Using MIMIC Models to Assess the Influence of Differential Item Functioning

John A. Fleishman, Ph.D.

Center for Financing, Access, and Cost Trends

Agency for Healthcare Research and Quality

The opinions expressed in this article are those of the author. No official endorsement of the Agency for Healthcare Research and Quality or the Department of Health and Human Services is intended or should be inferred.

Using MIMIC Models to Assess the Influence of Differential Item Functioning

Making comparisons is the essence of research. We may want to know if patients who received a certain treatment had better health status than those who received an alternate treatment. Or, we may examine the extent to which persons with low socioeconomic status have worse health status than those who are more advantaged. One threat to the validity of comparisons is measurement non-equivalence, or differential item functioning (DIF). If DIF is present, it is like one's ruler expands for one group and contracts for another; true group differences in the construct of interest may be distorted if one's measure is not invariant across groups being compared.

It is useful to think of measurement as attempting to obtain quantitative information regarding an unobserved, or latent variable. The latent variable can be a construct that is difficult or impossible to measure directly, such as intelligence, an attitude, or health status. We have a number of observed indicators, such as items in a questionnaire, but no single item in isolation is equivalent to the latent construct we are trying to measure. However, we can make inferences regarding the latent variable based on patterns of relationships among the observed indicators.

In many measurement applications, we are conceptually interested in group differences in a latent variable. Group differences in observed variables (e.g., the means of questionnaire items in each group) would be proportional to the mean differences in latent variables if the measures were invariant. A group difference in observed variables does not, per se, indicate DIF. DIF is present if there are group differences in observed indicators, over and above differences in the latent variable.[1] The goal is to disentangle group differences in the latent variable from group differences arising from DIF.

This presentation amplifies some comments made by Jeanne Teresi and provides an example of adjusting for DIF using multiple-indicator, multiple cause structural equation models, also termed MIMIC models.

Dealing with DIF in Different Stages of Research

As has been noted, IRT and DIF detection were developed in the context of educational testing. Test developers often have considerable resources to devote to constructing and pretesting test items. Educational tests, such as the SAT, may go through several rounds of revision, in which poorly performing items are modified or possibly discarded altogether. Tests are based on large pools of candidate items, so dropping one or even several items because they have DIF might not seriously impede progress.

In other areas of research, in contrast, investigators might not have the luxury of being able to develop instruments through several rounds of pretests and revisions, nor might they have a large pool if items from which to select. In many instances, researchers opt to use established measures for which reliability and validity information has already been gathered. In the area of health status measurement, for example, there are literally thousands of studies that use the SF-36, or its shortened version, the SF-12.[2,3] These measures are widely accepted and have already undergone a rigorous phase of development, although, for the most part, formal investigations of DIF are only just starting to be conducted.

Consider the situation of someone who is using an established measure in his or her study, and then decides, after data have been collected, to examine the possibility of DIF. Suppose some of the items are found to exhibit DIF. The researcher is not in a position to develop a new, DIF-free set of items. Moreover, the researcher has good reasons to be reluctant to drop items with DIF from the analysis. First, dropping items changes the instrument, making

results non-comparable to research that used the original instrument. Second, dropping items might adversely affect the content validity of the instrument.

As an example of the second point, consider the SF-12, a generic measure of health status. The SF-12 is conceptualized as measuring eight domains of health: general health, physical functioning, role performance (physical), pain, emotional well-being, role performance (emotional), vitality, and social functioning. The eight domains, in turn, are viewed as reflecting two superordinate factors: physical health and mental health.[4] The standard scoring of Version 1 of the SF-12 produces two summary measures: the physical component summary (PCS) and the mental component summary (MCS). In the SF-12, some domains, such as pain or vitality, are represented by only one item. If these items are found to have DIF and are discarded, these domains will not be represented. This would mean that the reduced measure would exclude certain key components of health status, resulting in incomplete measurement. And unlike the situation of educational test developers, there is no pool of items to replace the ones with DIF.

As an alternative to either ignoring DIF or altering an established measure by discarding items, one can attempt to adjust for DIF in the analysis by modeling DIF. That is, one retains the possibly problematic items, but one includes parameters representing DIF effects in a more encompassing statistical model. In this context, several DIF-detection methods have shortcomings. Some techniques, such as the Mantel-Haensel approach, do not enable one to control for other covariates. Standard applications of IRT-based methods also do not typically include covariates other than the focal group difference being tested. However, it is possible to embed DIF effects in a special type of structural equation model, the MIMIC model.[5]

Demographic Differences in Health Status

To illustrate the use of MIMIC models, consider an example involving the SF-12.[6] The study investigated sociodemographic differences in health status. People have proposed using generic measures of health status, such as the SF-12, to monitor population health and to assess health differences among population subgroups. Several studies have used the SF-12 in sample surveys and have reported differences in physical or mental health across different subgroups.[3, 7-9]

- Older persons have lower physical health than younger ones;
- Mental health scores do not decline with age, and may even rise;
- Women report lower physical and mental health than men;
- Persons with more education report better physical health than the less-educated;
- Educational differences in mental health are less pronounced than for physical health;
- Black respondents had higher mental health scores than Whites or Hispanics in one study.

The question is, to what extent do these findings reflect true differences between subgroups in physical or mental health, and to what extent might they reflect DIF? For example, women might be more willing than men to admit physical limitations or psychological distress. Such a tendency would make health status items less difficult for women to endorse, compared to men at the same level of underlying health status.

MIMIC Model

The standard, unidimensional IRT model postulates that there is an underlying characteristic, such as physical health status, that cannot be observed directly. This is the latent variable denoted as theta (θ) in IRT models. We can observe theta indirectly, through its

influence on observed indicators. This model can be represented diagrammatically as a factor analysis model. If the indicators are dichotomous and the latent variable has a normal distribution, this is the same as the normal ogive IRT model.[10] The loadings of the indicators on the factor correspond to the discrimination parameters in IRT. The intercepts for each item correspond to the difficulty parameters in the IRT formulation. Figure 1 shows a factor model for the SF-12.

In the case of the SF-12, there are two latent variables (factors): physical health and mental health. A second factor can easily be added to the confirmatory factor analysis model. Note that standard IRT software assumes that there is only one latent factor. Multidimensional IRT models have been developed, but their use has not been widespread in applied research. Thus, another advantage of the MIMIC model approach is that it is not difficult to accommodate multiple dimensions or latent variables.

Note that, in this representation, each factor influences some of the indicators, but not all of them. In order to identify this model, and obtain unique parameter estimates, we have to assume that certain paths are absent. In this model, underlying physical health affects responses to items 2, 3, 4, 5, and 8, but mental health does not. Underlying mental health affects responses to items 6, 7, 9, and 11, but physical health does not. Items 1, 10, and 12 are influenced by both physical and mental health. (Typically, one strives for a simple structure, in which each item reflects only one factor. Empirical research suggests, however, that these three items are influenced by both physical and mental health.[4]) This is a standard confirmatory factor analysis model. It assumes that the correlations among the items arise from their common dependence on the latent factors.

The MIMIC model (Figure 2) extends this by incorporating additional variables, which are assumed to influence the latent factors. In the present case, we have gender, age, education, and race/ethnicity. The effects of these variables on underlying physical and mental health are represented by arrows from these variables to the latent factors. This part of the model can be construed as two multiple regressions: regressing physical health and mental health on the sociodemographic variables. For example, if gender is coded such that males are zero and females 1, a negative coefficient for the regression of physical health on gender would indicate that women have lower underlying physical health than men. So, we have multiple indicators, which reflect the underlying factors, and we have multiple causes, which affect the underlying factors. Hence the name MIMIC model.

To this point, the model does not include any representation of DIF. This model says that age, gender, etc., differences will be observed in the 12 items because the sociodemographic variables affect latent physical and mental health, which then influence responses to certain observed items. Controlling for the underlying factors, there are no sociodemographic differences in responses to the observed items. DIF is incorporated by adding direct effects from the sociodemographic variables to the observed indicators, unmediated by the latent factors. This represents a systematic difference in responses, controlling for the latent factor, which is the definition of DIF. In Figure 2, the dashed line from gender to one SF-12 item represents one possible DIF effect. Multiple DIF effects can be incorporated into the model. By estimating parameters corresponding to these direct effects, one can statistically control for them and then examine the other parameters of more substantive interest, specifically, the effects of the sociodemographic variables on physical and mental health. This approach has also been used to examine DIF in measures of depression, functional disability, and cognitive functioning.[11-14]

Data.

Let's apply the MIMIC model to examine sociodemographic differences in physical and mental health, using the SF-12. Data come from the Medical Expenditure Panel Survey (MEPS) conducted in 2000 by the Agency for Healthcare Research and Quality. [15,16] MEPS data are based on a nationally representative sample of civilian, non-institutionalized US residents. A self-administered questionnaire containing the SF-12 was distributed to respondents over the age of 17. The response rate for eligible respondents was 93.5%.

Overall, 15,438 adults returned the questionnaire. We excluded 1,792 questionnaires because they were completed by someone other than the intended person. We also excluded 1,964 persons due to missing data on one or more SF-12 items. The final analytic sample included 11,682 cases. (These analyses differ slightly from those in Fleishman and Lawrence[6] because the current analyses do not incorporate indicators of specific medical conditions and the analytic sample differs by 56 cases.)

Results.

As a basis of comparison, we examined sociodemographic differences in physical and mental health status using standard analytic techniques. Table 1 shows mean PCS and MCS scores for different sociodemographic categories. Table 2 shows results of separate multivariate regression analyses for the PCS and MCS. The multivariate analyses are consistent with the unadjusted means. Men had higher PCS and MCS scores than women. For education, both physical and mental health were higher for more highly educated groups; people with less than a high-school education were especially low in PCS and MCS scores. For age, PCS scores declined monotonically among older age groups, as one would expect, but MCS scores did not. The 60-69 year-old group reported the highest mean MCS scores. For race/ethnicity the pattern

was less clear. In the regression analyses, racial-ethnic differences in PCS were not significantly different, but Blacks and Hispanics had higher MCS scores than whites.

In the MIMIC model framework, the latent physical and mental health factors do not have a perfect correspondence with the PCS and MCS variables. The PCS and MCS are each weighted combinations of all the SF-12 items. The weights were derived from principal components analysis of the SF-12 items. Principal components analysis is not equivalent to factor analysis; it does not posit the existence of latent variables. The physical factor, in contrast, reflects 8 of the 12 items, and the mental health factor reflects 7 of the items. The variances of the latent factors also will not equal the variances of the PCS and MCS scores. Thus, sociodemographic group differences are in a different metric in the MIMIC analyses. One would not expect identical results from the multiple regression and MIMIC models.

The MIMIC analyses were conducted using Mplus software.[17] Mplus is general structural equation modeling software, which incorporates the MIMIC model as one possible specification. Because the responses to the SF-12 items were on ordinal response scales (ranging from 2 to 6 categories), with distributions that were not symmetric, weighted least squares estimation was performed, not maximum likelihood (which assumes normally distributed variables).

Table 3 shows the coefficients for the sociodemographic variables in the MIMIC model without DIF. These coefficients can be interpreted in a manner analogous to the coefficients in the regression model presented above.

- Men report higher physical and mental health than women.
- Physical health declines with age. However, mental health does not differ significantly across age groups.

- Physical health is lower among those with less education, compared with those who attended college. Mental health is also lower among those with less than a high-school education, even after controlling for age.
- Blacks report higher mental health than whites, but racial-ethnic differences in physical health are not significant.

There were some differences between the MIMIC model results and the multiple regression results, all involving mental health. In the regression analysis, the effects of having a high school degree and of being in the two oldest age groups were statistically significant, but they were not in the MIMIC model.

Finally, we incorporated DIF effects into the model. To do this, one must assign one item for each factor as having no DIF. Otherwise, the model is not identified. This selection of a no-DIF item could be based on prior research that identifies a suitable anchor item. Otherwise, one can proceed empirically. Most structural equation modeling software will report derivatives for parameters fixed at zero. The derivative indicates the extent to which the fitting function would change if a parameter that was constrained to equal zero (as in the no-DIF model) was instead feely estimated. We selected one item on each factor with the lowest derivatives, implying that the amount of DIF in that item was relatively small. These items were "limited in work or other activities due to physical health (PHYSLIM)," and "accomplished less than you would like as a result of emotional problems (MENTLESS)." We then incorporated all the other possible DIF effects into the model. For 9 sociodemographic variables and 10 SF-12 items (excluding the two no-DIF items), there were 90 additional DIF parameters to estimate.

The selection of the no-DIF anchor items may not be clear-cut and may require a judgment call. It may be the case that no item appears to be absolutely devoid of DIF. It is

useful to examine the sensitivity of the model to different selections of anchor items. Candidates should be items that appear to have minimal DIF effects. We therefore re-estimated the model, assigning two other items with relatively small derivatives as the anchor items. These new anchor items were "Accomplished less than you would like at work or other daily activities as a result of physical health (PHYSLESS)," and "didn't do work or other activities as carefully as usual due to emotional problems (MENTLIM)."

For both models, we examined the magnitude and significance of the estimated DIF effects for each item. This procedure provided information useful in selecting anchor items. Both items used to anchor the physical health dimension (PHYSLESS and PHYSLIM) had DIF parameters that were generally small and non-significant; either could be a reasonable choice for an anchor item. MENTLIM had one strong DIF effect, with the variable indicating less than high-school education (parameter = -0.167, s.e=0.03, t=-5.644). MENTLESS also had one strong DIF effect with having less than a high-school education (parameter = 0.173, s.e.=0.031, t=5.616). These results did not provide a clear choice between MENTLESS and MENTLIM as anchors for mental health. Somewhat arbitrarily, we chose MENTLESS as the anchor for mental health, and used both PHYSLESS and PHYSLIM as anchors for physical health. We then incorporated into the model those DIF effects that were significant in either of the two preceding models. The resulting model included 39 additional DIF parameters.

Several criteria can be used to evaluate the fit of the no-DIF and the DIF models. A goodness-of-fit statistic, reflecting the discrepancy between the observed data (item means and covariances) and the model's predictions, can be referred to a chi-square distribution. However, because statistical power increases with sample size, chi-square goodness of fit tests in large samples should be viewed with caution because trivial differences often appear statistically

significant. Consequently, we also examined other indicators of goodness-of-fit. The Comparative Fit Index (CFI) compares the substantive model to a baseline null model of independence among the observed variables; values of 0.95 or higher suggest acceptable fit.[18] The Root Mean Square Error of Approximation (RMSEA) assesses misfit per degree of freedom; values less than 0.08 suggest an acceptable fit, while values less than 0.05 suggest very good fit.[19]

Table 4 shows goodness-of-fit statistics for both the no-DIF and the DIF models. Both models had significant chi-square values, suggesting that the models did not fit to within sampling error. However, with over 11,000 cases, fairly trivial departures from the true model may inflate the chi-square statistic. In terms of other fit indices, both models had acceptable CFI and RMSEA, with the values for the DIF model being slightly better than those for the no-DIF model. The difference in the chi-squares between the no-DIF and DIF models was significant, implying that the DIF model is an improvement on the no-DIF model.

When the model included parameters reflecting DIF, what happened to the estimates of sociodemographic differences in physical and mental health? Table 5 shows the revised estimates. For physical health status, we see the following: (1) The gender difference changed only slightly. (2) Racial/ethnic differences were slightly greater, but still not significant. (3) Effects for age were slightly smaller, but still significant. (4) Effects for education were slightly smaller, but still significant. In sum, after controlling for DIF, the overall picture of sociodemographic differences in physical health did not change radically, compared with the no-DIF model.

For mental health status, the gender effect was slightly smaller in magnitude, but still significant. For race/ethnicity, however, the effect for Blacks was now no longer significantly

different from Whites. For age, the effects of being in the 40-59 and the 60-69 age groups were smaller and non-significant. More notable, the effect for the 70+ age group, which was not significant in the no-DIF model, was now significantly negative. Finally, educational differences were virtually unchanged.

The overall picture is that sociodemographic differences in physical health did not appear to be distorted appreciably by DIF. For mental health, in contrast, effects of race/ethnicity and age did change, while gender and education remained stable. The MIMIC model has thus highlighted two points at which DIF might affect conclusions regarding sociodemographic differences in health.

Table 6 shows the specific DIF parameters estimated. The item "felt calm and peaceful" had a large number of relatively large DIF effects. Future studies might investigate the extent to which persons in different demographic groups interpret this item differently. DIF effects are prevalent for the age 70+ group. In particular, people in this group are especially likely to endorse feeling calm, having energy, health not interfering with social activities, and not feeling downhearted. This contributes to the differences between the no-DIF and DIF results for this group. Responses to these items might reflect a response shift or frame of reference effect, in which people respond that they're feeling good, compared to others in their age group or compared to a negative stereotype of elderly people.

Conclusions

This analysis demonstrates how one would examine the presence of DIF using the MIMIC model approach. This approach has several things to recommend it:

- DIF detection analyses can be embedded in a larger structural model that focuses on substantive issues of concern.

- Analyses can proceed even if one cannot exclude items with DIF or collect new data using a refined measure.
- The MIMIC model can incorporate multiple latent variables simultaneously.
- Comparing estimated effects of key variables when controlling for DIF and not controlling for DIF provides an intuitive sense of the extent to which DIF may be distorting substantively important comparisons.
- Standard IRT-based DIF assessment usually examines two groups at a time. If one has multiple groups, this requires multiple pairwise DIF assessments. In contrast, multiple groups can be handled simultaneously in the MIMIC model.

There are some disadvantages of using the MIMIC model to examine DIF. First, the model assumes that the loadings of the observed variables on the latent factors are the same in all groups. This essentially assumes that non-uniform DIF, shifts in the discrimination parameter from group to group, is not present. One approach for dealing with the possibility of non-uniform DIF is to conduct multiple-group MIMIC modeling. For example, we could compare a MIMIC model (excluding age) for the 18-39 year-old group versus the 70+ group. This approach, however, encounters the common problems of having to do multiple comparisons among groups with potentially small sample sizes. Second, although factor scores can be estimated for individual persons, the MIMIC model approach does not focus on producing an overall score for each person. If the study is focused on determining a score to be used to make decisions regarding specific individuals, then other DIF detection approaches may be preferable. Third, to date, available software for MIMIC models uses only data on means and covariances; IRT estimation software, in contrast, uses information on the specific pattern of responses to each item. Technically, MIMIC models are estimated using limited-information methods, while

IRT programs use full-information methods. (However, this may change in newer versions of commercially available structural equation modeling software.)

Despite its limitations, the MIMIC model is a useful approach for applied researchers who want to control for DIF in analyses of broader models that represent substantive research questions.

References

1. Millsap, R.E., Everson, H.T. Methodology review: Statistical approaches for assessing measurement bias. App Psych Meas, 1993; 17:297.

2. Ware JE, Kosinski M, Keller SD. A 12-item short-form health survey: Construction of scales and preliminary tests of reliability and validity. Med Care 1996; 34:220.

3. Ware JE, Kosinski M, Keller SD. How to score the SF-12 physical and mental health summary scales, 3rd Ed. Lincoln, RI: QualityMetric, Inc., 1998.

4. Keller SD, Ware JE, Bentler PM, et al. Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: Results from the IQOLA Project. J Clin Epidemiol 1998; 51: 1179.

5. Muthen BO. Latent variable modeling in heterogeneous populations. Psychometrika, 1989; 54:557-585.

6. Fleishman, JA, Lawrence WF. Demographic variation in SF-12 Scores: True Differences or Differential Item Functioning. Med Care, 2003; 41: III-75 – III-86.

7. Burdine JN, Felix MRJ, Abel AL, et al. The SF-12 as a population health measure: An exploratory examination of potential for application. Health Serv Res 2000; 35:885.

8. Johnson JA, Coons SJ. Comparison of the EQ-5D and SF-12 in an adult US sample. Qual Life Res 1998; 7: 155.

9. Jenkinson C, Chandola T, Coulter A, Bruster S. An assessment of the construct validity of the SF-12 summary scores across ethnic groups. J Pub Health Med 2001; 23: 187.

10. Muthen BO, Lehman J. Multiple group IRT modeling: Applications to item bias analysis. J Ed Stat 1985; 10: 133.

11. Fleishman JA, Spector WD, Altman BM. Impact of differential item functioning on age and gender differences in functional disability. Journals of Gerontology: Social Sciences 2002; 57(B): S275-S284.

12. Gallo JJ, Anthony JC, Muthen BO. Age differences in the symptoms of depression: A latent trait analysis. Journals of Gerontology: Psychological Sciences, 1994; 49: P251-P264.

13. Grayson DA, Mackinnon A, Jorm AF, Creasey H, Broe GA. Item bias in the Center for Epidemiologic Studies depression scale: Effects of physical disorders and disability in an elderly community sample. Journals of Gerontology: Psychological Sciences, 2000; 55B: P273-P282.

14. Jones RN, Gallo JJ. Education and sex differences in the Mini-Mental State Examination: Effects of differential item functioning. J Gero: Psychological Sciences 2002; 57B: P548-P558.

15. PUF Documentation Files: MEPS HC-0039, 2000 Full-Year Population Characteristics. Rockville, MD: Agency for Healthcare Research and Quality, November 2002. Available at http://www.meps.ahrq.gov/puf/DataResultsDoc.asp. Accessed on January 21, 2003.

16. Cohen SB, DiGaetano R, Goksel H. Estimation procedures in the 1996 Medical Expenditure Panel Survey Household Component. MEPS Methodology Report No. 5. Rockville, MD: Agency for Health Care Policy and Research, 1999. AHCPR Pub. No. 99-0027.

17. Muthen LK, Muthen BO. Mplus User's Guide. Los Angeles, CA: Muthen and Muthen, 1998.

18. Hu LT, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling 1999; 6: 1.

19. Browne MW, Cudeck R. Alternative ways of assessing model fit. In Bollen KA, Long JS (Eds.), Testing Structural Equation Models. Thousand Oaks, CA: Sage, 1993, pp 136-162.

Table 1. Variation in PCS-12 and MCS-12 Scores by Demographic Characteristics

Variable	Weighted Proportion	Mean PSC-12	Mean MCS-12
Gender			
Female	0.55	49.35	50.69
Male	0.45	50.83	52.28
Education			
< High School	0.17	47.03	49.67
High School	0.33	49.27	51.46
Some College	0.51	51.48	51.95
Race/Ethnicity			
White	0.77	49.87	51.43
Black	0.10	50.25	51.62
Hispanic	0.09	50.42	50.88
Other	0.03	51.63	51.82
Age Group			
18-39	0.45	52.78	51.07
40-59	0.37	49.76	51.16
60-69	0.09	46.08	53.00
70+	0.09	41.24	52.54

N=11,682.

Analyses based on weighted data.

Table 2. Regressions of PCS-12 and MCS-12 on Demographic Characteristics

	Dependent Variable	
Independent Variable	PCS-12	MCS-12
Male	1.36 (.19)***	1.58 (.15)***
Black	0.01 (.30)	0.68 (.34)*
Hispanic	0.05 (.25)	0.23 (.33)*
Other race	0.39 (.52)	0.56 (.74)
Age 40-59	-3.24 (.19)***	-0.04 (.21)
Age 60-69	-6.62 (.37)***	2.00 (.35)***
Age 70+	-11.01 (.43)***	1.92 (.41)***
No high school degree	-3.75 (.29)***	-2.56 (.30)***
High school degree	-1.90 (.21)***	-0.57 (.22)*
Constant	53.42	50.85

N=11,682. Analyses based on weighted data. Standard errors in parentheses.

* $p < .05$; *** $p < .001$.

Table 3. Effects of Demographic Variables on Physical and Mental Health in No-DIF Model

	Physical Factor	Mental Factor
Variable	No DIF	No DIF
Male	0.205 (.022)*	0.186 (.018)*
Black	0.020 (.034)	0.114 (.027)*
Hispanic	0.008 (.030)	0.082 (.024)
Other race	0.072 (.068)	0.108 (.054)
Age 40-59	-0.407 (.025)*	-0.034 (.020)
Age 60-69	-0.710 (.039)*	0.099 (.032)
Age 70+	-1.062 (.039)*	0.001 (.032)
No high school degree	-0.459 (.030)*	-0.224 (.024)*
High school degree	-0.238 (.025)*	-0.059 (.021)

N=11,682. Analyses based on weighted data. Standard errors in parentheses.

* -- Ratio of parameter to standard error exceeds 3.5.

Table 4. Goodness-of-Fit of No-DIF and DIF Models

	No DIF Model	DIF Model
Chi-square	3676.98	1845.699
Degrees of freedom	135	96
CFI	.989	.994
RMSEA	.047	.039

Table 5. Effects of Demographic Variables on Physical and Mental Health in DIF Model

Variable	Physical Factor	Mental Factor
Male	0.193 (.023)*	0.181 (.019)*
Black	0.079 (.034)	0.056 (.029)
Hispanic	0.025 (.032)	-0.037 (.027)
Other race	0.073 (.068)	0.108 (.054)
Age 40-59	-0.348 (.026)*	-0.064 (.022)
Age 60-69	-0.654 (.039)*	-0.111 (.040)
Age 70+	-1.048 (.040)*	-0.279 (.040)*
No high school degree	-0.350 (.031)*	-0.237 (.026)*
High school degree	-0.158 (.028)*	-0.077 (.022)

N=11,682. Analyses based on weighted data. Standard errors in parentheses.

* -- Ratio of parameter to standard error exceeds 3.5.

Table 6. Direct Effects (DIF) of Demographic Variables on SF-12 Items.

SF-12 Item	Male	Age 40-59	Age 60-69	Age 70+	Black
General health	-0.076*	-0.084*	-------	-0.004	-0.165*
Moderate activities	-----	-0.121*	-0.186*	-0.197*	-0.128*
Climbing stairs	0.111*	-0.173*	-0.245*	-0.278*	-----
Accomplished less - physical[a]	-----	-----	------	-----	-----
Limited in work[a]	-----	-----	-----	-----	-----
Pain	----	-----	-----	0.278*	-----
Accomplished less - emotional	-----	-----	-0.021	-0.093	-----
Worked less carefully[a]	-----	-----	-----	-----	-----
Felt calm	-----	0.062	0.335*	0.546*	0.242*
Felt downhearted	-----	-----	0.276*	0.445*	-----
Had energy	0.073*	-----	0.228*	0.235*	-----
Social Activities	-----	0.106*	0.329*	0.470*	-----

Table 6 (continued). Direct Effects (DIF) of Demographic Variables on SF-12 Items.

SF-12 Item	Hispanic	Other race	Education < 12	High school
General health	-0.175*	-------	-0.303*	-0.221*
Moderate activities	----	------	-0.167*	-----
Climbing stairs	----	------	-0.166*	-0.117*
Accomplished less - physical	-----	-----	-----	-----
Limited in work	-----	-----	-----	-----
Pain	------	-----	-----	-0.089*
Accomplished less - emotional	------	------	-0.117	------
Worked less carefully	-----	-----	-----	-----
Felt calm	0.314*	-----	0.129*	0.082*
Felt downhearted	-----	-----	-----	-----
Had energy	0.281*	-----	-----	-----
Social Activities	-----	-----	-----	-----

Dashed lines indicate that the parameter was not estimated (i.e., was fixed to equal zero).

A – Item selected as anchor item with no DIF.

* -- Ratio of parameter to standard error exceeds 3.5.

Figure 1.

Two-Dimensional Confirmatory Factor Analysis Model for the SF-12

Observed variables are shown as rectangles and latent variables as ovals. The effect of one variable on another is represented by an arrow from the first to the second. The arrows represent parameters to be estimated. The individual SF-12 items, on the right-hand side of the figure, are influenced by physical health or mental health. (For simplicity, the Figure does not show residual correlation between the physical and mental health factors, correlated errors between SF-12 items, or residual variation in SF-12 items.)

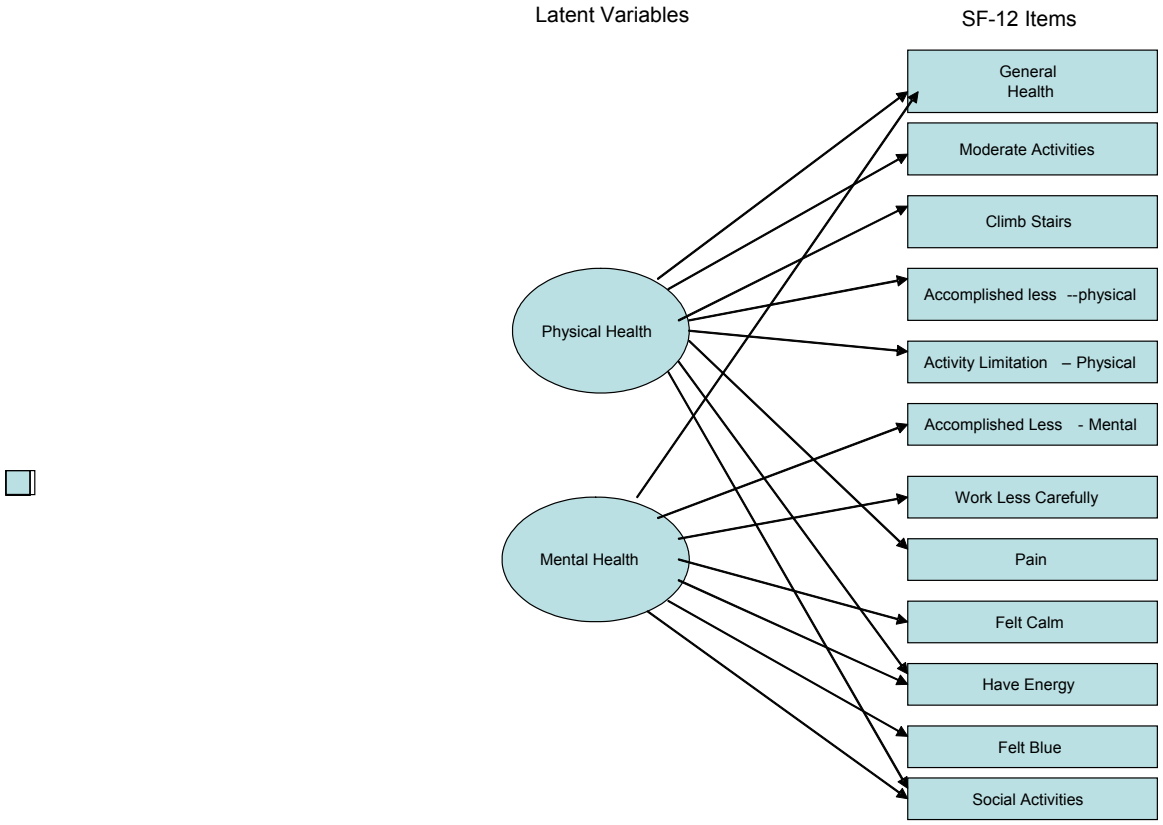
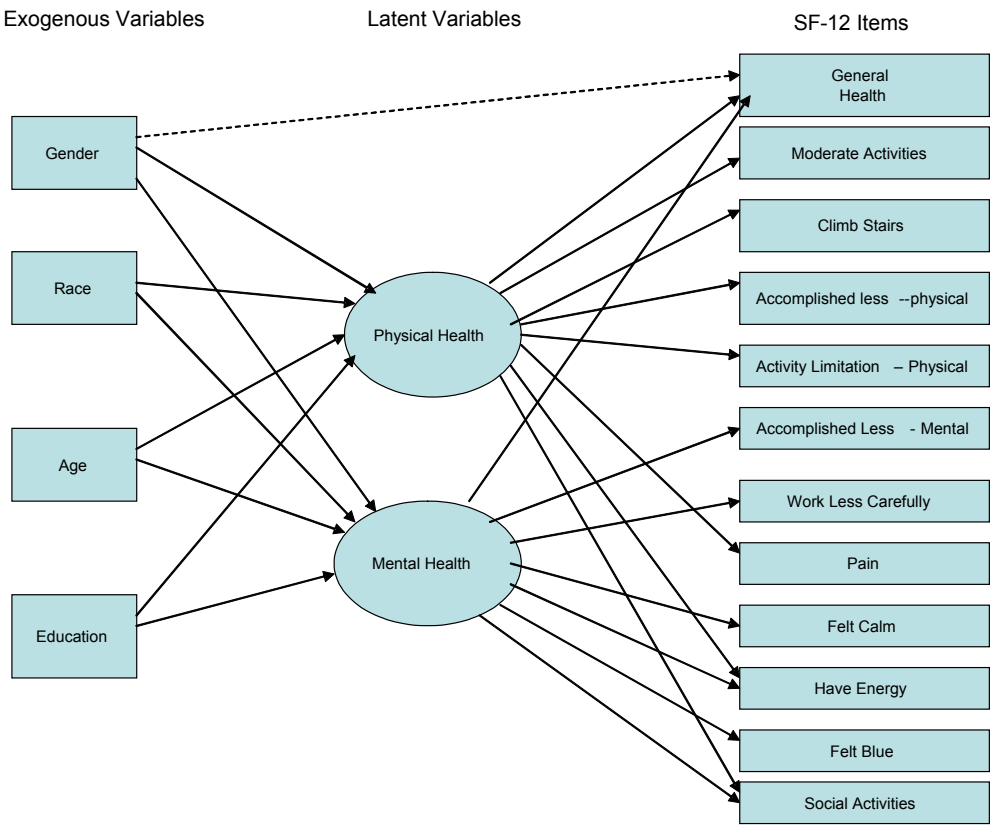

Latent Variables
SF-12 Items
Physical Health
Mental Health
General Health
Moderate Activities
Climb Stairs
Accomplished less --physical
Activity Limitation – Physical
Accomplished Less - Mental
Work Less Carefully
Pain
Felt Calm
Have Energy
Felt Blue
Social Activities

Figure 2.
Multiple-Indicator Multiple-Cause (MIMIC) Model of Demographic Variation in the SF-12.

Observed variables are shown as rectangles and latent variables as ovals. The effect of one variable on another is represented by an arrow from the first to the second. The arrows represent parameters to be estimated. The individual SF-12 items, on the right-hand side of the figure, are influenced by physical health or mental health. The two latent factors, in turn, are affected by the exogenous demographic variables on the left-hand side of the Figure. (For simplicity, the Figure does not show correlations among the exogenous variables, residual correlation between the physical and mental health factors, correlated errors between SF-12 items, or residual variation in SF-12 items.) The dashed arrow from "gender" to "general health" represents one possible DIF effect.

Figure 2.
M

Ob
va
pa
inf
the
Fig
the
va
po


Exogenous Variables
Latent Variables
SF-12 Items
Gender
Race
Age
Education
Physical Health
Mental Health
General Health
Moderate Activities
Climb Stairs
Accomplished less --physical
Activity Limitation – Physical
Accomplished Less - Mental
Work Less Carefully
Pain
Felt Calm
Have Energy
Felt Blue
Social Activities