either* physical or mental health problems.

The vitality scale is a subjective measure of general well-being. By design, it was intended to tap both positive health states (e.g., energy) as well as somatic expressions of physical illness and psychological distress (e.g., fatigue). As a result, the interpretation of vitality scores was expected to be complicated relative to both physical and mental health dimensions, and this was confirmed empirically in both psychometric and clinical tests of validity.

The strong convergent validity of the bodily pain scale in the psychometric test, yet poor convergent validity in medical tests, may be an artifact of the specific conditions that were represented in the severe medical group. The four medical conditions represented are not typically dominated by pain. Consistent with this explanation, previous studies have shown that the SF-36 severity of bodily pain item was the most valid measure in group discriminations involving patients with arthritis and back problems.[4] This issue warrants further study. Given the weak to low-moderate associations between the bodily pain scale and both psychometric and clinical criteria for mental health, our results suggest that differences in this scale can be attributed largely to the physical dimension of health.

The relatively poor convergent validity results for the general health perceptions scale in both psychometric and clinical tests of the mental health component suggest that this scale is most sensitive to the physical health dimension. Further, RV estimates for the general health perceptions scale tended to be higher in clinical than in psychometric tests of physical health. These differences in results across psychometric and clinical tests suggest this scale taps aspects of physical health including but not limited to those represented in the physical functioning scale. Consistent with this finding, previous research has found measures of general health perceptions to be highly sensitive to both serious and minor physical symptoms, regardless of whether they are associated with physical limitations or with disability.[40]

Although the results of psychometric and clinical tests were not identical, taken as a whole they were very similar and provide a basis for guidelines for interpreting each scale. Both psychometric and clinical tests provided consistent information about the underlying nature of each scale—physical and/or mental—as well as the degree to which each scale measured that component (pure versus complex). We achieved a greater understanding of the validity of score inferences, and the quality of those inferences, by combining distinct approaches to construct validation—assessment of convergent and discriminant validity across psychometric and clinical standards. These results underscore the usefulness of combining psychometric with clinical tests to better understand the interpretation of measures.

An important lesson of this research is that a multidimensional assessment of health is necessary to achieve a comprehensive understanding of the impact of disease on health-related quality of life. Relatively pure measures, such as the physical functioning and mental health scales, are highly sensitive to the psychometric and clinical criteria studied here and permit unambiguous interpretations. However, sole use of these measures results in an incomplete assess-

ment of health because they ignore variations in disability, personal evaluations of health, and general well-being. Therefore, despite the complexity of interpretation inherent in measures of role and social disability, vitality, and perceptions of health, they are essential qualities to measure to obtain a synergistic and comprehensive assessment of the burden of disease and/or treatment on patients' everyday functioning and well-being.

Further, multidimensional assessments of health are important because, unlike the groups deliberately formed for validity tests here, most patients have multiple coexisting conditions, both physical and mental. For example, medical comorbidity is common among patients with both chronic medical[4,41] and psychiatric[42] conditions, and psychiatric comorbidity is common among patients with medical conditions.[43,44] Moreover, given the extent of under-recognition of depressive disorders in primary care,[21,43,45] the prevalence of comorbid medical and psychiatric conditions may be greater than previously reported. Results from the incremental burden tests indicate that scales that measure both physical and mental dimensions may be most useful in these circumstances. For example, the general health perceptions scale was most valid in detecting the combined effect of having both a serious medical and psychiatric condition relative to uncomplicated patients. Analysis of a unidimensional measure will not capture the range of effects disease and/or treatment have on subjective states that have social meaning for the patient and possibly clinical significance for the practitioner.

One barrier to the meaningful use of general health status measures in clinical practice and research is the lack of information necessary to interpret scores.[46,47] Our results not only provide guidelines for interpreting score differences in each scale but also provide guidelines for establishing the size of large score differences. Because clinically severe groups were compared, results reported here help to gauge the size of differences in scores that should be considered very large. These estimates apply only to the MOS SF-36 scoring algorithms, which are documented elsewhere.[48] For example, a difference of 23 points on the physical functioning scale (nearly one standard deviation) reflects the impact of a complicated chronic medical condition on everyday physical functioning. A difference of 27 points on the mental health scale (1.3 standard deviation units) reflects the impact of serious depressive symptoms. Pending further research, the mean differences reported here are offered as benchmarks for gauging very large effect sizes for the SF-36 scales. While these differences might appear to be so large as to render measurement meaningless, physicians greatly underestimate patient-reported disabilities in physical and social functioning,[49,50] and mental health differences of this magnitude are routinely underdetected in primary care.[21,43,45]

Results from tests of validity based on comparisons between groups known to differ clinically have great potential in documenting the sizes of small and large differences in general health scales as well as in advancing understanding of the meaning of those differences. Such tests should be extended to include more subtle disease-specific criteria to define the sizes of very small score differences and tests of the convergent and discriminant validity of scales in detecting those differences. Tests based on small and large clinical changes over time will also advance understanding of how to use and interpret general health scales. The results reported here clearly indicate that the issue is not as simple as whether or not a health status scale is valid. At least for the SF-36 scales, validity for purposes of measuring one dimension of health tends to go hand in hand with poor validity for another. Thus, in selecting measures of health status, prior-

ity should be given to those proven to be most relevant to the desired use and interpretation.

Acknowledgments

The authors gratefully acknowledge the following: Audrey Burnam, PhD, Sheldon Greenfield, MD, Richard Kravitz, MD, Alvin R. Tarlov, MD, Kenneth Wells, MD, and Mark B. Wenneker, MD for assistance in defining the medical and psychiatric groups compared; Cameron Cushing, MS, Stephanie Kieszak, MA, and J. F. Rachel Lu, MS, for analytic assistance; helpful critiques provided by two anonymous reviewers, James D. Lankin for editorial assistance; and Rebecca Voris, Jennifer Lin, and Kathleen Clark for administrative support.

References

1. Tarlov AR, Ware JE, Greenfield S, et al. The Medical Outcomes Study: an application of methods for monitoring the results of medical care. JAMA 1989;262:925.

2. Stewart AL, Hays RD, Ware JE. The MOS Short-Form General Health Survey: reliability and validity in a patient population. Med Care 1988;26:724.

3. Ware JE, Sherbourne CD, Davies AR. Developing and testing the MOS 20-item Short-Form Health Survey: a general population application. In: Stewart AL, Ware JE, eds. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Chapel Hill, NC: Duke University Press, 1992.

4. Stewart AL, Greenfield S, Hays RD, et al. Functional status and well-being of patients with chronic conditions: results from the Medical Outcomes Study. JAMA 1989;262:907.

5. Wells KB, Stewart A, Hays RD, et al. The functioning and well-being of depressed patients: results from the Medical Outcomes Study. JAMA 1989; 262:914.

6. Bindman AB, Keane D, Lurie N. Measuring health changes among severely ill patients: the floor phenomenon. Med Care 1990;28:1142.

7. Katon WJ, Buchwald DS, Simon GE, et al. Psychiatric illness in patients with chronic fatigue and those with rheumatoid arthritis. J Gen Intern Med 1991; 6:277.

8. Wu AW, Rubin HR, Mathews WC, et al. A health status questionnaire using 30 items from the Medical Outcomes Study. Med Care 1991;29:786.

9. Wachtel T, Piette J, Mor V, et al. Quality of life in persons with AIDS as measured by the Medical Outcomes Study instruments. Ann Intern Med 1992; 116:129.

10. Kravitz RL, Greenfield S, Rogers W, et al. Differences in the mix of patients among medical specialties and systems of care: results from the Medical Outcomes Study. JAMA 1992;267:1617.

11. Parkerson GR, Broadhead WE, Tse, CJ. Comparison of the Duke Health Profile and the MOS Short-Form in health young adults. Med Care 1991;29:679.

12. Anderson JS, Sullivan V, Usherwood TP. The Medical Outcomes Study Instrument (MOSI)—Use of a new health status measure in Britain. J Fam Practice 1990;7:205.

13. Ware JE, Sherbourne CD. The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual framework and item selection. Med Care 1992;30:473.

14. Garrett HE. Statistics in Psychology and Education. New York: Longmans, Green and Co., 1926.

15. Messick S. The once and future issues of validity: assessing the meaning and consequences of measurement. In: Wainer H, Braun H. Test Validity. Hillsdale, NJ: Lawrence Erlbaum Assoc., Publishers, 1988.

16. Guion RM. On trinitarian doctrines of validity. Professional Psychology 1980;11:385.

17. Nunnally JC. Psychometric Theory. New York: McGraw-Hill Publishing Company, 1978.

18. Ware JE, Davies-Avery A, Brook RH. Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Vol. VI, Analysis of Relationships Among Health Status Measures. Santa Monica, CA: The RAND Corporation, 1980 (publication number R-1987/6-HEW).

19. Hays RD, Stewart AL. The structure of self-reported health in chronic disease patients. Psychological Assessment 1990;2:22.

20. Rogers W, McGlynn E, Berry S et al. Methods of sampling. In: Stewart AL, Ware JE, eds. Measuring Functioning and Well-Being: The Medical Outcomes Study Approach. Chapel Hill, NC: Duke University Press, 1992.

21. Wells KB, Hays RD, Burnam MA, et al. Detection of depressive disorder for patients receiving prepaid or fee-for-service care: results from the Medical Outcomes Study. JAMA 1989;262:3298.

22. Burnam MA, Wells KB, Leake B, et al. Development of a brief screening instrument for detecting depressive disorders. Med Care 1988;26:775.

23. McHorney CA, Ware JE, Rogers W, et al. The validity and relative precision of MOS short- and long-form health status scales and Dartmouth COOP charts: results from the Medical Outcomes Study. Med Care 1992;30:MS253.

24. Bergner M, Bobbitt RA, Carter WB, et al. The Sickness Impact Profile: development and final revision of a health status measure. Med Care 1981;19:787.

25. Greenwald HP. The specificity of quality-of-life measures among the seriously ill. Med Care 1987;25:642.

26. Hall JA, Epstein AM, McNeil BJ. Multidimensionality of health status in an elderly population: construct validity of a measurement battery. Med Care 1989;27:S168.

27. Brooks WB, Jordan JS, Divine GW, et al. The impact of psychologic factors on measurement of functional status. Med Care 1990;28:793.

28. Comrey AL. A first course in factor analysis. New York: Academic Press, 1973.

29. Wenneker MB, Greenfield S, McHorney CA, et al. The validity of a severity scale for hypertension in predicting functional status and well-being: results from the Medical Outcomes Study. Clin Res 1990;38:228A.

30. Wenneker MB, McHorney CA, Kieszak SM, et al. The impact of diabetes severity on quality of life: results from the Medical Outcomes Study. Clin Res 1991;39:612A.

31. Dixon WJ, Massey FJ. Introduction to statistical analysis. New York: McGraw-Hill Book Company, Inc. 1951.

32. Snedecor GW, Cochran WG. Statistical methods, 8th ed. Ames, Iowa: Iowa State University Press, 1967.

33. Liang MH, Larson MG, Cullen KE et al. Comparative measurement efficiency and sensitivity of five health status instruments for arthritis research. Arthritis Rheum 1985;28:542.

34. Broadhead WE, Blazer DG, George LK, et al. Depression, disability days, and days lost from work in a prospective epidemiologic survey. JAMA 1990;264: 2524.

35. World Health Organization. Constitution of the World Health Organization. In: Basic Documents. Geneva: World Health Organization, 1948.

36. Bergner M. Measurement of health status. Med Care 1985;23:696.

37. Spitzer WO. State of Science 1986: quality of life and functional status as target variables for research. J Chron Dis 1987;40:465.

38. Ware JE. Standards for validating health measures: definition and content. J Chron Dis 1987;40:473.

39. Patrick DL, Erickson P. Assessing health-related quality of life for clinical decision making. In: Walker SR, Rosser RM, eds. Quality of Life: Assessment and Application. Lancaster: MTP Press Limited, 1988.

40. Shapiro MF, Ware JE, Sherbourne CD. Effects of cost sharing on seeking care for serious and minor symptoms: results of a randomized controlled trial. Ann Intern Med 1986;104:246.

41. Lohr KN, Kamberg CJ, Keeler EB, et al. Chronic disease in a general adult population: findings from the RAND Health Insurance Experiments. West J Med 1986;145:537.

42. Wells KB, Rogers W, Burman A, et al. How the medical comorbidity of depressed patients differs across health care settings: results from the Medical Outcomes Study. Am J Psychiatry 1991;148:1688.

43. Nielsen AC, Williams TA. Depression in ambulatory medical patients: prevalence by self-report questionnaire and recognition by nonpsychiatric physicians. Arch Gen Psychiatry 1980;37:999.

44. Rodin G, Voshart K. Depression in the medically ill: an overview. Am J Psychiatry 1986;143:696.

45. Prestidge BR, Lake CR. Prevalance and recognition of depression among primary care outpatients. J Family Practice 1987;25:67.

46. Nelson EC, Berwick DM. The measurement of health status in clinical practice. Med Care 1989; 27:S77.

47. Deyo RA, Patrick DL. Barriers to the use of health status measures in clinical investigation, patient care, and policy research. Med Care 1989;27:S254.

48. International Resource Center for Health Care Assessment. How to score the MOS 36-item short-term health survey (SF-36). Boston: The Health Institute, 1992.

49. Nelson E, Conger B, Douglass R, et al. Functional health status levels of primary care patients. JAMA 1983;249:3331.

50. Calkins DR, Rubenstein LV, Cleary PD, et al. Failure of physicians to recognize functional disability in ambulatory patients. Ann Intern Med 1991;114:451.

Differential item functioning Item bias SF-36 Quality of life. This work was completed while Dr. McHorney was at Indiana University. Dr. McHorney is now with Outcomes Research & Management, Merck & Co., Inc.Â Results from the Whitehall II StudyBrit Med J31512731279PubMedGoogle Scholar. 9. McHorney, C, Ware, J, Raczek, A 1993The MOS 36-item short-form health survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructsMed Care31247263PubMedGoogle Scholar. 10. McHorney, CA, Ware, JE,Jr, Lu, JF, Sherbourne, CD 1994The MOS 36-item Short-Form Health Survey (SF-36): III.

Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs Author(s): Colleen A. McHorney, John E. Ware, Jr., Anastasia E. Raczek Source: Medical Care, Vol. 31, No. 3 (Mar., 1993), pp. 247-263 Published by: Lippincott Williams & Wilkins Stable URL: http://www.jstor.org/stable/3765819 .Â The MOS36-Item Short-FormHealthSurvey (SF-36): II.Psychometricand ClinicalTests of Validityin Measuring. Physicaland MentalHealthConstructs. COLLEENA.Â For purposes of clinical tests of validity, clinical criteria defined mutually exclusive adult patient groups differing in severity of medical and psychiatric conditions. The 36-item short form of the Medical Outcomes Study questionnaire (SF-36) was designed as a generic indicator of health status for use in population surveys and evaluative studies of health policy. It can also be used in conjunction with disease-specific measures as an outcome measure in clinical practice and research (1). Conceptual Basis. The SF-36 derived from the work of the Rand Corporation of Santa Monica during the 1970s. Rand's Health Insurance Experiment compared the impact of alternative health insurance systems on health status and utilization (2; 3, p2:3). The outcome measure... The MOS 36 item short form health survey questionnaire (SF-36) is widely acknowledged as the gold standard generic measure of health status; few studies however have evaluated its use for clinical trials in multiple sclerosis. Its clinical appropriateness, internal consistency reliability, validity, and responsiveness was investigated across a broad range of patients with multiple sclerosis.Â Convergent and discriminant construct validity was supported by the direction, magnitude, and pattern of correlations with other health measures.Â CONCLUSIONS The SF-36 has some limitations as an outcome measure in multiple sclerosis.