FINAL REPORT

HOS/VA Comparison Project

Part 1. Measurement Equivalence of Medicare HOS SF-36 & VA Veterans SF-36

Avron Spiro III, Austin F. Lee, Lewis E. Kazis,
Donald R. Miller, Xinhua S. Ren, Milanda Zhang

Boston University School of Public Health

CONTENTS

VI. Results of equivalence models by gender

The Centers for Medicare & Medicaid Services' Office of Research, Development, and Information (ORDI) strives to make information available to all. Nevertheless, portions of our files including charts, tables, and graphics may be difficult to read using assistive technology.

Persons with disabilities experiencing problems accessing portions of any file should contact ORDI through e-mail at ORDI_508_Compliance@cms.hhs.gov

Part 1. Measurement Equivalence of Medicare HOS SF-36 & VA Veterans SF-36

0. ABSTRACT

Objective: Determine whether the versions of the SF-36 administered in the Health Outcomes Survey (HOS) and Department of Veterans Affairs 1999 National Survey of Veterans Health (VA99) are comparable. There are some notable differences (see below) in the versions of the SF-36 used and in the populations sampled.

Methodology: Compare factor structure of SF-36 between samples from Medicare's Health Outcomes Survey and VA's 1999 National Survey of Veterans Health, to assess measurement equivalence. Multiple-group confirmatory factor analysis was used to test a sequence of nested hypotheses representing varying degrees of equivalence. Samples included adults aged 65 and older, 167,092 from Cohort 1 of Medicare HOS, and 477,477 from the 1999 VA survey. Analyses were conducted separately for men and women.

Results: Because of differences between HOS and VA in gender and age distributions, separate analyses were conducted for men and women. There were notable differences between HOS and VA samples, with HOS being older on average, having more women, and more non-white women. The VA participants reported more medical conditions.

Measurement equivalence analyses were conducted using multiple-groups factor analysis to compare the SF-36 between the HOS and VA samples, separately for men and women. The results of an extensive series of analyses suggested that, for both men and women, the best-fitting model is one that specified equivalence of factor loadings between HOS and VA on 6 of 8 factors. Because of differences in the response format of RP and RE items between HOS and VA versions of the SF-36, we did not constrain loadings on these items to equality across samples. The analyses also suggested that imposing further degrees of equivalence between HOS and VA was not appropriate, e.g., the intercepts and unique variances differed significantly between HOS and VA men and women.

Conclusions: The degree of measurement equivalence found, that of partial metric equivalence (i.e., equal factor loadings on 6 of 8 scales, excepting RP and RE), suggests that SF-36 scales can be computed and compared between HOS and VA. Although other aspects of the SF-36 differ between the two samples, for example, unique variances and factor covariances, the establishment of partial metric equivalence indicates that quantitative comparisons between these two samples are appropriate.

Measurement Equivalence of Medicare HOS SF-36 & VA Veterans SF-36

The SF-36 was developed as a brief self-report measure of health status. It can be used to assess outcomes in clinical trials, monitor quality of care in medical practice and response to treatment or intervention, compare the inputs and outputs of different medical systems, characterize or compare the health of populations, or examine changes within a population over time. The latter two uses are relevant for the present project, which is to assess the equivalence of the SF-36 across two healthcare systems, to determine whether it can be used to compare change in these systems over time.

However, before such questions can be addressed, we must first establish the equivalence of the instrument across those systems. That is, do SF-36 items relate to scales in the same manner across patients in different health care systems? If they do, then comparisons can be conducted across systems; if the relations of items to factors differ among groups, then comparisons of groups are misleading at best.

The goal of this project is to examine the extent of measurement equivalence (Byrne et al., 1989; Meredith & Horn, 2001; Widaman & Reise, 1997) between two versions of the SF-36, the version administered in the Medicare Health Outcomes Survey (based on the MOS SF-36), and the Veterans SF-36 (Kazis et al., 1998, 2002) developed and administered in the Veterans Health Administration, US Department of Veterans Affairs. These versions of the SF-36 have been adopted by their respective Federal Agencies to assess the outcomes of health care.

However, recipients of VA healthcare are likely to differ in important ways from those enrolled in Medicare HMO programs. Persons in the two healthcare systems are not likely to be random samples from the same population. For example, we have reason to expect important differences in the age and sex structure of the two systems, and to expect that the health of VA patients is worse than that of Medicare HMO enrollees (e.g., Agha et al., 2000; Kazis et al., 1998; Peabody et al., 1998; Petersen et al., 2000; Rogers et al., 2002).

Given this, it becomes important to know whether health status can be compared across those who use the health care provided by these agencies. Thus, in this paper, we describe the use of multi-group confirmatory factor analysis (Joreskog, 1971) to examine the extent to which the two versions of the SF-36 are comparable across systems, using data from the Medicare 1998 Health Outcomes Survey and from the VA 1999 National Survey of Veterans Health.

1. Samples

1.1. HOS Cohort 1

The HOS was first fielded in May 1998 as part of HEDIS 3.0 by the National Committee on Quality Assurance (NCQA)/CMS. Simple random samples of 1,000 beneficiaries who had been enrolled for at least 6 months (and were not ESRD patients) were selected from each of 268 plans in 287 market areas. (For plans with fewer than 1000 members, all eligible members were selected).

Potential respondents were mailed a pre-notification letter, followed 1 week later by a survey. A reminder was mailed 2 weeks later, followed by a second copy of the survey 2 weeks after that. After a second reminder, a number of attempts were made to contact the potential respondent by phone.

The sample for Cohort 1 was 279,135 persons, of whom 167,092 (60%) provided completed surveys. Of these, we included 166,104 whom we deemed "valid" cases (i.e., survey disposition = M/T 10/11; did not have a marker variable for invalid survey (i.e., *invsrv*=0); and were age 65 and older. Of this sample, 42.5% (n=70,610) were men; 57.5% (n=95,494) women.

1.2. 1999 National Survey of Veterans Health

The 1999 VA data were obtained from a stratified random sample of 3,421,388 VA enrollees (based on the VA enrollment file) as of 1999. Of those enrolled, 1,406,049 were sampled; 887,775 (63.14%) completed the survey.

Data collection took place between July 1999 and January 2000. A modified Total Design Methodology (TDM) approach developed by Dillman was used. This approach is based upon rewarding the respondent, establishing trust and reducing respondent costs. It uses four carefully spaced mailings: (1) a pre-notification letter, (2) cover letter and Veterans SF-36 questionnaire, (3) reminder post card, and (4) second wave of questionnaire mailings to the non-respondents of the first wave mailings. All occurred over 12 weeks, with a 14 week follow-up period for questionnaire receipts.

Information on all 1.4 million sampled enrollees was obtained from VA administrative data (i.e., Patient Treatment and Outpatient Files) to provide socio-demographic characteristics and other administrative data (e.g., Service Connected Disability Status). ICD-9-CM codes for diagnoses were also obtained from these files. Individual identifiers were subsequently stripped to maintain confidentiality. The ICD-9-CM diagnoses were linked to medical and mental conditions based on literature review and a consensus panel of clinicians.

Of the respondents to the VA survey, 477,477 (53.8%) were aged 65 and over. Because the VA is a health care system for veterans, the vast majority (93.4%) were men (n=445,816); only 6.6% (n=9329) were women.

1.3. Differences between HOS and VA samples

Because of the difference between the Medicare HOS sample and the VA for the proportion of women (57.5% in HOS; 6.6% in VA), all analyses have been stratified by sex. Thus, in the analyses below, we compare HOS to VA, separately for men and for women.

2. Measures

2.1. Mode of Administration

There were some differences in mode of administration between samples. The HOS version was administered by mail or by phone; the VA version was administered by only mail. The VA version was administered in both English and Spanish versions; the HOS only in English.

The timing of survey administration also differed; the VA survey was fielded July 1999 to January 2000; in the HOS, Cohort 1 was fielded May 1998 (and Cohort 2 in March 1999, Cohort 3 in April 2000). It is also important to note that the VA survey was a one-time cross-sectional survey; in the HOS, participants are surveyed again 2 years later (if they are still enrolled in the same health plan).

2.2. SF-36

As the SF-36 is well documented in the literature, we have not detailed the content of the MOS SF-36 and the Veterans SF-36 other than to highlight the differences between the two versions. There are several important differences in the versions of the SF-36 used in the HOS (Appendix I) and the VA (Appendix II). In particular, the VA administered the (Veterans SF-36; Kazis et al., 1998) which differs from the HOS SF-36 (NCQA, 1998) in two respects. The Veterans SF-36 uses 5-point response choices for the 4 RP (role limitations due to physical problems), and 3 RE (role limitations due to emotional problems) items from 'no, none of the time to yes, all of the time', and has two (rather than 1) health transition items, one for physical and one for emotional health. (The latter difference is not relevant for the present comparison, because the health transition items are not used in scoring the SF-36).

The other differences are in the order of presentation of items between the two studies. In both, the SF-36 items were the first ones presented; however, the order of items within survey measures differed between the studies. In the HOS, the form was in the standard SF-36 order (general health (GH1), health transition item (HTran), physical functioning PF (1-10), role limitations due to physical problems RP (1-4) with dichotomized yes/no choices, Role limitations due to emotional problems RE (1-3) with dichotomized yes/no choices, Social functioning (SF1), Bodily pain (BP1-2), Vitality VT (1-4), MH (1-5), Social Functioning (SF2), general Health (GH2-5), *while the order in the VA form was slightly different* (GH1, PF (1-10), RP (1-4)-5 pt choices, RE (1-3) -5pt choices , SF1, BP (1-2), VT (1-4), MH (1-5), SF2, GH2-GH5, physical transition item and emotional transition item).[1] The reader is referred to appendix I and II for the exact wording of items, response choices and ordering of items.

2.3. Demographics

There were differences between the two studies in the source of various demographic variables. In the HOS, age, race and gender were obtained from administrative data and by self-report (we used the administrative data in our analyses), and education by self-report. In the VA, age and gender were obtained from administrative data, and race and education were obtained by self-report. Because of some concerns about the number of VA respondents 99 years and older (perhaps because missing birth date was interpreted as 1/1/00), we truncated age at 98.

2.4. Comorbidities

The HOS and VA samples differed in the nature and extent of disease variables. The HOS measure inquired whether a doctor had ever told the respondent they had any of 13 conditions, all medical. The VA asked whether a doctor had ever told the respondent they had any of 15

[1] Items on different pages of the form are separated by a slash, "/".

conditions, including 3 mental health conditions (depression, PTSD, schizophrenia). In addition, using the VA electronic patient record, administrative data on inpatient stays (over the previous 10 years) and outpatient visits (over previous 3 years) were examined, ICD-9-CM codes were obtained, and combined to define 30 medical and 6 mental health conditions.

3. Analytic Method

To compare the two versions of the SF-36 between HOS and VA, we used confirmatory factor analysis to test a sequence of hypothesized models, differing in the nature of constraints across groups. Analyses were conducted separately for men and women, in each case comparing between the HOS and VA samples.

Below, we present some notation useful for describing our approach, as well as a conceptually oriented discussion of what we attempted to do to test equivalence using multiple-group confirmatory factor analysis. The technical details of the estimation are then presented.

3.1. Conceptual issues and notation

The goal of factor analysis is to take a set of observed scores and construct a parsimonious model of the relations among these scores. This is done by proposing one or more latent constructs (i.e., factors) that account for the covariation among the observed scores. There are fewer latent constructs than observed variables, often many fewer. In the case of the SF-36, we propose 8 latent constructs based upon the prior history of use of this assessment to account for the covariation among the 35 scored items, the transition item(s) is not included in this analysis.

Many people are familiar with exploratory factor analysis, in which the data largely determine the model; in the case of confirmatory factor analysis, a model is proposed a priori (as a pattern of loadings of observed variables on latent factors) and the fit of the model to the data is tested using a chi-square statistic (large values indicate poor fit of the model to the data).

In the factor analysis model, a score on an observed item (x) is a function of one or more latent constructs,

$$x = \tau + \Lambda \xi + \delta,$$

where x is a px1 vector of observed variables, ξ a mx1 vector of latent constructs, τ a px1 vector of item intercepts, δ a px1 vector of residuals, and Λ a pxm matrix of factor loadings. That is, the score on observed variable x is modeled as in a regression model as a weighted function of latent variables ξ (where the weights are in matrix Λ), plus an intercept (τ) and a residual δ. As in regression, a one unit change in ξ leads to a change of Λ in x.

Some additional notation:
p = number of items
m = number of latent variables
S = covariance matrix of observed variables
Σ = covariance matrix implied by the factor analytic model
Φ = var(ξ), covariance matrix of latent variables
Θ = Cov(δ), covariance matrix of residuals (generally diagonal), and

$\mu = \tau + \Lambda \kappa$, where κ is mx1 vector of latent means (of the ξ) and μ is px1 vector of item means

The covariance matrix implied by the latent constructs can then be represented as a function of the latent variables,

$$\Sigma = \Lambda \Phi \Lambda' + \Theta.$$

Statistical tests are conducted by comparing the estimated covariance matrix Σ implied by the hypothesized model to the observed covariance matrix S, using one of several methods to estimate *S* such as generalized least squares or maximum likelihood. In the latter case, a measure (*F*) of discrepancy between Σ and *S* is calculated as $\log ||\Sigma|| + \text{trace}(S\Sigma^{-1}) - \log ||S|| - p$, where p = number of variables. When multiplied by N-1, *F* is distributed as a chi-square with degrees of freedom equal to $(p*(p+1)/2) - t$, where *t* is the number of parameters estimated. This chi-square can be used to determine how well the covariance matrix Σ specified by the model fits the covariance matrix of the observed data S.

3.2. Testing Degrees of Equivalence

To test the extent of equivalence in measurement properties of the SF-36 between the HOS and the VA, we tested a sequence of models representing varying degrees of equivalence. The sequence was based primarily on Vandenberg and Lance (2000; cf. also Steenkamp & Baumgartner, 1998). By comparing the fit of successive models, and computing the change in various fit measures, one can test whether a model with more equivalence constraints across groups significantly worsens the fit of the model to the data. Such a worsening of fit would indicate that the previous, less constrained model provides a better fit.

The comparison of successive models can be conducted when one model is 'nested' within another. A given model is nested within another model when the former can be derived from the latter by placing restrictions on parameters within the former (e.g., in the case of comparing degrees of equivalence, imposing a set of equality restrictions), without introducing new parameters to be estimated (Bentler & Bonett, 1980; Chenug & Rensvold, 2002). The new, more restricted model is nested in the more general (less restricted) model, and has fewer degrees of freedom. The difference in chi-square between the two models can be computed, and when tested against the difference in degrees of freedom of the two models, indicates whether the imposition of additional restrictions resulted in a significant worsening of the fit of the model to the data. For example, when a model representing metric equivalence is compared to a model representing configural equivalence, and the difference in chi-square is significant at the difference in df, it indicates that the more restricted model (which imposes equality constraints on factor loadings across groups) significantly worsens the fit of the model to the data; thus one would reject the null hypothesis of no difference between the two models and conclude that configural rather than metric equivalence obtains in the data.

The tests can be divided into two groups, one testing measurement equivalence; the other structural equivalence (e.g., Cheung & Rensvold, 2002; Vandenberg & Lance, 2000). The former group is the prerequisite for conducting the latter, which are generally those of substantive interest. They are presented in order of increasing degree of equivalence.

Measurement Equivalence

- configural equivalence -- same pattern of salient (i.e., non-zero) and fixed zero factor loadings across groups, but the corresponding loadings are not constrained to equality across groups. If accepted, this hypothesis indicates that items define the same scales, but that the weights of items on scales can differ among groups. *More simply, accepting this hypothesis implies that the various groups associate the same items with the same constructs.*

- metric equivalence – the values of corresponding non-zero factor loadings are constrained to be identical among groups. This hypothesis indicates that the weights of items on scales are identical across samples. (NB: in this study, because response choices differ for the RP and RE scales between HOS and VA studies, the various parameters for these two factors are NOT necessarily constrained to be identical in this or subsequent tests). *Accepting the hypothesis of metric equivalence implies that the strength of the relationship between each item and its underlying construct is identical across groups.*

- scalar intercept equivalence -- item intercepts are identical, $\tau_i = \tau_j$. If accepted, this hypothesis indicates that differences in the means of observed variables are due to differences in the latent construct means, or that the items do not have different intercepts in the different groups.

- equivalence of unique variances – residual item variances are identical, $\theta_i = \theta_j$. If this hypothesis is accepted, it indicates that unique variances are identical across groups (i.e., the items are equally reliable).

If the model of metric equivalence (at a minimum) is accepted, and if intercept equivalence is also accepted, then we can conclude that our measure (i.e., the SF-36) is invariant across samples. That is, the acceptance of metric and scalar equivalence are prerequisite to testing group differences in latent constructs represented by the SF-36 scales. (Note that, in the present situation, due to the differences in response choices between the HOS and the VA versions of the RP and RE scales, we do not expect to retain the hypothesis of intercept equivalence between samples for these scales. However, it is possible that the hypothesis of partial intercept equivalence can be retained for the remaining 6 SF-36 scales).

Structural Equivalence

- equivalence of factor variances, $\varphi_{ii} = \varphi_{jj}$. This hypothesis tests whether the variances of latent constructs are equal over groups. If accepted, it indicates that the variances of the latent constructs are identical.

- equivalence of factor covariances, $\varphi_{ij} = \varphi_{kl}$. If accepted, this hypothesis, in combination with the previous, indicates that correlations among constructs are equal across groups. This is a very strong hypothesis, unlikely to be retained even among random samples from the same population (e.g., Meredith & Horn, 2001).

- equivalence of factor means $\kappa_i = \kappa_j$. This hypothesis tests whether the latent construct means are equal across the groups.

Note also within this sequence of tests, the model of partial measurement equivalence can also be a hypothesis of interest, e.g., that some but not all parameters are invariant across groups (Byrne et al., 1989; Steenkamp & Baumgartner, 1998). In the present case, given the differences in response formats for role items between the HOS SF36 and the Veterans SF36 (2 point vs. 5 point, respectively), this might be an appropriate context for considering partial equivalence, i.e., that the remaining 6 factors are invariant, but the 2 role scales vary across groups.

The Table below indicates the sequence of tests as conducted in LISREL, where an "X" indicates that the matrix (or, in the case of partial equivalence, some of its parameters) is constrained between groups.

	Factor Parameter Matrix Constrained *				
Model of equivalence	**Factor pattern (LX)**	**Item intercepts (TX)**	**Unique variances (TD)**	**Factor variance-covariance matrix (PH)**	**Factor means (KA)**
Configural					
Metric	X				
Scalar	X	X			
Unique	X	X	X		
Factor variance	X	X	X	X (variances only)	
Factor covariance	X	X	X	X	
Factor mean	X	X	X	X	X

* Note. The abbreviation in parentheses refers to the matrix as specified in the LISREL program, e.g., LX is the LISREL abbreviation for the factor pattern matrix.

3.3. Estimation

We used the LISREL program (Joreskog & Sorbom, 1996), version 8.52, and conducted maximum likelihood estimation on covariance matrices.

3.4. Identification

Identification constraints are necessary for estimation of a confirmatory factor model in even a single group; a common set of constraints we adopted was to fix one arbitrarily selected loading on each factor to 1.0 to provide a scale for the factor. Also, each item was permitted to load only one factor, that specified by the standard SF-36 scoring algorithm. For example, all 10 PF items were allowed to load only the physical functioning factor; the loadings of these items on other factors were fixed at zero.

In multiple group factor analysis, it is also necessary to impose constraints between groups, especially when means are included in the analysis. Thus, in one group(HOS), the vector of factor means, κ, was set to 0; values in the other group (VA) are relative to this.

3.5. Model Fit

Because the chi-square test of fit is affected by sample size N (i.e., $\chi^2 = (N-1)F$, where F is the minimum fit function), and because the sample size was so exceptionally large in these analyses, we relied primarily on other goodness of fit indices which are (much) less sensitive to

sample size. Fit tests are used to decide on the adequacy with which a given model fits the data. Most agree that the chi-square test is not particularly useful in assessing model fit, and that other indices should be used.

Fit indices can be divided into two classes; absolute and incremental (Hu & Bentler, 1999). The former assess the adequacy with which the model reproduces the data; the latter how well a given model improves on a more restricted, nested baseline model. Absolute indices assess the degree of discrepancy between the covariance matrix implied by the model (Σ) and that observed from the data (S), and should be small, i.e., less than 0.08. Incremental indices are similar to variance explained measures such as R^2 and should be larger than .95.

A good rule of thumb is to consider at least one index of each kind in evaluating the fit of a model to the data. Absolute indices include the root mean square error of approximation (RMSEA; Steiger & Lind, 1980) and standardized root mean squared residual (SRMR; Hu & Bentler, 1999); incremental indices include the comparative fit index (CFI; Bentler 1990), Tucker-Lewis (1973) index (TLI; also known as the non-normed fit index, NNFI). Absolute fit indices should be small, indicating little discrepancy between the model and the data. For RMSEA and SRMR, values less than 0.05 indicate good fit of the model to the data, and values between 0.05 and 0.08 acceptable fit. For incremental indices, large values are preferred, indicating that the given model accounts for much of the variation in the data. Thus, values greater than 0.90 indicate reasonable fit, and values greater than 0.95 indicate excellent fit (Cheung & Rensvold, 2002; Hu & Bentler, 1999). Recently, Hu and Bentler (1999) proposed a two-index decision rule for assessing fit, based on a series of simulation studies. For maximum likelihood estimation, they recommend that the standardized root mean squared residual (SRMR) be less than 0.09, and either root mean squared error of approximation less than 0.06 or non-normed fit index or comparative fit index 0.95 or larger.

In the context of assessing measurement equivalence, fit tests are also used to compare different models (if they are nested within each other), to determine whether a given model results in an improvement or a worsening of fit relative to a more or less constrained model. In particular, the difference in χ^2 between nested models (e.g., $\Delta\chi^2_{21} = \chi^2_2 - \chi^2_1$) is itself distributed as a χ^2. When tested against the difference in degrees of freedom ($\Delta df = df_2 - df_1$) of two nested models, it can be used to determine whether the more constrained model improves or worsens the fit of the model to the data.

Because the differences in chi-square are affected by sample size, Cheung and Rensvold (2002) suggested an examination of the change in measures of fit other than χ^2, such as CFI. Based on a simulation study involving 20 different fit indices, they recommend that a value of $\Delta CFI \leq -0.01$ "indicates that the null hypothesis of equivalence should not be rejected" (p. 251). Widaman and Riese (1997) suggested considering the ratio of $\Delta\chi^2 / \Delta df$ to assess whether change in fit between models; if large, this indicates that the given set of constraints leads to a relatively large increase in the amount of misfit per degree of freedom. In other words, the constraints being imposed are worsening the fit of the model, relative to the degrees of freedom.

4. Descriptive Findings

Here we describe some of the descriptive findings and comparisons between the HOS and VA data. For reasons described above and elaborated below, all results are presented separately by gender within study, resulting in 4 groups (men and women in HOS, men and women in VA).

4.1. Age Distribution

We first examined the age distribution within each sample (HOS, VA), stratified by gender (for reasons noted above). **Figure 1** (age distrib all.xls) presents the age distribution of the entire sample from each survey (including those under age 65).

In the HOS, men and women show similar distributions, with the peak around 67, and very low prevalence below age 65. In the VA, men show a trimodal age distribution, with peaks around ages 51, 68, 75, representing cohorts of Vietnam, Korea, and WWII veterans, respectively. Women in the VA show a bimodal distribution, with peaks at 44 and 78, representing Vietnam and WWII eras of service, respectively. Because of selection into the VA for health care services, age is intermingled with period of service. Note that in the VA, some respondents were aged 18 to 64.

In **Figure 2** (age distrib.xls), age distributions are shown for respondents $\geq$ 65 years. In the HOS, men and women show similar distribution, with peaks near 67, and long tails. In the VA, men have peaks at 67 and 75; whereas for VA women, there is a very large peak around age 77. All subsequent analyses are based only on respondents aged 65 and older.

4.2. Demographics

Table 1 (demographics.xls) presents selected demographics by study. Note, as mentioned above, that some of the demographic variables are based on respondent self-report, while others are obtained from administrative data.

For age **(Table 1 and figure 3)**, the means are similar between studies; however, the standard deviation is larger in the HOS and the maximum age is about 6-10 years older (note that age was truncated at 98 in the VA as the validity of ages above 98 was questioned). In each study, over 75% of respondents by gender were aged 65-79 (HOS 83.5% males, 79.4% females, VA 82.0% males, 77.0%, females respectively).

For race (**Figure 4**), the VA includes more non-white men, and fewer Black women, than the HOS; a larger proportion of VA enrollees are missing race (4% vs. <1% for HOS). Note that race was obtained from administrative data in the HOS, and by self-report in the VA.

For education (**Figure 5),** which was obtained by self-report in both studies, there was more missing data in the VA (10%) than in the HOS (2%). There were more low educated men in the VA than in HOS (20% vs. 13% with $\leq$8 years) and somewhat fewer highly educated (31% vs. 40% with > 12 yr). Women in the HOS were less highly educated than were women in the VA, perhaps because large numbers of female VA enrollees were nurses.

4.3. Comorbidities

Generally there were higher prevalence's of self-reported conditions in the VA than in the HOS; the only exception was that depression was higher for women only in HOS than in VA (30% vs. 25%), see **Figure 6**.

4.4. SF-36 Item statistics

Table 2 (valid items.xls) presents the distribution of valid SF-36 items by group. There was a higher percent of respondents omitting all 35 items in the VA (2.6%) than in the HOS (0.07%); a higher percent missing 1 to 3 items (0.4%-1% in VA, vs. 0.1% in HOS); and a lower percent with all 35 items complete (about 60% vs. 75%-80% for HOS). In HOS, men were more likely to complete all items than were women (80% vs. 74%); in VA, the difference with all 35 items complete was lower than in HOS, but more similar between men (61%) and women (58%).

Figure 7 (means.xls; chart missing data items) presents, for each item, the percent of missing data separately for the 4 groups (men and women within each study). In the VA, 10-12% of respondents by gender were missing responses on the GH1 and SF2 items, perhaps because of the layout of the SF-36 items in the VA survey (see above and appendices I and II). In general, the rates of missing item responses were higher in the VA than in the HOS.

Among men in the VA, the most common missing data patterns were (a) omission of GH1 (4.98% of all respondents), (b) omission of SF2 (3.07%), or (c) omission of all 35 items (2.68%). Remaining missing data patterns were each by less than 1% of respondents. For women in the VA, these same three patterns were most common, accounting for 4.96%, 3.98%, and 2.63% of responses. Among men in the HOS, only 1 pattern had more than 1% of respondents -- omitting 20 items and answering only 10 PF + 5 GH (1.07%). Among HOS women, only 1 pattern had more than 1% of respondents -- omitting item PF1 (1.37%). These results are important to consider in the development of future strategies for imputation methods.

Figure 8 (means.xls, chart means) plots item responses for the 4 groups. The item mean profiles are generally similar among the 4 groups, with the exception due to the use of 5-point responses for role scales in the VA. In general, item means for HOS respondents are higher than for VA (especially for PF).

Appendix III (Item distrib.xls) presents frequency distributions by study and gender for the SF-36 items. Also shown are order of items in the survey and the percent missing data.

5. Measurement Equivalence

We tested, separately for men and women, measurement equivalence between HOS and VA.. In these analyses, we computed 4 covariance matrices (men and women, each for VA and HOS Cohort 1), on persons aged 65 and older (**Appendix IV**). Listwise deletion was used (i.e., cases with any of the 35 SF-36 items missing were deleted). Scores on some items were reversed, so that all are in the same direction, with higher scores indicating better health.

Table 3 shows the sample sizes, by study and gender, for those with complete data on the SF-36, the total sample, and the percent of those from the total sample with complete data. Note that the percent of women with complete data is lower than that of men, and that more respondents in the HOS had complete data than in the VA, but that, overall, more women than men had complete data (72.6% vs. 63.7%).

Table 3. Sample Size by Study and Gender

	Complete SF-36 data		Total Sample		% with complete data	
	Men	**Women**	**Men**	**Women**	**Men**	**Women**
HOS Cohort 1	56,799	70,746	70,610	95,494	80.4	74.1
VA	272,171	5,398	445,816	9,329	61.0	57.9
Total	328,970	76,144	516,426	104,823	63.7	72.6

As previously noted, these very large sample sizes pose a problem with respect to assessing the fit of models to data, given their excess power to reject the null hypothesis of good fit of the model to the data. That is, since chi-square is defined as (N-1)*F, where F is a goodness of fit function, the very large sample sizes will greatly inflate the value. Thus, it was important to rely on fit indices other than chi-square to decide whether or not to accept a given model as adequately fitting the data.

To the best of our knowledge, the largest sample size in a published study of health outcomes was studied by Marshall et al. (2001), who had a sample of approximately 15,000 persons who completed the CAHPS. However, in this study, the cohort was divided into 4 groups ranging in size from 609 to 7983. In simulation studies of factor analysis (e.g., Hu & Bentler, 1999), the largest sample size considered was 5000. Our analysis for men, with over 300,000 respondents having complete data, is by far the largest sample ever subjected to factor analysis in any domain, to our knowledge.

The initial model of configural equivalence was obtained by specifying that each item of the SF-36 loaded on only one scale, according to standard scoring algorithms (Ware et al., 1993; Kazis et al., 1998). That is, all 10 PF items loaded on the physical functioning factor, the 4 RP items on the role physical factor, etc. For each factor, the loading for one item was fixed at 1.0, to provide a scale for the factor. The items were PF2 (limited in moderate activities), RP1 (cut down on amount of time), BP1 (amount of pain), GH1 (is your health excellent, very good, etc.), VT2 (have a lot of energy), SF2 (how much of the time did health interfere with social activities), RE1 (cut down on amount of time), and MH5 (have you been happy). Unique variances, item intercepts, and factor variances and covariances were freely estimated. Example LISREL scripts for specifying models of configural equivalence and of intercept equivalence are given in **Appendix V**[2].

We did not consider alternative hypotheses about numbers of factors or loadings on items on different factors, nor did we attempt to improve the fit of the initial model of configural equivalence to the data by allowing items to load on more than one factor (or, for that matter, on

[2] To adapt these programs for women, changes would need to be made in the titles, the NO= parameter on the DA line, and in the file name on the 3 lines describing the input data, for both HOS and VA portions of the programs.

a different factor than specified by the standard scoring algorithm). Prior work has provided a strong foundation for this a priori model (Keller et al., 1998; Ware et al., 1993).

Successive models of increasing equivalence between groups were then tested, following the sequence specified in Section 3.2 above. For each model considered (detailed results are shown in **Appendix VI**), various measures of fit are shown in Table 4, first for men and then for women. Chi-square and df are presented; no *p* values are given because all are highly significant, indicating that the various models do not fit the data well according to this very powerful test. The minimum fit function (F) is provided, followed by two measures of absolute fit (RMSEA, SRMR) and two measures of relative fit (CFI, NNFI).

Because, as noted above, the chi-square is influenced by sample size (i.e., chi-square = (N-1)*F), we rely on the other indices to determine the adequacy of the fit of the models to the data. We consider the values of these indices in relation to the cutoffs recommended by Hu and Bentler (1999), who suggest minimum values near 0.95 for incremental fit indices, and for the absolute indices, cutoffs of 0.08 for SRMR and 0.06 for RMSEA. We also considered the change in chi-square relative to the change in df ($\Delta\chi^2$ / Δdf) as advocated by Widaman and Reise (1997).

We discuss, in some detail, the fit of the configural equivalence model for men (shown in the first row of Table 4); the description applies to the remaining models and to those for women, shown in the lower half of Table 4. We then discuss the differences between successive models of increasingly stricter equivalence, presented in Table 5.

For men, the fit of Model 1 (Table 4a, row 1), specifying configural equivalence, was highly significant, with a chi-square of 860,283.405 on 1064 degrees of freedom. The minimum of the fit function is 2.615, and, as noted above, is multiplied by N-1 to obtain the chi-square. The absolute fit indices convey somewhat different results regarding the adequacy with which the model fits the data; the RMSEA of .0789 is larger than the cutoff of 0.06 recommended by Hu and Bentler (1999). (Note that others, e.g., Steiger & Lind, 1980 recommend a cutoff of 0.08). However, the SRMR of .0433 is lower than the 0.08 cutoff. Further, the two indices of incremental fit, CFI and NNFI, both exceed the recommended cutoff of 0.95. *Based on these results, our conclusion is that the configural equivalence model provides a reasonably good fit to the data, suggesting that for men the same items are related to the same factors in both HOS and VA. A similar finding holds for women (Table 4b, row 1).*

In subsequent rows of Table 4, tests are shown for different models of equivalence. After considering configural equivalence, we tested (Model 2) partial metric equivalence, forcing equality constraints on factor loadings for all items except RP and RE (because of the different response scales used for these items in the HOS and VA). Subsequent models in Table 4 add additional equivalence constraints to the model of partial metric equivalence; in these subsequent models, equality constraints were not imposed on RP and RE items.

The next model examined, Model 3, of partial intercept equivalence, constrained intercepts across HOS and VA for all items except RP and RE. The fit of this model to the data was within the recommended cutoffs for all 4 fit indices.

Model 4, which specifies partial equality of unique variances across samples, provides an adequate fit to the data for both men and women. The remaining models, 5-7, test various aspects of structural rather than measurement equivalence (as defined in Section 3.2 above). In model 7, we estimated factor means, and allowed them to differ between HOS and VA samples on 6 of 8 factors (all but RP and RE). This model used fewer degrees of freedom than Model 6; thus in the comparisons of model given in Table 5, the test is whether relaxing the equivalence constraints for the 6 factor means results in an improvement in fit.

Table 4. Fit of Equivalence Models across HOS Cohort 1 and VA by Gender

4a. MEN				**Absolute Fit**		**Relative Fit**	
Model of equivalence	**Chi-square**	**df**	**min F**	**RMSEA**	**SRMR**	**CFI**	**NNFI**
1. Configural	860,283.405	1064	2.615	0.0789	0.0433	0.982	0.980
2. Partial metric	868,338.265	1086	2.640	0.0786	0.0449	0.982	0.980
3. Partial intercept	904,137.142	1114	2.748	0.0788	0.0453	0.981	0.979
4. Partial unique variance	936,731.281	1142	2.847	0.0790	0.0457	0.980	0.979
5. Partial factor variance	942,617.650	1148	2.865	0.0793	0.0476	0.980	0.979
6. Partial factor covariance	946,726.677	1163	2.878	0.0789	0.0468	0.980	0.979
7. Partial factor mean	917,990.812	1157	2.791	0.0780	0.0540	0.980	0.981

4b. WOMEN				**Absolute Fit**		**Relative Fit**	
Model of equivalence	**Chi-square**	**df**	**Min F**	**RMSEA**	**SRMR**	**CFI**	**NNFI**
1. Configural	175,170.039	1064	2.301	0.0728	0.0425	0.980	0.978
2. Partial metric	175,798.056	1086	2.309	0.0722	0.0571	0.980	0.978
3. Partial intercept	177,511.311	1114	2.331	0.0715	0.0528	0.980	0.979
4. Partial unique variance	178,511.451	1142	2.344	0.0708	0.0566	0.980	0.979
5. Partial factor variance	179,058.147	1148	2.352	0.0709	0.1790	0.980	0.979
6. Partial factor covariance	179,437.066	1163	2.357	0.0705	0.1600	0.980	0.979
7. Partial factor mean	177,942.662	1157	2.337	0.0704	0.0936	0.980	0.979

Note. Fit = minimum fit function; RMSEA = root mean squared error of approximation; SRMR = standardized root mean squared residual; CFI = comparative fit index; NNFI = non-normed fit index.

There are some interesting differences in fit of corresponding models to data between men and women, in part because the women's sample is about 25% the size of the men's sample. Thus, the chi-square values are lower. In addition, note that the minimum value of the fit function is generally smaller for women (about 2.3) vs. that for men, about 2.6 to 2.8. This larger value for men, combined with the much larger sample size, leads to substantially larger chi-square values for men. The measures of absolute fit, RMSEA and SRMR, are generally smaller for women (with the exception of SRMR for Models 5 and 6, testing equality of factor variances and covariances, respectively). The relative fit indices, CFI and NNFI, were quite similar between men and women. Overall, all values of RMSEA were larger than the 0.06 cutoff recommended by Hu and Bentler (1999), although the values of SRMR were all smaller than the cutoff of 0.08 (except for Models 5 and 6 for women). Values of both indices of confirmatory fit exceeded the 0.95 minimum recommended by Hu and Bentler (1999).

Examining the fit of the equivalence models in Table 4 is revealing, but the key results are those in Table 5, which compare the fit of successive pairs of models, to determine at what point in the model-fitting sequence the imposition of equality constraints across HOS and VA samples worsens the fit of the model to the data. As noted above, a key statistic for evaluating the impact of imposing additional equality constraints across samples is the change in chi-square. However, as also noted above, the chi-square values are influenced by sample size, and in the present case, sample sizes are quite large. Thus, it becomes important to consider changes in other measures of fit, such as those shown in Table 5.

In considering the results of model comparisons shown in Table 5, we adopted the following strategy. First, we examined the difference in chi-square between successive models, which is itself a chi-square, and can be tested against the difference in degrees of freedom. When this is significant, it indicates that including additional equality constraints across HOS and VA samples worsened the fit of the model to the data. Note that, because chi-square is itself influenced by sample size, which in the present case, was quite large, these changes in chi-square are also affected by sample size. Thus, we also considered changes in other measures of fit between successive models, to determine at what point in the sequence of tests they indicate a worsening in fit.

Table 5. Differences in Fit of Successive Models of Measurement Equivalence, by Gender

5a. MEN

Models Compared	$\Delta \chi^2$	Δ df	$\Delta\chi^2/\Delta$ df	Δ RMSEA	ΔSRMR	ΔCFI	ΔNNFI
21. Partial metric vs. configural	8,054.86	22	366.130	-0.0003	0.0016	0.0000	0.0000
32. Partial intercept vs. partial metric	35,798.88	28	1,278.531	0.0002	0.0004	-0.0010	-0.0010
43. Unique vs. intercept	32,594.14	28	1,164.076	0.0002	0.0004	-0.0010	0.0000
54. Factor variance vs. unique	5,886.37	6	981.062	0.0003	0.0019	0.0000	0.0000
65. Factor covariance vs. factor variance	4,109.03	15	273.935	-0.0004	-0.0008	0.0000	0.0000
76. Factor means vs. factor covariance	28,735.86	6	4,789.311	0.0009	-0.0072	0.0000	-0.0020

5b. WOMEN

Models Compared	$\Delta \chi^2$	Δ df	$\Delta\chi^2/\Delta$ df	Δ RMSEA	ΔSRMR	ΔCFI	ΔNNFI
21. Partial metric vs. configural	628.02	22	28.55	-0.0006	0.0146	0.0000	0.0000
32. Partial intercept vs. partial metric	1,713.25	28	61.19	-0.0007	-0.0043	0.0000	0.0010
43. Unique vs. intercept	1,000.14	28	35.72	-0.0007	0.0038	0.0000	0.0000
54. Factor variance vs. unique	546.70	6	91.12	0.0001	0.1224	0.0000	0.0000
65. Factor covariance vs. factor variance	378.92	15	25.26	-0.0004	-0.0190	0.0000	0.0000
76. Factor means vs.	1,494.40	6	249.07	-0.0001	-0.0664	0.0000	0.0000

factor covariance							

For men, the first row of Table 5 (comparison 21) suggests that the fit of Model 2, partial metric equivalence, is not much worse than that of Model 1, configural equivalence. Although the $\Delta\chi^2$ suggests a worsening of fit, the influence of sample size is less in the other indices, all of which suggest a relatively small impact of imposing equality constraints on factor loadings for 6 of 8 factors (omitting RP and RE) across HOS and VA. Importantly, the $\Delta\chi^2$ / Δdf ratio is relatively small.

The next comparison, 32, tested the impact of imposing equality constraints on intercepts (for 28 of 35 items, omitting 4 for RP and 3 for RE). The change in chi-square was very large, over 35,000, and the $\Delta\chi^2$ / Δdf ratio was quite large (> 1200), suggesting that this set of constraints worsened the fit of the model to the data. The model adding equivalence of unique variances (Model 4) also showed a notable worsening of fit, compared to Model 3 (Comparison 43). Compared to the model including equivalence of unique intercepts, the model constraining factor variance across samples (Model 5) was only somewhat worse (Comparison 54); the model adding equality constraints on factor covariances was only slightly worse. Finally, Model 7, which allowed factor means to vary between sample (except for RP and RE) was a significant improvement (Comparison 76). However, because Model 3 was significantly worse than Model 2, we selected Model 2 as providing the best fit to the data for men.

Based on these results, we conclude that the best-fitting model for the men is Model 2, which imposed equality constraints on items for 6 of 8 factors (all except RP and RE, due to differences in response formats). In other words, factor loadings for 6 of 8 factors are equivalent between men in Medicare HMO's who completed the HOS version of the SF-36 and men receiving VA care who completed the Veterans SF-36. Other parameter matrices (e.g., intercepts, unique variances, factor covariances, factor means) were not equivalent across HOS and men.

For women, the results also suggest that the model of partial metric equivalence provides the best fit to the data; the improvement of this model over the configural equivalence model was reasonable (Comparison 21). The partial intercept equivalence model did not result in much improvement (Comparison 32), nor did the unique equivalence model (Comparison 43). The model of invariant factor variances was not notably worse than that of unique equivalence (Comparison 54), and the model adding equivalence of factor covariances also seemed to provide a reasonable fit (Comparison 65). The model allowing differences in 6 of 8 factor means (all but RP and RE) was a significant improvement. However, given that Model 3 provided a worse fit than Model 2, Model 2 was accepted as providing the best fit to the data.

Given the acceptance of Model 2, partial metric equivalence between HOS and VA, for both men and women, we conducted a second set of equivalence analyses. In these, we forced equivalence of <u>all</u> parameters between HOS and VA, including the RP and RE items. Table 6 presents results for several models, separately for men and for women. As the starting point, the model of configural equivalence described above is given; the tests of differences in models (Table 7) use this configural equivalence model as a baseline.

Table 6. Fit of Full Equivalence Models across HOS Cohort 1 and VA by Gender

6a. MEN				**Absolute Fit**		**Relative Fit**	
Model of equivalence	**Chi-square**	**df**	**min F**	**RMSEA**	**SRMR**	**CFI**	**NNFI**
1. Configural	860,283.40	1064	2.615	0.0789	0.0433	0.982	0.980
2a. Full metric	870,395.56	1091	2.646	0.0786	0.0450	0.982	0.980
3a. Full intercept	1,138,138.36	1126	3.460	0.0806	0.0460	0.976	0.974

6b. WOMEN				**Absolute Fit**		**Relative Fit**	
Model of equivalence	**Chi-square**	**df**	**min F**	**RMSEA**	**SRMR**	**CFI**	**NNFI**
1. Configural	175,170.04	1064	2.301	0.0728	0.0425	0.980	0.978
2a. Full metric	175,941.52	1091	2.311	0.0721	0.0585	0.980	0.978
3a. Full intercept	194,317.96	1126	2.552	0.0727	0.0559	0.978	0.977

For men (Table 6a), Model 2a specified that <u>all</u> factor loadings (including RP and RE items) were equivalent between groups; the fit of this model to the data was acceptable. Model 3a, which added the set of constraints specifying equivalence of item intercepts across HOS and VA, did not fit the data well; the chi-square was very large, as was the minimum fit function. *Similarly, for women (Table 6b), the full metric model fit reasonably well*. The model of full intercept invariance did not provide a good fit to the data.

Turning to Table 7, we conducted difference tests between these three models for men and for women. For men (Table 7a), the model of full metric equivalence l was not much worse than that of configural invariance (Comparison 2a-1). A comparison of the models of full and partial metric equivalence (Comparison 2a-2) also did not suggest a worsening of fit for the model of full metric equivalence; however, the model of full intercept equivalence was notably worse than that of full metric equivalence (Comparison 3a-2a). *Thus, Model 2a was accepted as providing the best fit to the data for men,* and no further model comparisons were conducted.

For women (Table 7b), the model of full metric equivalence model was not much worse than that of configural invariance (Comparison 2a-1). A comparison of the models of full and partial metric equivalence (Comparison 2a-2) also did not suggest a worsening of fit for the model of full metric equivalence. The model of full intercept equivalence was notably worse than that of full metric equivalence (Comparison 3a-2a); *thus, Model 2a was accepted as providing the best fit to the data for women,* and no further model comparisons were conducted.

Table 7. Differences in Fit of Successive Models of Full Measurement Equivalence, by Gender

7a. MEN **Models Compared**	**$\Delta\chi^2$**	**Δ df**	**$\Delta\chi^2/\Delta$ df**	**Δ RMSEA**	**ΔSRMR**	**ΔCFI**	**ΔNNFI**
2a 1. Full metric vs. configural	10,112.15	27	374.52	-0.0003	0.0017	0.0000	0.0000
2a 2. Full metric vs. partial metric	2,057.29	5	411.46	0.0000	0.0001	0.0000	0.0000
3a 2a. Full intercept vs. full metric	267,742.80	35	7,649.79	0.0020	0.0010	-0.0060	-0.0060

7b. WOMEN

Models Compared	$\Delta \chi^2$	Δ df	$\Delta\chi^2/\Delta$ df	Δ RMSEA	ΔSRMR	ΔCFI	ΔNNFI
2a 1. Full metric vs. configural	771.48	27	28.57	-0.0007	0.0160	0.0000	0.0000
2a 2. Full metric vs. partial metric	143.46	5	28.69	-0.0001	0.0014	0.0000	0.0000
3a 2a. Full intercept vs. full metric	18,376.44	35	525.04	0.0006	-0.0026	-0.0020	-0.0010

The final models accepted for men and women, of full metric equivalence, are presented in Tables 8 and 9, respectively. Factor loadings that were fixed at 1.0 to provide a scale for the factor are shown in **bold**; entries for RP and RE are in *italics*, to distinguish these items which used different response formats in the two studies. Factor variances are given in **bold**; factor covariances are below the main diagonal and factor correlations are above.

In sum, the results of the equivalence analyses indicate that the factor structure of the SF-36 was equivalent between the HOS and VA samples, for both men and women. Despite differences in the response formats of the RP and RE items between the samples, the contributions of each item to its hypothesized scale was identical in both samples. Other components of the factor analysis model, such as item intercepts, factor covariances, and unique variances, were not identical between samples. However, given that the minimum requirement of metric equivalence was met, these results imply that meaningful comparisons can be conducted between HOS and VA samples using the SF-36 scales.

Table 8. Full Metric Equivalence Model, MEN

			Unique Variances		Item Intercepts	
Item Label	**Item Description**	Factor Loadings	HOS	VA	HOS	V A
PF1	Vigorous	0.577	0.364	0.256	1.721	1.401
PF2	Moderate	**1.000**	0.168	0.182	2.400	1.952
PF3	Lift/carry	0.941	0.145	0.201	2.577	2.173
PF4	Climb several	0.985	0.216	0.197	2.212	1.739
PF5	climb one	0.990	0.122	0.168	2.573	2.190
PF6	bend/kneel	0.852	0.257	0.254	2.249	1.858
PF7	Walk mile	1.031	0.255	0.225	2.177	1.723
PF8	Walk several	1.134	0.154	0.169	2.412	1.906
PF9	Walk one	0.949	0.129	0.203	2.653	2.322
PF10	bathe/dress	0.585	0.153	0.278	2.805	2.568
RP1	***limited***	***1.000***	*0.079*	*0.351*	*1.707*	*3.079*
RP2	***do less***	*1.010*	*0.076*	*0.285*	*1.575*	*2.826*
RP3	***kind***	*1.048*	*0.059*	*0.225*	*1.598*	*2.810*
RP4	***difficulty***	*1.060*	*0.060*	*0.214*	*1.606*	*2.800*
BP1	amount	**1.000**	0.643	0.528	4.285	3.531
BP2	interfere	1.086	0.059	0.095	4.018	3.226
GH1	general	**1.000**	0.329	0.283	3.148	2.434
GH2	sick easier	0.879	0.579	0.842	4.276	3.764
GH3	healthy as	1.282	0.674	0.648	3.542	2.773
GH4	health worsen	0.840	0.953	0.952	3.393	2.892
GH5	health excellent	1.418	0.423	0.446	3.365	2.469
VT1	pep	0.973	0.631	0.505	3.554	2.794
VT2	energy	**1.000**	0.579	0.519	3.606	2.772
VT3	worn out	0.842	0.629	0.979	4.510	3.784
VT4	tired	0.820	0.539	0.836	4.129	3.456
SF1	extent interfere	1.107	0.292	0.357	4.257	3.470
SF2	time interfere	**1.000**	0.327	0.441	4.288	3.589
RE1	***limited***	***1.000***	*0.030*	*0.232*	*1.829*	*3.650*
RE2	***do less***	*1.044*	*0.058*	*0.219*	*1.748*	*3.413*
RE3	***not careful***	*0.957*	*0.052*	*0.379*	*1.823*	*3.635*
MH1	nervous	1.079	0.714	1.011	5.166	4.583
MH2	down in dumps	1.155	0.335	0.496	5.483	4.969
`MH3	calm/peaceful	1.125	1.019	0.972	4.343	3.748
MH4	downhearted	1.117	0.396	0.560	5.261	4.755
MH5	happy	**1.000**	0.857	0.968	4.581	4.137

Table 8 (continued): Factor Covariances (and correlations, above diagonal), Men

FULL Metric invariance

HOS	**PF**	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**
PF	**0.343**	0.692	0.631	0.712	0.692	0.700	0.488	0.471
RP	0.160	**0.156**	0.698	0.688	0.712	0.740	0.616	0.474
BP	0.367	0.274	**0.987**	0.642	0.661	0.723	0.502	0.509
GH	0.310	0.202	0.474	**0.552**	0.832	0.738	0.528	0.609
VT	0.468	0.325	0.758	0.714	**1.334**	0.755	0.545	0.668
SF	0.352	0.251	0.617	0.471	0.749	**0.738**	0.694	0.739
RE	0.094	0.080	0.164	0.129	0.207	0.196	**0.108**	0.664
MH	0.196	0.133	0.359	0.321	0.548	0.451	0.155	**0.504**

VA	**PF**	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**
PF	**0.421**	0.777	0.661	0.736	0.715	0.708	0.574	0.481
RP	0.606	**1.446**	0.723	0.764	0.775	0.780	0.701	0.526
BP	0.489	0.991	**1.301**	0.685	0.689	0.762	0.621	0.573
GH	0.389	0.749	0.637	**0.664**	0.845	0.776	0.623	0.627
VT	0.591	1.187	1.001	0.877	**1.623**	0.776	0.630	0.670
SF	0.504	1.029	0.954	0.694	1.085	**1.204**	0.776	0.768
RE	0.483	1.093	0.919	0.659	1.041	1.104	**1.683**	0.711
MH	0.296	0.600	0.620	0.485	0.810	0.800	0.876	**0.901**

Table 9. Full Metric Equivalence Model, WOMEN

			Unique Variances		Item Intercepts	
Item Label	**Item Description**	Factor Loadings	HOS	VA	HOS	V A
PF1	Vigorous	0.643	0.326	0.233	1.607	1.351
PF2	Moderate	**1.000**	0.198	0.191	2.221	1.917
PF3	Lift/carry	0.940	0.187	0.209	2.367	2.107
PF4	Climb several	1.014	0.219	0.194	2.010	1.688
PF5	climb one	0.976	0.158	0.180	2.416	2.175
PF6	bend/kneel	0.886	0.251	0.244	2.129	1.890
PF7	Walk mile	1.072	0.252	0.223	1.990	1.676
PF8	Walk several	1.144	0.179	0.182	2.239	1.952
PF9	Walk one	0.903	0.172	0.213	2.544	2.336
PF10	bathe/dress	0.462	0.191	0.248	2.770	2.656
RP1	***limited***	***1.000***	*0.084*	*0.347*	*1.670*	*3.322*
RP2	***do less***	*1.095*	*0.082*	*0.306*	*1.515*	*2.979*
RP3	***kind***	*1.150*	*0.061*	*0.210*	*1.562*	*3.031*
RP4	***difficulty***	*1.152*	*0.060*	*0.195*	*1.571*	*3.046*
BP1	amount	**1.000**	0.608	0.519	4.042	3.565
BP2	interfere	1.042	0.098	0.101	3.833	3.345
GH1	general	**1.000**	0.303	0.272	3.074	2.699
GH2	sick easier	0.790	0.649	0.783	4.210	4.039
GH3	healthy as	1.127	0.691	0.724	3.593	3.207
GH4	health worsen	0.767	0.908	0.987	3.503	3.204
GH5	health excellent	1.401	0.441	0.475	3.292	2.710
VT1	pep	0.934	0.573	0.441	3.438	2.966
VT2	energy	**1.000**	0.577	0.461	3.436	2.872
VT3	worn out	0.774	0.733	0.928	4.371	3.999
VT4	tired	0.772	0.634	0.786	3.967	3.574
SF1	extent interfere	1.049	0.314	0.308	4.155	3.623
SF2	time interfere	**1.000**	0.341	0.406	4.169	3.717
RE1	***limited***	***1.000***	*0.044*	*0.210*	*1.795*	*3.969*
RE2	***do less***	*1.111*	*0.060*	*0.222*	*1.703*	*3.706*
RE3	***not careful***	*0.962*	*0.057*	*0.301*	*1.789*	*3.937*
MH1	nervous	1.055	0.831	0.793	4.934	4.870
MH2	down in dumps	1.083	0.402	0.412	5.396	5.246
`MH3	calm/peaceful	1.172	0.933	0.910	4.157	3.900
MH4	downhearted	1.120	0.455	0.485	5.084	4.926
MH5	happy	**1.000**	0.836	0.854	4.523	4.347

Table 9 (continued): Factor Covariances (and correlations, above diagonal), Women

Full metric invariance									
HOS		**PF**	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**
	PF	**0.400**	0.712	0.673	0.723	0.711	0.678	0.440	0.430
	RP	0.168	**0.139**	0.732	0.684	0.721	0.736	0.591	0.457
	BP	0.463	0.297	**1.184**	0.686	0.696	0.739	0.496	0.486
	GH	0.355	0.198	0.580	**0.603**	0.814	0.742	0.515	0.603
	VT	0.550	0.329	0.927	0.773	**1.497**	0.767	0.543	0.665
	SF	0.408	0.261	0.765	0.548	0.892	**0.904**	0.671	0.719
	RE	0.096	0.076	0.186	0.138	0.229	0.220	**0.119**	0.673
	MH	0.211	0.132	0.410	0.363	0.631	0.530	0.180	**0.601**
		PF	**RP**	**BP**	**GH**	**VT**	**SF**	**RE**	**MH**
VA	**PF**	**0.414**	0.784	0.687	0.743	0.702	0.678	0.483	0.380
	RP	0.547	**1.176**	0.739	0.762	0.765	0.782	0.629	0.456
	BP	0.514	0.932	**1.353**	0.683	0.674	0.741	0.552	0.466
	GH	0.395	0.682	0.656	**0.682**	0.805	0.761	0.574	0.552
	VT	0.590	1.084	1.025	0.869	**1.708**	0.772	0.595	0.610
	SF	0.489	0.950	0.966	0.704	1.130	**1.256**	0.726	0.677
	RE	0.356	0.781	0.735	0.543	0.891	0.931	**1.311**	0.703
	MH	0.208	0.421	0.462	0.388	0.679	0.646	0.685	**0.725**

6. Discussion and Future Plans

There are a number of avenues of further exploration that could be pursued with these data.

- It would be of interest to determine whether the SF-36 is equivalent across successive HOS cohorts. Given that the samples are drawn from the population of Medicare HMO users (which is a dynamic population), one might expect a greater extent of equivalence than was observed in the present case between HOS and VA
- Examination of equivalence of SF-36 across men and women might be of interest, certainly within the two studies. However, attempts to compare men and women across HOS and VA are likely to indicate a great many differences, given the very different age/gender distributions, and the relatively unique nature of women who use VA healthcare.
- An additional question is whether or not the SF-36 is equivalent by age-group; thus, one could stratify the analyses by age group, e.g., 65-74, 75-84, 85+
- The matrices in the present study were based on cases with complete data (i.e., listwise deletion). However, a number of methods are available for taking missing data into account, e.g., multiple imputation or full-information maximum likelihood.
- Is the SF-36 equivalent in the HOS across mode of administration (mail vs. phone)?

7. References

Agha Z, Lofgren RP, VanRuiswyk JV, Layde PM (2000). Are patients at Veterans Affairs medical centers sicker? A comparative analysis of health status and medical resource use. Archives of Internal Medicine, 160, 3252-57.

Bentler, PM. (1990). Comparative fit indices in structural models. Psychological Bulletin, 107, 238-46.

Byrne, BM, Shavelson, RJ, & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105, 456-466.

Cheung GW & Rensvold RB (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. Structural Equation Modeling, 9, 233-55.

Hu, LT & Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.

Joreskog, K. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-26.

Joreskog, K., & Sorbom, D. (1996). LISREL 8: User's reference guide. Chicago: Scientific Software International

Kazis, L. E., Miller, D. R., Clark, J., Skinner, K., Lee, A., Ren, X. S., Spiro, A. III, Rogers, W. H., & Ware, J.E. (2002). Improving the response choices on the SF-36 role functioning scales (Veterans SF-36): Results from the Veterans Health Study. Medical Care, in press.

Kazis LE, Miller DR, Skinner KM, Lee A, Rogers WH, Spiro III A, Fincke, BG, Selim A, Linzer M. (1998). Health-related quality of life in patients by the Department of Veteran's Affairs. Archives of Internal Medicine, 158, 626-632.

Keller SD, Ware, JE, Bentler PM et al. (1998). Use of structural equation modeling to test the construct validity of the SF-36 health survey in ten countries: Results from the IQOLA project. Journal of Clinical Epidemiology, 51, 1179-88.

Marshall GN, Morales LS, Elliott M, Spritzer K, Hays RD (2001). Confirmatory factor analysis of the Consumer Assessment of Health Plans Study (CAHPS) 1.0 core survey. Psychological Assessment, 13, 216-29.

Meredith W & Horn J (2001). The role of factorial invariance in modeling growth and change. In LM Collins & AG Sayer (Eds.), New methods for the analysis of change (pp. 203-240). Washington DC: American Psychological Association.

NCQA (1998). HEDIS® 3.0/1998, Volume 6: Health of Seniors Manual. Washington, DC: National Committee for Quality Assurance.

Peabody JW, Luck J (1998). How far down the managed care road? A comparison of primary care outpatient services in a Veterans Affairs medical center and a capitated multispecialty group practice. Archives of Internal Medicine, 158, 2291-99.

Petersen LA, Normand SL, Daley J, McNeil BJ (2000). Outcome of myocardial infarction in Veterans Health Administration patients as compared with Medicare patients. New England Journal of Medicine, 343, 1934-41.

Rogers WH, Kazis, L., Miller, D., Skinner, K., Clark, J., Spiro, A. III, & Fincke, G. (2002). Comparing the health status of VA and non-VA ambulatory patients: The Veterans Health and Medical Outcomes Studies. Medical Care, in press.

Steenkamp J-B & Baumgartner H (1998). Assessing measurement invariance in cross-national consumer research. Journal of Consumer Research, 25, 78-90.

Steiger, JH & Lind J (1980, May). Statistically based tests for the number of common factors. Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.

Tucker, LR & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Vandenberg RJ & Lance CE (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3, 4-70.

Ware JE, Snow KK, Kosinski M, Gandek B. SF-36 health survey: manual and interpretation guide. Boston: The Health Institute, New England Medical Center; 1993.

Widaman KF & Reise SP (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In KJ Bryant, M Windle & SG West (Eds.), The science of prevention: Methodological advances from alcohol and substance abuse research (pp. 281-324). Washington DC: American Psychological Association.

8. Acronyms

BP	Bodily pain (SF-36 subscale)
CFI	Confirmatory fit index
CHF	Chronic heart failure
CLD	Chronic lung disease
COPD	Chronic obstructive pulmonary disease
df	degrees of freedom
GH	General health (SF-36 subscale)
HOS	Health Outcomes Survey (formerly, Health of Seniors)
ICD9CM	International Classification of Diseases, Version 9, Clinical Modification
MH	Mental health (SF-36 subscale)
MI	Myocardial infarction
MOS	Medical Outcomes Study
NNFI	Non-normed fit index (aka TLI)
NSVH	National Survey of Veterans Health
PF	Physical functioning (SF-36 subscale)
PTSD	Post-traumatic stress disorder
RE	Role limitations due to emotional functioning (SF-36 subscale)
RMSEA	Root mean squared error of approximation
RP	Role limitations due to physical functioning (SF-36 subscale)
SF	Social functioning (SF-36 subscale)
SRMR	Standardized root mean squared residual
TLI	Tucker-Lewis index (aka NNFI)
VA	US Department of Veterans Affairs