Medical Outcomes Study Short Form 36: Testing and Cross-Validating a Second-Order Factorial Structure for Health System Employees

Pamala J. Reed

Objective. To test the factorial validity of the SF-36.

Data Source. Sample data collected in 1995 and 1996 using telephone interviews with health system employees as part of a study of health status.

Methods of Analysis. Confirmatory factor analysis and structural equation modeling techniques were used to evaluate the data.

Principal Findings. The results of this study suggest that (1) Mental Health and Physical Health are not independent; (b) Mental Health cross-loads onto Physical Health; (c) general health loads onto Mental Health instead of Physical Health; (d) many of the error terms are correlated; (e) the physical function subscale is not reliable across the samples or the "age" or "education" subgroups; and (f) the mental health subscale path from Mental Health is not reliable across some subgroups. This hierarchical factor pattern was replicated across both samples.

Conclusions. This study supports the second-order factorial structure of the SF-36. Adding the covariance path between the variables Physical Health and Mental Health improved model fit. Two paths from the second-order latent variables to the first-order latent variables differ from the original hypothesized structure of the SF-36. Health perception was influenced by Mental Health rather than Physical Health, and mental health was influenced by both Mental Health and Physical Health. This cross-loading suggests that the perception of Physical Health greatly affects mental health. Scale instabilities in the SF-36 across subgroups suggest that a comparison of mean scores or summary scores is inappropriate. Data interpretation can be improved if multigroups structural equation modeling is used.

Key Words. SF-36, structural equation modeling, confirmatory factor analysis, factorial validity

It has been recognized for some time that a need and a great demand exist for measures of physical and mental health, social and role functioning, and other general health concepts for use in evaluating healthcare (Stewart, Hays, and Ware 1988; Schroeder 1987). One of the most recently developed and promising tools is the Medical Outcomes Study (MOS) Short Form 36 (SF-36) (Ware and Sherbourne 1992). It has become the tool of choice for measuring health status (Dexter et al. 1996). Rigorous psychometric analysis of the SF-36, however, has been limited to scale reliability, precision, and validity, and the use of exploratory factor analysis or principal component analysis to identify the two latent health constructs, Physical and Mental Health (Hays and Stewart 1990; McHorney, Ware, and Raczek 1993).

The SF-36 is used to compare self-reported health status across groups (Johnson, Goldman, Orav, et al. 1995; McHorney et al. 1994). Implicit in this process is the assumption that the hypothesized item groupings (structure) of the SF-36 are invariant across groups. However, no evidence exists to support this assumption, and there is at least some evidence that the SF-36 may not be invariant across groups. McHorney et al. (1994) reported that SF-36 responses from certain subgroups consisting of people who are elderly, African American, poorly educated, or in poverty were found to be less reliable than those from other study participants. Others have reported that different ethnic or cultural groups may interpret items differently (Coulton, Hyduk, and Chow 1989; Deyo 1984), which can undermine the reliability of hypothesized item groupings. With the use of confirmatory factor analytic (CFA) techniques, Wolinsky and Stump (1996) found that the SF-36 consisted of a ninth factor, referred to as "health optimism," for older, African American and white women, and African American men, and the original eight factors for older, white men. This suggests that the determination of equivalency across subgroups for the SF-36 is necessary in order to provide a meaningful interpretation of the data across groups. Little work has been done to evaluate the factorial reliability and validity of the SF-36. The purpose of this article, therefore, is to report the results of a second-order CFA of the factor validity of the SF-36 in a sample of adult healthcare workers.

The MOS Short Form 36 (SF-36) consists of one multi-item scale measuring eight health concepts: (1) physical functioning (PF); (2) role limitation due to physical health problems (RP); (3) bodily pain (BP); (4) general health

Address correspondence and requests for reprints to Pamala J. Reed, Dr.P.H., M.P.H., Assistant Professor, University of Tennessee–Memphis, College of Pharmacy, Department of Pharmacy Practice and Pharmacoeconomics, 847 Monroe Avenue, Suite 200, Memphis, TN 38163. At the time of this study Dr. Reed was at the University of Texas–Houston, School of Public Health.
This article, submitted to *Health Services Research* on July 1, 1997, was revised and accepted for publication on February 24, 1998.

perception (GH); (5) vitality (VT); (6) social functioning (SF); (7) role limitations because of emotional health problems (RE); and (8) general mental health (MH); and one item measuring self-reported health transition. Criterion validity (how well the scale measures what it purports to measure) and relative precision have been well established for the eight scales found in the SF-36 (Ware and Sherbourne 1992; McHorney, Ware, Rogers, et al. 1992). The 36 items making up the SF-36 were derived from long-form measures of general health.

The long-form measures of general health status that serve as the foundation of the SF-36 were constructed to capture two major dimensions of health, Physical Health and Mental Health. These two dimensions have been empirically confirmed in both general and patient populations (Ware, Davies-Avery, and Brook 1980; Hays and Stewart 1990). Psychometric tests of construct validity have indicated that the SF-36 also captures these two health dimensions (McHorney, Ware, and Raczek 1993). McHorney et al. used principal component analysis to identify the two latent health constructs, Physical and Mental Health.

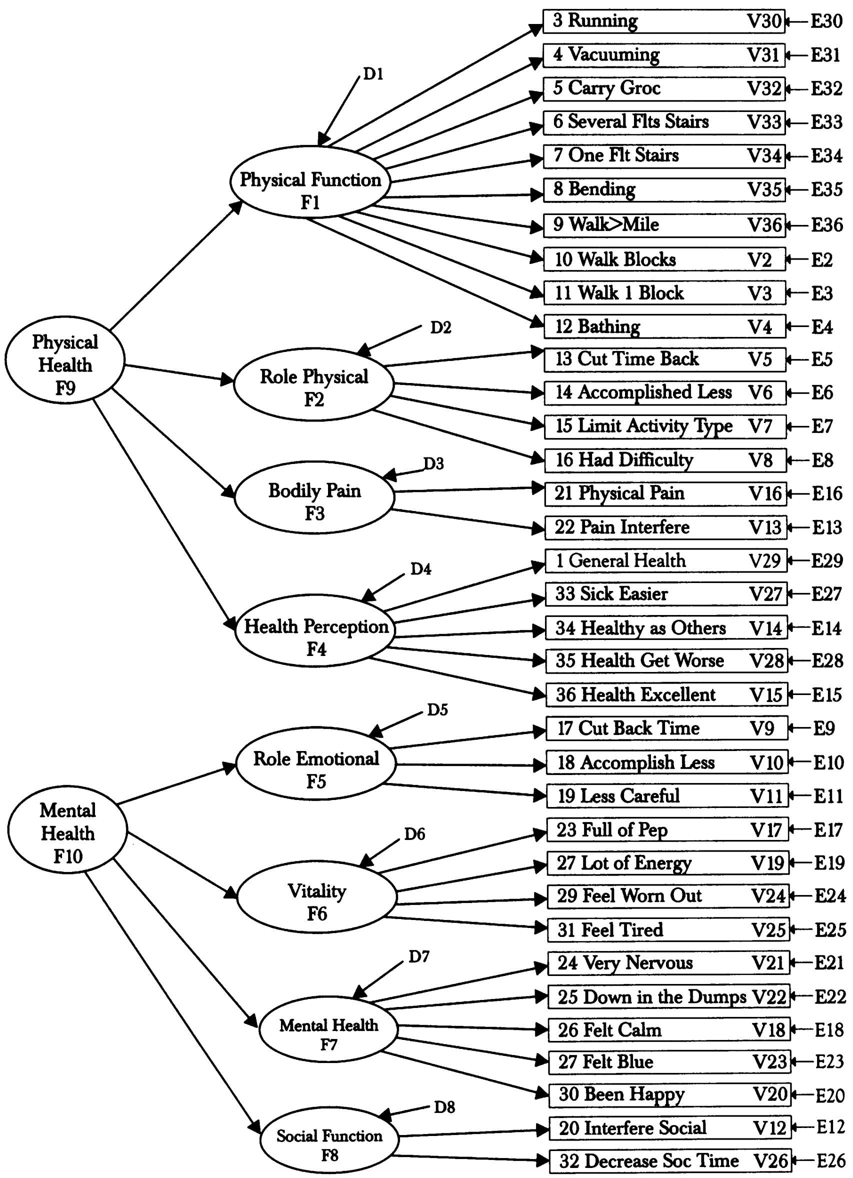
Exploratory factor analysis and principal component analysis have been used to determine the item groups of the SF-36 and to extract the two health constructs (second-order variables). Exploratory factor analysis has been used to evaluate the number of first-order variables (scales) and to see if there are any second-order variables. These techniques do not rely on any a priori theory regarding the item groupings or possible second-order variables; they can suggest underlying patterns in the data (Bollen 1989). However, it is well documented that several deficiencies are associated with exploratory factor analysis and principal component analysis when they are used as a factor-analytic strategy (Snook and Gorsuch 1989; Gorsuch 1990). Structural equation modeling (SEM) offers a more appropriate alternative. SEM consists of model fitting, testing, and equating based on the analysis of covariance structures within the framework of a confirmatory factor-analytic model. Structural equation modeling seeks to test data against the hypothesized or theoretical model.

As depicted in Figure 1, the model under study represents a covariance structure model. Because some readers may be unfamiliar with the symbols conventionally associated with such models, a brief explanation is in order. Covariance structure models can be decomposed into two submodels: a structural model and a measurement model. The structural model defines the pattern of relations among unobserved hypothetical constructs. In the CFA of an assessment instrument, for example, the structural model formulates the pattern of factor intercorrelations as defined by the theory on which the instrument is based. Typically, the structural submodel is identified schematically by the presence of interrelated circles, each of which represents a hypothetical construct (or factor). Turning to Figure 1, we see a hierarchical ordering of circles positioned in such a way that, if the page were turned sideways, the "Physical Health" and "Mental Health" circles would be on the top, with four other circles beneath each one. We interpret this schema as representing two second-order factors (Physical Health and Mental Health), and eight first-order factors (physical functioning; role limitation due to physical health problems, bodily pain; general health perception; vitality; social functioning; role limitations because of emotional health problems; and general mental health). The single-headed arrows leading from the higher-order factor to each of the lower-order factors are regression paths that indicate the causal impact of Physical Health on physical functioning, role limitation due to physical health problems, bodily pain, and general health perception factors, and Mental Health on vitality, social functioning, role limitations because of emotional health problems, and general mental health factors; they represent the second-order factor loadings. Finally, the angled arrow leading to each first-order factor represents residual error in a first-order prediction from the higher-order factors of Physical Health and Mental Health.

The measurement model defines relations between observed variables and unobserved hypothetical constructs. In other words, it provides the link between item scores on the assessment instrument and the underlying factors they were designed to measure. The measurement model, then, specifies the pattern by which each item loads onto a particular factor. This submodel can be identified by the presence of rectangular boxes, each of which represents one SF-36 item, as indicated by the number shown. The single-headed arrows leading from the first-order factor to the boxes are regression paths that link each of the factors to its respective set of observed scores; these coefficients represent the first-order factor loadings. For example, Figure 1 postulates that items 3, 4, 5, and so on load onto the Physical Function factor. All of the possible relations that are not specified by a path are fixed at a value of zero. Finally, the single-headed arrow pointing to each box represents observed measurement error associated with the item variables.

One important omission in Figure 1 is the presence of double-headed arrows among the first-order factors, thereby indicating their intercorrelation. This is because in second-order factor analysis, all covariation among the first-order factors is considered explained by the second-order factors. However, it is possible to have a covariance path between second-order factors.

Figure 1: Hypothesized Structural Equation Model for the SF-36



Multigroup structural equation modeling (MSEM) is uniquely suited for exploring whether mediating relationships among predictors of outcomes may vary by population subgroups because of its ability to test a theoretical model for its applicability to different groups simultaneously, the test for invariance. MSEM models do not require cumbersome interaction terms and nested models to estimate hypothesized group differences in path-analytic model coefficients or model fit. A single χ^2 goodness-of-fit statistic evaluates a set of complex models, one for each group. To validate the usual assumption that groups are equivalent, groups can be required to have identical estimates for all parameters (a fully constrained or universal model). Differences among groups can be evaluated for their appropriateness by "freeing" some parameters (allowing one or more groups to vary uniquely), "fixing" some parameters (setting parameters to zero), and/or "constraining" (requiring two or more groups to have equal parameters) any or all parameters for different groups. MSEM analyses often begin by estimating a fully constrained model, then relaxing constraints to allow for group-specific differences in particular parameters based on theory or inductive evidence (e.g., the Lagrange multiplier test). The Lagrange multiplier tests for the impact of freeing one or more constraints on model fit. The analysis program used in this analysis, EQS (Bentler 1992b; Byrne 1994), has been developed to allow parameters to be added in a stepwise process by order of multivariate significance. This procedure is similar to forward stepwise multiple regression analysis. Therefore, constraints that are found to be statistically significant ($p \leq .05$) and theoretically sound can be relaxed. Additional review of the multivariate LM test provides information on the effect of any change on all of the parameters decreasing the chance of a Type I error.

The purpose of the present study was to test the factorial validity of the model depicted in Figure 1. Specifically, the study (a) tested for the validity of a second-order SF-36 factorial structure, as shown in Figure 1; (b) on evidence of model misfit, determined the best-fitting model; (c) tested for the invariance of factorial measurement and structure across two independent samples from the same population to validate the measurement model; and (d) tested for the invariance of factorial measurement and structure across race, age, and educational subgroups using MSEM methodology.

METHODS

Sample

Data were collected in 1995 and again in 1996, using the same population base and the same survey tool. The population consists of the employees of a large teaching hospital. A 12 percent turnover rate was recorded during the period between the two surveys. The survey tool consisted of the SF-36, 65 questions on disease prevention practices, and two demographic questions.

The 1996 sample data were used to evaluate the fit of the hypothesized model and tests invariance across subgroups. Employees were eligible if they were employed at the time of the 1996 survey and were enrolled with the health maintenance organization (HMO) as their healthcare provider. There were 3,611 persons eligible for participation, and 833 (23 percent) were randomly chosen to be interviewed. Of these 833, 486 (58.3 percent) were accessible by telephone; 75 (9 percent) did not list a telephone; 202 (24.2 percent) telephone numbers were not valid; 69 (8.3 percent) did not answer; of the 486 people contacted by telephone, 87 (10.6 percent) refused to participate. Like many hospitals, this group of employees is made up of a disproportionate number of women (75 percent). Four of the respondents refused to answer any of the SF-36 questions and were dropped from the analysis. One other respondent did not answer a sufficient number of the SF-36 questions to allow for a reliable replacement of the missing data by regression imputation and was also dropped from this analysis. Missing data were 1.8 percent or less for any individual item, and were distributed randomly across items and respondents. Regression imputation was used to replace missing values, resulting in complete data on 394 respondents. Therefore, 394 employee responses (i.e., 81.1 percent of the contacted employees) are used in this analysis.

The 1995 data were used for the cross-validation sample. All full-time employees were eligible to participate in the 1995 study. There were 3,753 employees eligible for participation, and 823 (21 percent) were chosen by a stratified (by job title) random sampling method to be interviewed. Of these 823, 461 (55.3 percent) were accessible by telephone; 362 were not contacted. All of the 461 employees contacted by telephone agreed to participate. Like the 1996 sample, this group of hospital employees is made up of a disproportionate number of women (77 percent). The high level of participant cooperation resulted in a data set with no missing data points. Therefore, all 461 employee responses are used in this analysis.

Instrumentation

The SF-36 items for both the 1996 and 1995 sample data were scored and transformed as recommended. A detailed description and the exact wording of the SF-36 questions are available elsewhere (Ware et al. 1993), and are not repeated here. Mean scores are reported in Table 1 for the 1995 and 1996 samples and by subgroups, age, race, and education, for the 1996 sample. It is not surprising to note that these mean scores are much higher than those from the Medical Outcomes Study (McHorney et al. 1994). For example, MOS physical function mean score is 73 as opposed to 90 (1995 sample) and 96 (1996 sample) in these data. However, they are very similar to those of a healthy HMO population, physical function mean score of 91, obtained in 1990 by a random sample of members by the Geisinger Health Plan in Danville, Pennsylvania.

Analysis of the Data

Tests for the factorial validity of the SF-36, and for its invariance across independent samples and subgroups, were based on the analysis of covariance structures within the framework of the CFA model. The analyses are based on covariance structures. The covariance structure of the observed variables, therefore, constitutes the crucial parametric information. Analyses were conducted in two major stages using the EQS software program (Bentler 1992b; Byrne 1994) designed for performing the complex analysis required for structural equation modeling. First, CFA procedures were conducted for the 1996 sample data testing the hypothesized second-order factorial structure shown in Figure 1. Presented with findings of inadequate fit, an examination of the parameters identified by the Lagrange multiplier test (LM test) as those that would contribute most to a significantly better-fitting model was undertaken. If the inclusion of these parameters was deemed to be substantively and psychometrically reasonable, the model was respecified accordingly. Invariance testing across groups assumes well-fitting single group models. Second, the final best-fitting model from Stage 1 was tested for its invariance across the 1995 sample and to test the invariance across subgroups using the 1996 sample.

Multiple criteria were used in the assessment of model fit: (a) the χ^2 likelihood ratio statistic; (b) the Satorra-Bentler Scaled Statistic (S-Bχ^2) (Satorra and Bentler 1988a,b); (c) the normed and non-normed fit indexes (NFI, NNFI) (Bentler and Bonett 1980); and (d) the Comparative Fit Index (CFI) (Bentler 1990). The S-Bχ^2 incorporates a scaling correction for the χ^2 statistic when distributional assumptions are violated. Its computation takes into account the model, the estimation method, and the sample kurtosis values (Hu, Bentler, and Kano 1992). The S-Bχ^2 has been shown to more loosely approximate χ^2 than the usual test statistic, to have robust standard errors, and to perform as well as or better than the usual asymptotically distribution free methods generally recommended for non-normal multivariate data (Bentler 1992a; Hu, Bentler, and Kano 1992). The CFI is a revised version of the

Table 1: SF-36 Mean Scores for Health System Employees

Characteristics	*Physical Function*	*Role Limited by Physical Health*	*Bodily Pain*	*Health Perception*	*Role Limited by Emotional Health*	*Vitality*	*General Mental Health*	*Social Function*
1995 Sample	90	92	87	81	95	68	82	91
1996 Sample	96	87	84	80	92	64	82	85
1996 Subgroups								
Age								
40 < yrs old	97	88	86	81	93	63	83	85
40 ≥ yrs old	95	86	82	80	92	66	84	86
Ethnicity								
White	95	87	83	80	91*	63	83	85
Nonwhite	96	87	84	81	94*	65	85	85
Education								
High school grad or less	91*	87	82*	78*	92	63*	83*	83*
Some college or more	97*	87	85*	82*	93	65*	85*	87*

* Denotes a statistically significant difference between subgroups, $p < .05$.

Bentler and Bonett (1980) NFI that adjusts for degrees of freedom. The CFI ranges from 0 (poor fit) to 1.00 (perfect fit) and is derived from the comparison of a restricted model (i.e., one in which structure is imposed on the data) with a null model (one in which each observed variable represents a factor). The CFI provides a measure of complete covariation in the data; a value >.90 indicates a psychometrically acceptable fit to the data. The corrected CFI value (CFI*) computed from the S-Bχ^2 statistic for the null model is also reported. It is important to note, however, that the S-Bχ^2 statistic is not yet available for multigroup analyses in the current version of the EQS program; these values are therefore reported only for the single-group analysis.

RESULTS

The CFA model in the present study, based on the original model of the SF-36 (Ware and Sherbourne 1992), hypothesized a priori that (a) responses to the SF-36 could be explained by eight first-order factors and two second-order factors of Physical Health and Mental Health; (b) each item would have a nonzero loading on the first-order factor it was designed to measure and zero loadings on the other seven first-order factors; (c) error terms associated with each item would be uncorrelated; and (d) covariation among the eight first-order factors would be explained fully by their regression onto the second-order factors (Figure 1).

Preliminary analyses identified one multivariate outlier for the 1996 sample and one multivariate outlier for the 1995 sample; deletion of these cases resulted in a sample size of 393 (1996 sample) and a sample size of 460 (1995 sample). As expected, both samples demonstrated evidence of univariate positive kurtosis of 5.20 for the 1996 sample and 3.63 for the 1995 sample. Given that skewness and kurtosis values are zero for data that are normally distributed (albeit values ranging from -1.00 to $+1.00$ may be considered to be approximately normal [Muthén and Kaplan 1985]) it is easy to see how far the present data deviate from these criteria. Although non-normality is not likely to affect the maximum likelihood estimates, it can lead to downwardly biased standard errors that result in an inflated number of statistically significant parameters (Muthén and Kaplan 1985). Given the abnormally high degree of kurtosis associated with the present data, it was deemed critical that the final assessment of statistical fit be based on the S-Bχ^2, and on its related CFI* value, both of which correct for this violation.

Stage 1: Tests of the SF-36 Hypothesized Model

The SF-36 hypothesized model was estimated using the maximum likelihood method, and the chi-square (χ^2) value for the model was statistically significant, $\chi^2(552, N = 394) = 2238.78, p < .001$; $\chi^2/\text{df} = 4.06$, which indicates a poor fit. A number of other results also indicate that there is a problem with the model's fit. The Bentler and Bonett (1980) NFI for this model is .681, the NNFI is .717, Bentler's CFI (1992a) is .737 and, most importantly, the S-Bχ^2 = 1093.47 and the CFI* = .705 (Table 2). These results suggest that the hypothesized model of the SF-36 does not well describe the relationships among the measured items of the SF-36 in this population.

A review of the multivariate LM test χ^2 statistic revealed that model respecification could yield a substantially better fit if the exogenous latent variables, Physical Health and Mental Health, were allowed to covary, which suggests that they are correlated. Adding such a path would be consistent with the evidence from medical sociology, and there are good theoretical arguments for covariance paths in such models (Harman 1976; Nunnaly and Bernstein 1994). Therefore, a covariance path was added between these latent variables and the model was re-estimated, revised model 1 (Table 2). This reparameterization resulted in a substantially better-fitting model, as demonstrated in Table 2 by the significant decrease in χ^2 and S-Bχ^2 and the significant increase in the other fit parameters; however, the fit is still not adequate.

Further review of the multivariate LM test statistics indicated that additional model improvement could be attained by changing the path from Physical Health to general health perception to a path from Mental Health to general health perception. This path suggests that Mental Health explains

Table 2: Summary of Fit Statistics for Second-Order Models of the SF-36 Factorial Structure

	Model Fit							*Change*	
Model	χ^2	*df*	*NFI*	*NNFI*	*CFI*	*S-B*χ^2	*CFI**	χ^2	*df*
Hypothesized model	2238.78	552	0.681	0.717	0.737	1093.5	0.705		
Revised model 1	2000.62	551	0.715	0.756	0.774	992.67	0.760	238	1
Revised model 2	1956.32	550	0.721	0.763	0.781	970.82	0.771	67	1
Final model	1122.43	528	0.836	0.893	0.905	687.56	0.910	834	22

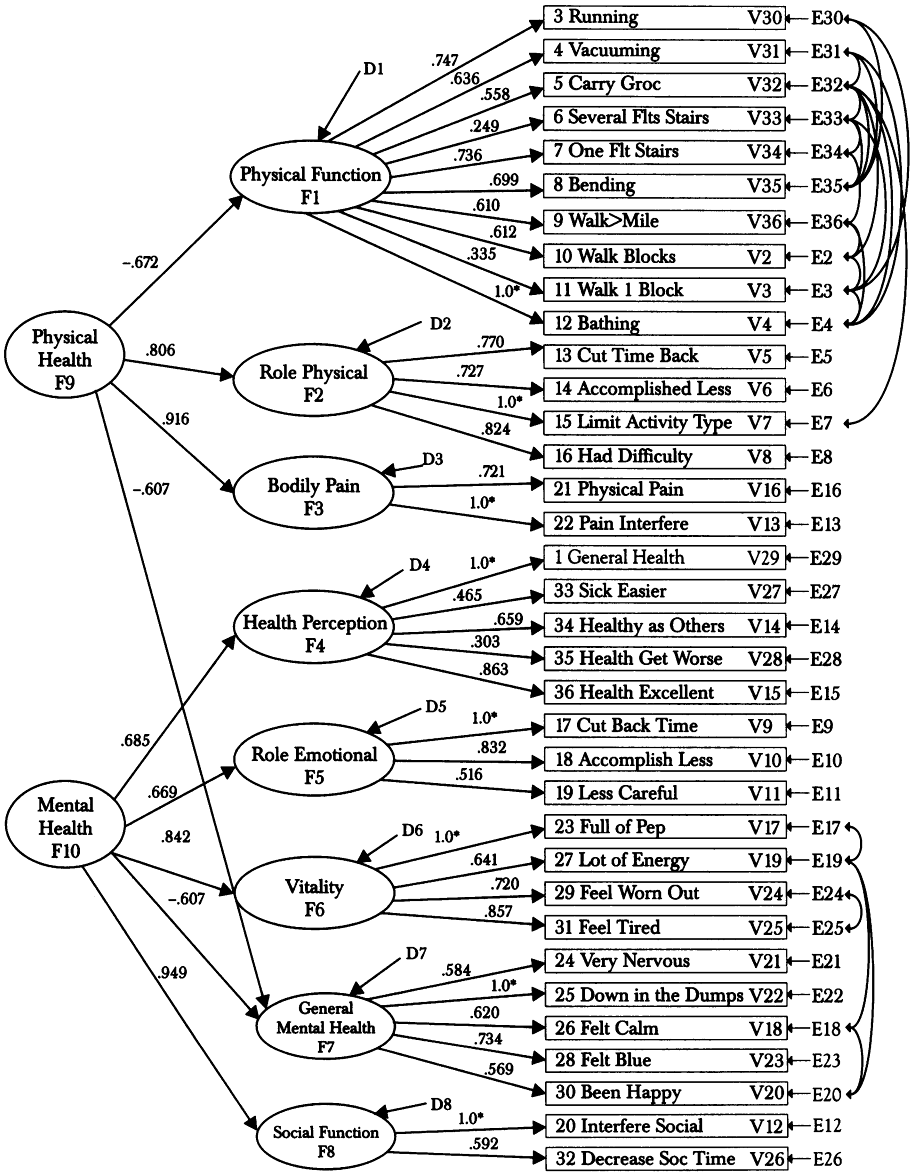
Note: $N = 395$; df = degrees of freedom; NFI = normed fit index; NNFI = non-normed-fit index; CFI = comparative fit index; S-B χ^2 = Satorra-Bentler Scaled Statistic; CFI* = comparative fit index based on the S-B χ^2.

general health perception better than Physical Health. The multivariate LM test statistic also indicated that an additional path from Physical Health to general mental health would result in a substantial improvement in model fit. As shown in Table 2, revised model 2, which includes this reparameterization, resulted in a substantially better-fitting model, although the fit is still not adequate.

Additional review of the multivariate LM test statistics revealed that model respecification could yield a substantially better fit if the error terms associated with 22 pairs of the items were free to covary. These error covariances involved Item 19 (full of pep) and Item 17 (full of energy); Item 20 (feel happy) and Item 18 (feel calm); Item 19 (full of energy) and Item 18 (feel calm); Item 20 (feel happy) and Item 19 (full of energy); Item 32 (carry groceries) and Item 7 (limit type of activity because of physical limitations); Item 25 (feel worn-out) and Item 24 (feel tired); and 16 pairs of items measuring physical activities, for instance, Items 4 and 3, bathing and able to walk one block; Items 34 and 33, climb one flight of stairs and climb several flights of stairs. All of the error covariances are displayed in Figure 2; they are represented by the double-headed arrows going from the error terms on the right side of the page. Since the multivariate LM test statistics associated with these item pairs were distinctively larger than all remaining ones, and because findings or error covariances are not unusual in the validation of assessment instruments in general—and because many of these measures are interdependent (e.g., climbing several flights of stairs leads to a positive response to climbing one flight of stairs) or measure very similar feelings (e.g., feeling worn-out and feeling tired)—the model was respecified to include the estimation of these parameters. As shown in Table 2, this reparameterization, the measurement model, resulted in a substantially better-fitting and quite adequate model; all 22-error covariances were statistically significant.

Although a review of the multivariate LM test statistics indicated that still further model improvement could be attained by estimating several additional error covariances, these error parameters tend to be associated with the idiosyncratic interpretation of item content, and thus reflect "noise" in the data rather than any sound structural change. In the interest of parsimony, it was determined to cease further post hoc model-fitting. A summary of the measurement model is presented schematically in Figure 2, standardized estimates are reported; asterisked loadings denotes a parameter fixed to 1.0 for purposes of statistical identification.

Figure 2: Measurement Model for the SF-36



Stage 2: Tests for Invariance

In testing for invariance across samples or subgroups, sets of parameters are put to the test in a logically ordered and increasingly restrictive fashion. Depending on the model and hypotheses to be tested, the following sets of parameters are typically of interest in answering questions related to group invariance: (a) factor loading paths; (b) factor variances/covariances; (c) structural regression paths; (d) factor residuals; and (e) error variances/covariances. Except in particular instances, the equality of error variances and covariances is probably the least important hypothesis to test (Bentler 1992a). Although the Jöreskog tradition of invariance testing holds that the equality of these parameters should be tested, it is now widely accepted that to do so represents an overly restrictive test of the data.

In EQS, we can test simultaneously for the invariance of both the first- and second-order factor loadings. Therefore, in testing for invariant factorial structure, all first- and second-order factor loadings were constrained equally across subgroups and then were tested statistically in a simultaneous analysis of the data; error covariances were allowed to be freely estimated. Judgment of replicability was based on two criteria: (a) goodness-of-fit of the constrained model; and (b) probability level of the equality constraints as determined by the Lagrange multiplier test (LM test), with equality constraints of $p < .05$ being untenable. Readers are reminded that χ^2 values for the multigroup models are based on the uncorrected, rather than on the corrected Satorra-Bentler Scaled (S-Bχ^2) statistic (Satorra and Bentler 1988a,b). Therefore, the χ^2 values are expected to be substantially larger than would be the case for the S-Bχ^2 statistic.

Cross-Validation. The use of an additional sample from the same population will allow for the determination of the validity of the measurement model. Goodness-of-fit for this two-group constrained model yielded a χ^2 (1,063, $N = 853$) = 2707.14, $p < .001$; $\chi^2/\text{df} = 2.55$, and CFI = .88.

Examination of the probability values revealed one second-order factor loading to be nonequivalent across the two samples. Relaxing the constraint on the physical function scale improved the fit significantly; χ^2 decreased by 151.68, $p < .001$. All of the remaining constraints were found to be equal across the two samples.

White/Nonwhite. The 1996 sample data were divided into two groups: Group 1 consists of 207 respondents who identified themselves as Caucasian and Group 2 consists of 180 respondents who identified themselves as either African American, Asian, Hispanic, or other; seven respondents did not provide information on race and were not included in this analysis. Preliminary analysis did not reveal any multivariate outliers in either group. However, both groups demonstrated evidence of univariate positive kurtosis of 4.29 for Group 1 and 6.42 for Group 2. Group 2 is significantly more kurtotic than Group 1. Goodness-of-fit for this two-group constrained model yielded a χ^2 (1065, $N = 387$) = 2294.77, $p < .001$; χ^2/df = 2.15, and CFI = .83 (Table 3).

Examination of the multivariate probability values revealed one second-order factor loading to be untenable. Relaxing the constraint on the general mental health scale path from Mental Health improved the fit significantly; χ^2 decreased by 9.77, $p < .002$, thereby arguing for its nonequivalence across white healthcare workers and nonwhite healthcare workers. All of the remaining constraints were found to be equal across the two groups.

Education. The 1996 sample data were divided into two groups: Group 1 consists of 193 respondents who reported having no more than a high school education and Group 2 consists of 200 respondents with more than a high school education; one respondent did not provide information on educational level and was dropped from this analysis. Preliminary analysis did not reveal any multivariate outliers in either group. However, both groups demonstrated evidence of univariate positive kurtosis of 5.05 for Group 1 and 5.94 for Group 2. Goodness-of-fit for this two-group constrained model yielded a χ^2 (1,065, $N = 393$) = 2,190.59, $p < .001$; χ^2/df = 2.06, and CFI = .89 (Table 3).

Examination of the multivariate probability values revealed two second-order factor loadings to be untenable. First, relaxing the constraint on the general mental health scale path from Mental Health improved the fit significantly; χ^2 decreased by 15.81, $p < .001$. Second, relaxing the constraint on the PF scale path from Physical Health improved the fit significantly; χ^2 decreased by 7.74, $p < .005$. These findings argue for nonequivalence across the two groups of healthcare workers. All of the remaining constraints were found to be equal across the two groups.

Table 3: Summary of Fit Statistics for Subgroup Analysis

	Model Fit						*Change*	
Model	χ^2	*df*	N	p	χ^2	*CFI*	*Free Parameter*	$\Delta\chi^2$ *
White/Nonwhite	2294.77	1065	387	<.001	2.15	0.83	Mental health	9.77
Age	1807.31	893	394	<.001	2.02	0.850	Physical function	4.97
Education	2190.59	1065	393	<.001	2.06	0.89	Mental health	15.81
							Physical function	7.74

Note: df = degrees of freedom; NFI = normed fit index; NNFI = non-normed-fit index; CFI = comparative fit index; *All reported changes in χ^2 statistically significant.

DISCUSSION

Findings from the present study offer support for the second-order factorial structure of the SF-36 as proposed by Ware and Sherbourne (1992). However, the results of this study suggest that (a) Mental Health and Physical Health covary and are not independent; (b) MH cross-loads onto Physical Health; (c) GH loads onto Mental Health instead of Physical Health; (d) many of the error terms are correlated; (e) the PF scale is not reliable across the two independent samples or across the "age" or "education" subgroups; and (f) the MH scale path from Mental Health is not reliable across the "white/nonwhite" or "education" subgroups.

Adding the covariance path between the second-order latent variables, Physical Health and Mental Health, markedly improved the fit of the model. This suggests that the current practice of principal components analysis with orthogonal rotation may be misleading. Researchers who use the components as independent variables in research should be aware that they may be substantially correlated and that they are not independent predictors of outcomes.

Two paths from the second-order latent variables to the first-order latent variables differ from the original hypothesized structure of the SF-36. In this population health perception was influenced by Mental Health rather than Physical Health and general mental health was influenced by both Mental Health and Physical health. The cross-loading of general mental health onto Physical Health suggests that a person's perception of Physical Health has a greater effect on that person's general mental health than has been predicted. It is possible that these findings are unique to this generally healthy population of healthcare workers. Krause and Jay (1994) have reported that all respondents to self-rated health items, such as those in the SF-36 items designed to measure health perception, do not use the same frame of reference when answering these questions. Some respondents think about specific health problems, while others think in terms of physical functioning or health behaviors. This issue of specific referents was not evaluated in this study. However, a more interesting explanation is the possible effect of negative affectivity on measures of self-reported health status. Watson and Pennebaker (1989) reported negative affectivity as a "general nuisance factor" when self-report health measures are used. They recommended the inclusion of an established trait negative affectivity marker for identifying and isolating its influence in health research. Further research to evaluate negative affectivity and its impact on the SF-36 and other similar tools should be pursued. Future research using

SEM should provide more information regarding the generalizability of these paths to other populations. This hierarchical factor pattern was replicated across both samples of healthcare workers, suggesting that the post hoc model fitting was not data specific.

Although the best-fitting factor model for healthcare workers included correlated errors for particular pairs of items, such findings are not unexpected in multigroup analysis in general. These parameters are typically unstable, and represent systematic rather than random measurement error in item responses that may reflect bias such as yea/nay-saying and social desirability (Aish and Jöreskog 1990), or idiosyncratic interpretation of item content. Because the majority of the correlated errors (16 of 22) are found in the item measures for PF, it is not surprising to find that the PF scale is unstable across these two samples even though they are independent samples from the same population. In general, this is a healthy population of healthcare workers who may have found it difficult to respond to the items measuring low-level physical function (e.g., "can you bathe yourself?"). In the subgroup analysis for "age," variance in the three measures of low-level activity was insufficient in the younger group to include them in the model. This suggests that other samples or subgroups could have a similar difficulty; for example, severe arthritis or heart disease patients may respond to the variables measuring high-level physical function (running) with the same negative response—"never." Creating a scale that is capable of discriminating varying levels of physical function may have led to problems with structured error and/or idiosyncratic interpretation of the item content. Further study in this area should determine whether this problem with the physical function scale is found in other populations.

However, in two sets of subgroups, "age" and "education," the physical function scale was found to be variant. This would suggest that the perception of Physical Health varies in its impact on physical function between younger and older subgroups and between the two subgroups divided by educational level. This indicates that direct comparisons of mean scale scores or the use of these scores to create summary scores (Ware, Kosinski, Bayliss, et al. 1995) would not be meaningful since the latent variable (scale) is unreliable and is therefore not valid.

In two of the subgroups, "white/nonwhite" and "education," the path from Mental Health to general mental health was found to be variant. This suggests that the effect of the perception of Mental Health on general mental health varies between whites and nonwhites and between those with a high school education or less and those with more than a high school education.

This instability suggests that comparisons involving these scales and between these subgroups are not meaningful.

What does this mean, then, for the SF-36? First, there is substantial support for some of the original hypothesized structure. In this population, there were eight scales with two latent factors and six of the eight scales (RP, BP, GH, RE, VT, and SF) appear to be reliable. However, the comparison of mean scale scores for the physical function scale and the general mental health scale is not meaningful because the two scales are not reliable. The remaining six scales are reliable and valid, lending themselves to meaningful comparisons. The use of summary scale scores in this population would not result in any meaningful interpretation of comparisons. Summary scale scores assume that the PF, RP, BP, and GH scales are representative of Physical Health and that the RE, VT, MH, and SF scales are representative of Mental Health (as demonstrated in Figure 1). However, that is not the case in this population, as shown in Figure 2. Two of the scales, PF and MH, are not reliable, hence not valid, and GH was representative of Mental Health as opposed to Physical Health. These findings violate the assumptions of the summary scale score model (Ware, Kosinski, and Keller 1994). The SF-36 is a valuable assessment tool, and evaluations of the SF-36 using CFA can provide researchers and practitioners with important information about the characteristics of the SF-36.

The importance of this study is in the use of SEM and MSEM in evaluating sample data from the use of the SF-36. SEM is uniquely suited to the analysis of latent variable structures and is widely used in other fields of study. The use of latent variable models for self-reported outcome measures has become widespread. Using MSEM for the testing of invariance across groups or subgroups is a far superior method of determining differences between groups or subgroups than that of mean scores or summary scores. These results also add to the existing evidence that argues for the hierarchical factorial structure as the best model for the SF-36. From a practical, as well as psychometric perspective, it seems imperative that construct validity research related to the SF-36 establishes whether this same hierarchical structure and invariance hold for other populations.

REFERENCES

Aish, A. M., and K. G. Jöreskog. 1990. "A Panel Model for Political Efficacy and Responsiveness: An Application of LISREL 7 with Weighted Least Squares." *Quality and Quantity* 24 (3): 405–26.

Bentler, P. M. 1990. "Comparative Fit Indices in Structural Models." *Psychological Bulletin* 107 (2): 238–46.

———. 1992a. *EQS Structural Equation Program Manual.* Los Angeles: BMDP Statistical Software.

———. 1992b. "On the Fit of Models to Covariances and Methodology to the Bulletin." *Psychological Bulletin* 112 (3): 400–404.

Bentler, P. M., and D. G. Bonett. 1980. "Significance Tests and Goodness-of-Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88 (4): 588–606.

Bollen, K. A. 1989. *Structural Equations with Latent Variables.* New York: John Wiley & Son.

Byrne, B. M. 1994. *Structural Equation Modeling with EQS and EQS/Windows: Basic Concepts, Applications, and Programming.* Thousand Oaks, CA: Sage Publications, Inc.

Coulton, C. J., C. M. Hyduk, and J. C. Chow. 1989. "An Assessment of the Arthritis Impact Measurement Scales in 3 Ethnic Groups." *Journal of Rheumatology* 16 (8): 1110–15.

Dexter, P. R., T. E. Stump, W. M. Tierney, and F. D. Wolinsky. 1996. "The Psychometric Properties of the SF-36 Health Survey Among Older Adults in a Clinical Setting." *Journal of Clinical Geropsychology* 1 (1): 225–31.

Deyo, R. A. 1984. "Pitfalls in Measuring the Health Status of Mexican Americans: Comparative Validity of the English and Spanish Sickness Impact Profile." *American Journal of Public Health* 74 (6): 569–73.

Gorsuch, R. L. 1990. "Common Factor Analysis Versus Component Analysis: Some Well and Little Known Facts." *Multivariate Behavioral Research* 25 (1): 33–39.

Harman, H. H. 1976. *Modern Factor Analysis, 3d Ed., Rev.* Chicago: University of Chicago Press.

Hays, R. D., and A. L. Stewart. 1990. "The Structure of Self-Reported Health in Chronic Disease Patients." *Psychological Assessment* 2 (1): 22–30.

Hu, L. T., P. M. Bentler, and Y. Kano. 1992. "Can Test Statistics in Covariance Structure Analysis Be Trusted?" *Psychological Bulletin* 112 (2): 351–62.

Johnson, P. A., L. Goldman, E. J. Orav, T. Garcia, S. D. Pearson, and T. H. Lee. 1995. "Comparison of the Medical Outcomes Study Short-Form 36-Item Health Survey in Black Patients and White Patients with Acute Chest Pain." *Medical Care* 33 (2): 145–60.

Krause, N. M., and G. J. Jay. 1994. "What Do Global Self-Rated Health Items Measure?" *Medical Care* 32 (9): 930–42.

McHorney, C. A., J. E. Ware, J. F. R. Lu, and C. D. Sherbourne. 1994. "The MOS 36-Item Short-Form Health Status Survey (SF-36): III. Tests of Data Quality, Scaling Assumptions, and Reliability Across Diverse Patient Groups." *Medical Care* 32 (1): 40–66.

McHorney, C. A., J. E. Ware, and A. B. Raczek. 1993. "The MOS 36-Item Short-Form Health Status Survey (SF-36): II. Psychometric and Clinical Tests of Validity in Measuring Physical and Mental Health Constructs." *Medical Care* 31 (3): 247–63.

McHorney, C. A., J. E. Ware, W. Rogers, A. B. Raczek, and J. F. R. Lu. 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 (Supplement): MS253–65.

Muthén, B., and D. Kaplan. 1985. "A Comparison of Methodologies for the Factor Analysis on Non-normal Likert Variables." *British Journal of Mathematical and Statistical Psychology* 38 (1): 171–89.

Nunnaly, J. C., and I. H. Bernstein. 1994. *Psychometric Theory*, 3d. Ed. New York: McGraw-Hill.

Satorra, A., and P. M. Bentler. 1988a. "Scaling Corrections for Chi Square Statistics in Covariance Structure Analysis." *American Statistical Association 1988 Proceedings of the Business and Economic Sections*, pp. 308–13. Alexandria, VA: American Statistical Association.

———. 1988b. *Scaling Corrections for Statistics in Covariance Structure Analysis*. UCLA Statistics Series 2. Los Angeles: University of California at Los Angeles, Department of Psychology.

Schroeder, S. A. 1987. "Outcome Assessment 70 Years Later: Are We Ready?" *The New England Journal of Medicine* 316 (3): 160–62.

Snook, S. C., and R. L. Gorsuch. 1989. "Component Analysis Versus Common Factor Analysis: A Monte Carlo Study." *Psychological Bulletin* 106 (1): 148–54.

Stewart, A. L., R. D. Hays, and J. E. Ware. 1988. "The MOS Short-Form General Health Survey: Reliability and Validity in a Patient Population." *Medical Care* (26): 724-31.

Ware, J. E., A. Davies-Avery, and R. H. Brook. 1980. *Conceptualization and Measurement of Health for Adults in the Health Insurance Study: Vol. VI, Analysis of Relationships Among Health Status Measures*. Pub. No. R-1987/6-HEW). Santa Monica, CA: The RAND Corporation.

Ware, J. E., M. Kosincki, M. S. Bayliss, C. A. McHorney, W. H. Rogers, and A. Raczek. 1995. "Comparison of Methods for the Scoring and Statistical Analysis of SF-36 Health Profile and Summary Measures: Summary of Results from the Medical Outcomes Study." *Medical Care* 33 (4): AS264–79.

Ware, J. E., M. Kosinski, and S. D. Keller. 1994. *SF-36 Physical and Mental Health Summary Scales: A User's Manual*. Boston: The Health Institute.

Ware, J. E., and C. D. Sherbourne. 1992. "The MOS 36-Item Short-Form Health Survey (SF-36): I. Conceptual Framework and Item Selection." *Medical Care* 30 (6): 473–83.

Ware, J. E., K. K. Snow, M. Kosinski, and B. Gandek. 1993. *SF-36 Health Survey Manual and Interpretation Guide*. Boston: Health Institute, New England Medical Center Hospitals.

Watson, D., and J. W. Pennebaker. 1989. "Health Complaints, Stress, and Distress: Exploring the Central Role of Negative Affectivity." *Psychological Review* 96 (2): 234–54.

Wolinsky, F. D., and T. E. Stump. 1996. "A Measurement Model of the Medical Outcomes Study 36-Item Short-Form Health Survey in a Clinical Sample of Disadvantaged, Older, Black, and White Men and Women." *Medical Care* 34 (6): 537–48.