Published in final edited form as:

J Psychiatr Res. 2009 August ; 43(12): 1025–1035. doi:10.1016/j.jpsychires.2008.12.007.

Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: An extension of the bi-factor item response theory model for use in differential item functioning

Frances M. Yang^{*}, Doug Tommet, and Richard N. Jones

Harvard Medical School, Brigham and Women's Hospital, Department of Psychiatry, Institute for Aging Research, Hebrew SeniorLife, 1200 Centre St., Boston, MA 02131, United States

Abstract

This study evaluates the measurement noninvariance, or differential item functioning (DIF), in the Center for Epidemiological Studies-Depression (CES-D) items attributable to age, sex, and race/ethnicity among community-dwelling older adults. Participants ($N = 2773$) were from the New Haven site of the Established Populations for Epidemiologic Studies in the Elderly. Statistical analyses included exploratory factor analysis, bi-factor confirmatory factor analysis, and a bi-factor multiple indicator and multiple causes (MIMIC) model to address measurement noninvariance. Blacks, compared with whites, were more likely to endorse items loading on the *interpersonal* factor, which include “people dislike me” and “people are unfriendly.” Women were less likely to endorse the *interpersonal* items and the “I felt like a failure” item (odds ratio [OR] = 0.63, 95% confidence interval [CI]: 0.42, 0.94) than men. But women had a higher proportional odds than men for endorsing the “crying” item (OR = 1.86, 95% CI: 1.17, 2.96). Those 75 years and older (relative to those aged 65–74) were less likely to endorse the “I felt like a failure” item (OR = 0.65, 95% CI: 0.43, 0.97). However, measurement noninvariance found in both the “crying” and “failure” items were attributable to women and to those aged 75 and older were trivial after controlling for the underlying level of depressive symptomatology. Therefore, the *interpersonal* items showed measurement noninvariance attributable to sex and race. The bi-factor MIMIC model is useful for examining measurement noninvariance due to sociodemographics in a multidimensional depression instrument.

Keywords

Center for Epidemiological Studies-Depression (CES-D); Older adults; Multiple indicator and multiple causes (MIMIC) model; Differential item functioning; Item response theory (IRT); Bi-factor model; Health disparities

© 2008 Elsevier Ltd. All rights reserved.

*Corresponding author. Tel.: +1 617 363 8270; fax: +1 617 363 8936. E-mail addresses: francesyang@hrca.harvard.edu, fmyang@bics.bwh.harvard.edu (F.M. Yang).

Conflict of interest statement

All authors declare that they have no conflicts of interest.

Contributors

Under the direction of Jones, Yang constructed the research questions, designed the study, conducted the analyses, and wrote each draft of the manuscript until completion for submission; Jones provided guidance and interpretation of the analyses, and edited each draft of the manuscript. Tommet contributed to the statistical analysis and editing of the manuscript. Jones revised each draft critically for important intellectual content, particularly with regards to methodology. All authors have read and approved the final manuscript.

1. Introduction

Health disparities persist among older Americans, despite declining disability (Martin et al., 2007). Untangling health disparities is a complex research and public health issue, but addressing this issue allows us to understand differences in the access to and use of health care services (Barreto and Segal, 2005; Cook et al., 2007; Dobalian and Rivers, 2008; Huguet et al., 2008). Specifically, for mental health care, there are race and ethnic disparities that persist due to distrust of health care providers, language barriers, economic barriers, lack of health insurance, and cultural barriers by minorities (Dobalian and Rivers, 2008). Among Medicare beneficiaries the prevalence of depression doubled between 1992 and 1998, and persons aged 75 and older, of “Hispanic or other” ethnicity, and without additional supplemental health insurance were significantly undertreated for depression (Lipton et al., 2003). Older adults are less likely than younger adults to visit a mental health professional (MHP) (e.g., “psychiatrist, psychologist, psychiatric nurse, or clinical social worker”) (Dobalian and Rivers, 2008; Mojtabai and Olfson, 2004; Robins and Regier, 1991). Among men and women with a history of depression, men were more likely to have visited an MHP than women. While African Americans and Hispanics were less likely to have visited an MHP in the prior year than were whites (Dobalian and Rivers, 2008).

In addition to disparities in mental health services, persistent underdiagnosis and undertreatment of psychiatric illnesses might also be due to measurement issues (Teresi et al., 2006). Item bias in psychiatric scales due to age, sex, or race/ethnic differences can contribute further to ongoing health disparities. Though many psychiatric and mental health scales have been validated using classical test theory, few have been examined on the item level to determine sources of measurement bias using item response theory (IRT), which has been well developed in the field of education. For example, Gallo and Rabins (1999) have shown that older adults may have a clinically significant depression without fulfilling either of the first two required criteria for major depression (dysphoria, anhedonia) in the Diagnostic and Statistical Manual of Mental Disorders (DSM), 4th edition (American Psychiatric Association, 1994). Instead, older adults, especially those with strong cultural influences (Kleinman, 1988), tend to present somatic complaints unexplained by comorbid illnesses to the clinician (Gallo and Rabins, 1999). The expression of the item or symptom level information by diverse older adults may differ due to cultural, historical, social, and medical influences. Therefore, “sampling items from identified subdomains of some larger domain,” as suggested by Gibbons and colleagues (2007, p. 4), is one measurement strategy that might help clinicians using psychiatric scales to efficiently and accurately identify depression (the “larger domain”) from somatic complaints (an “identified subdomain”).

Healthy People 2010: Understanding and Improving Health reports that disparities in mental health disorders are due to the differences in prevention, diagnosis, and treatment across age, sex, and race/ethnic groups in the US (National Center for Health Statistics, 2001). The mechanisms and processes that drive persistent mental health disparities may begin with bias in mental health assessment. This analysis extends the detection of measurement noninvariance, or differential item functioning (DIF), due to age, sex, and race/ethnic differences in the Centers for Epidemiologic Studies-Depression (CES-D) items (Yang and Jones, 2007) by using a bi-factor item response theory model.

1.1. Modeling multidimensional psychiatric scale data: using the second-order model versus the bi-factor item response theory (IRT) model to address mental health disparities

To address the complexity of eliminating health disparities in the US, efforts are under way to understand the measurement of self-reported health in diverse populations (Manly, 2006; Stahl and Hahn, 2006). Item response theory (IRT) examines individual questions in an instrument. IRT, also referred to as latent trait theory, estimates the latent trait level (severity level) of an

individual responding to a test item, the severity of the test item (or symptom), and the accuracy with which the item measures the latent trait (Lord, 1952; Rasch, 1960).

A common issue in health, mental health, and quality of life is that there exists a global, general, or quality of life construct, respectively; however, there may be several subdomains such as physical, functional, and emotional health that are included in the overall measure. The challenge of applying IRT to psychiatric scale data is that a fundamental assumption of traditional IRT models is the unidimensionality of a scale construct. However, most psychiatric scale data are fundamentally multidimensional, such as negative and positive affect. Recent formulations of IRT models have been articulated that allow for accurate assessment of dimensional constructs, which include both second-order models and bi-factor models. The bi-factor model recognizes a response item as being dependent upon both a single general factor (which loads in all items in a scale) and a specific subdomain factor (Gibbons et al., 2007; Gibbons and Hedeker, 1992). Besides the bi-factor model, there are other models that might explore better the dimensionality of a scale (Brown, 2006). The second-order model describes one or more distinct higher order factors, which items in a scale load onto, which represent constructs related to each other through a single general factor (Chen et al., 2005). The bi-factor and the second-order models are both multidimensional (hierarchical) IRT models that can be the same model when proportionality constraints are imposed (Chen et al., 2006) using the Schmid-Leiman (1957) transformation method and the results are interpreted in a similar fashion (Gustafsson and Balke, 1993). However, the two models are mathematically different with several potential advantages of the bi-factor model over the second-order model, thoroughly and concisely described by Chen et al. (2006), which include estimating fewer parameters and reducing model complexity that are applicable for detecting measurement noninvariance in multidimensional psychiatric scales.

Simulations have shown that false DIF detection can result from violations of the assumptions of unidimensionality inherent in most methods used for DIF detection (see Mazor et al., 1998). Thus, proper testing and modeling of dimensionality is an important aspect of DIF analysis. The bi-factor model is useful for both modeling and explaining DIF in the presence of a strong underlying factor, accompanied by multidimensionality, as is the case for many “broad” psychiatric scales operationalized by heterogeneous indicators (Reise et al., 2007). The advantages of the bi-factor approach for psychiatric measurement research was presented by Gibbons et al. (2009); these authors demonstrated the superior fit of the bi-factor model in comparison with other more traditional approaches in the analyses of a psychiatric diagnostic screening questionnaire (PDSQ; Zimmerman and Mattia, 2001).

1.2. Detecting measurement noninvariance or differential item functioning (DIF) in the CES-D due to sociodemographic differences

Measurement noninvariance in the CES-D and other depression scales that might be due to age, sex, and race/ethnic group differences can occur at a given level of depressive severity. These differences can result in “differential item functioning (DIF)”, which is the empirical investigation of how items perform across groups of individuals (Teresi, 2006a,b), and may impact scale scores and inferences about differences across groups. Non-uniform (magnitude and direction of DIF varies as a function of the underlying trait) and uniform DIF (magnitude is proportional across all levels of the underlying trait) are two types of DIF, similar to the epidemiologic concepts of effect modification and confounding, respectively (Crane et al., 2004).

In this study, we use the multiple indicator, multiple causes (MIMIC) model, an extension of the confirmatory factor analytic model for dichotomous dependent variables with covariates (Gallo et al., 1994; Muthén, 1989). MIMIC modeling was originally presented as a way to model unobserved quantities measured by formative causes and reflective indicators (Hauser

and Goldberger, 1971). We prefer the MIMIC model over other strategies of DIF detection because it is a straight forward extension to a multidimensional model, permits control of multiple covariates, and the implementation using Mplus software (Muthén & Muthén, Los Angeles, CA) permits the estimation of two parameter normal ogive IRT model and the inclusion of complex sampling weights. Modeling DIF with a MIMIC approach will detect uniform DIF, or proportional differences across group in the probability of an indicated response over the latent trait. Other approaches, including multiple group CFA (MGCFA) can detect non-uniform DIF (Jones, 2003). However, the correct identification of specific type of DIF is poor (based on our own ongoing simulation studies and the recent work of Finch and French (2008)). The MIMIC model can detect DIF regardless of whether the DIF is caused by group differences in IRT discrimination parameters or difficulty parameters, so long as the DIF produces a shift in the conditional probability of an indicated response across group. Absent *a priori* hypothesis about the type of DIF each item is expected to have, since it is difficult for the model to correctly distinguish between the types of DIF (Finch and French, 2008) a safer inference is in the determination of any DIF.

The bi-factor MIMIC model can be a useful and parsimonious tool to summarize DIF across multiple items. Fig. 1 illustrates a hypothetical unidimensional MIMIC model with one general construct (g) with the DIF effect found in the third item (y_3) of the scale due to the covariate (x). The hypothetical bi-factor MIMIC model (Fig. 2) adds one specific construct (measured by items y_1 – y_3 of the scale) to the unidimensional MIMIC model in Fig. 1. In our descriptions of such models, effects relating covariates and items can be described in terms of a total effect, a general effect, a specific effect, and a DIF effect. The total effect represents the sum of all pathways from the covariate to the item via both the general and specific factors, as well as directly from the covariate to an item (Fig. 2). The general effect is the indirect effect between the covariate and an item as mediated by the general latent factor. The DIF effect is the direct effect from the covariate to an item. The unidimensional DIF effect is decomposed in the bi-factor model into a specific effect and a (residual) DIF effect. The specific effect is the pathway from the covariate to the specific factor to the item (Fig. 2).

1.3. Objectives of the study

In this study, we apply the bi-factor MIMIC model to the Center for Epidemiological Studies-Depression (CES-D) scale. Factor analytic studies of the full 20-item CES-D scale frequently report four dimensions: depressed affect, low positive affect, somatic complaints, and interpersonal difficulties (Blazer et al., 1998; Hertzog et al., 1990; Iwata and Roberts, 1996; Knight et al., 1997; Radloff, 1977). Most researchers ignore the multidimensionality of the CES-D when conducting studies of depressive symptoms, including those that attempt to discover item bias or differential item functioning (DIF). For example, using different DIF detection methods, Cole et al. (2000) and Yang and Jones (2007) found three items with DIF due to sex and race/ethnic differences in a unidimensional CES-D scale from the New Haven Established Populations for Epidemiological Studies in the Elderly (EPESE).

Analytic methods for detecting DIF are potentially important and useful in evaluating the validity of depression measures and other health outcome measures in diverse populations (McHorney, 2003). In this study, we reexamine measurement noninvariance due to age, sex, and race/ethnicity in the New Haven EPESE using the bi-factor model for the CES-D scale. Hence, the purpose of this study is to probe the meaning of DIF that might be due to age, sex, and race/ethnic differences in the CES-D using a bi-factor model approach to DIF detection. We hypothesize that there is both a general depressive factor and the following four independent factors: *depressed affect*, *interpersonal difficulties*, *low positive affect*, and *somatic complaints*. Based on relevant findings from previous studies (Cole et al., 2000; Yang and Jones, 2007), we anticipate that the interpersonal items of “people dislike me” and “people

were unfriendly” will be the largest nuisance factor, which represents the items least correlated with the other CES-D items. We hypothesize that there will be measurement noninvariance in the CES-D items due to age, sex, and race/ethnicity.

2. Methods

2.1. Sample

The New Haven Established Populations for Epidemiologic Studies of the Elderly (EPESE) public use data were obtained from the Interuniversity Consortium for Political and Social Research (ICPSR). The New Haven EPESE was a population-based survey of adults aged 65 and older who provided information about risks for chronic diseases, functional loss, mortality, hospitalization, and placement in long-term care facilities. The New Haven EPESE used a stratified random sample of clusters of older adults who lived in public housing, private housing, and elsewhere in the community. Men were oversampled to attempt to achieve balance in the sex distribution of the sample. Further details of the New Haven and other EPESE locations are presented elsewhere (Taylor et al., 1998). The exclusion criteria for this study were a missed response on at least one CES-D question or background variable (age, sex, or race) ($N = 472$). The listwise complete sampling on CES-D responses is not a requirement of the modeling software, but we adopt it to maintain consistency with previous studies of DIF in the CES-D using this sample (Cole et al., 2000; Yang and Jones, 2007). The final sample in this study consisted of respondents who provided complete data on all CES-D questions and background variables ($N = 2733$). Over half of the participants were ages 65–74 (62%), women (58%), white (81%), did not graduate high school (68%), and were single (59%) (Table 1).

2.2. Instrumentation

2.2.1. Depression assessment—The New Haven EPESE respondents were presented with the original version of the 20 CES-D (Radloff, 1977) items with a four-category response (rarely or none of the time, some of the time, much of the time, most or all of the time) and polytomously scored 0–3. Each item was preceded by: “During the past week, I...” followed by a depressive symptom, for example the “bothered” item was asked: “was bothered by things that don’t usually bother me.” The following are the rest of the CES-D items and the shortened names for the items are included in parentheses: “did not feel like eating; my appetite was poor (eating),” “could not shake off the blues even with the help of my family or friends (blues),” “felt that I was just as good as other people (good),” “trouble keeping mind on task (mind),” “felt depressed (depressed),” “everything an effort (effort),” “felt hopeful (hopeful),” “life had been a failure (failure),” “felt fearful (fearful),” “restless sleep (sleep),” “was happy (happy),” “talked less than usual (talked),” “felt lonely (lonely),” “people were unfriendly (unfriendly),” “enjoyed life (enjoy),” “had crying spells (cry),” “felt sad (sad),” “felt that people disliked me (dislike),” and “could not get going (not get going).” For the study data, internal consistency estimates (using Cronbach’s alpha) for each of the sub-groups are as follows: less than age 75 (0.87), age 75 and greater (0.85), male (0.85), female (0.86), white (0.86), and black (0.84). The Cronbach’s alpha for each sub-group are very close to the entire sample (0.86) (see Table 2).

Radloff (1977) first developed the CES-D to measure the frequency of depressive symptoms in the general population through survey data collection (c.f., Steffick (2000)). Ross and Mirowsky (1984) found the following four-factor structure in the original 20-item CES-D: depressed affect, enervation, lack of positive affect, and interpersonal problems. Since the year 2000, other works have been published on psychometric properties of the CES-D scale in diverse older adults (Chan et al., 2004; Cole et al., 2000; Foley et al., 2002; Gregorich, 2006; Mui et al., 2001; Yang and Jones, 2007, 2008).

2.2.2. Background variables—Age was treated as a binary variable with those who were ages 65–74 (=0) compared with those who were aged 75 and older (=1), which was recoded from discrete categorical variables. Both sex (male = 0, female = 1) and race (black = 1, white = 0) were also binary variables. Callahan and Wolinsky examined the factor structure of the CES-D due not only to sex, but also due to race/ethnicity in an older population, showing that five items contributed to disparities (Callahan and Wolinsky, 1994). The first study to simultaneously examine age, sex, and race/ethnic bias in the CES-D for an older population by using an extension of the Mantel–Haenszel method (Cole et al., 2000). Cole and colleagues (2000) found that older Blacks were more prone to endorse interpersonal symptoms in the CES-D than older whites; while older women, regardless of race/ethnicity, were likely to endorse the “crying” symptom, which was consistent with earlier findings (c.f., Stommel et al., 1993). Eliminating five CES-D symptoms with sex bias helped reduce the mean difference in levels of depressive symptomatology between adult men and women (Stommel et al., 1993).

2.3. Statistical analysis

The analytic approach involved four main stages. We first conducted an exploratory factor analysis (EFA) to explore the dimensions of the CES-D measure and to examine its consistency with the four-factor first order model suggested in previous studies (Blazer et al., 1998; Hertzog et al., 1990; Iwata and Roberts, 1996; Knight et al., 1997; Radloff, 1977). We used the results to inform the construction of a bi-factor model in a confirmatory factor analysis (CFA). This sequence of the EFA preceding the CFA has been previously demonstrated by Edelen and Reeve (2007) for applying IRT modeling to a depression scale similar to the CES-D. We compared the factor loadings of the CES-D items on the general depression factor and the specific factors to determine the salience of secondary dimensions. For purposes of identifying the bi-factor model, the general and shared specific factors are orthogonal. We then added covariates to the model, implementing a structural equation model (SEM) approach to the IRT model for purposes of detecting measurement noninvariance.

2.3.1. Parameter estimation and model fit—All statistical analyses were conducted with Mplus software, version 5.1 (Muthén and Muthén, 1998–2008). For our EFA, CFA, and SEM models we use two approaches to polytomous dependent variables, one based on limited information weighted least squares mean and variance adjusted (WLSMV) that implements a multivariate probit regression model, equivalent to a two parameter normal ogive IRT model extended to polytomous dependent variables (a graded response model (Samejima, 1969)) and the other based on a full information maximum likelihood estimator with robust standard errors (MLR) that implements a two parameter logistic IRT model extended to polytomous items (Muthén and Muthén, 1998–2008).

We use two methods because each offers some special advantages, but both approaches allow for the inclusion of multiple covariates, adjustment for complex sampling weights, inclusion of the full sample in the context of missing data under an assumption that data that are missing at random. The limited information WLSMV approach provides fit statistics commonly found in the structural equation modeling literature and convenient and efficient algorithms for building measurement models for the purpose of detecting measurement noninvariance have been described (Jones, 2003; Muthén and Lehman, 1985; Yang and Jones, 2007). The fit statistics describe in various ways the ability of the model to account for the means/thresholds and covariances (among dependent variables) and regression parameters (dependent variables on covariates) observed in the sample data. We like to use the root mean square error of approximation (RMSEA) (Browne and Cudek, 1993; Muthén et al., 1998) and the comparative fit index (CFI) (Bentler and Chou, 1988; Muthén, 1998) and the overall model chi-square to guide model fit. The RMSEA provides a measure of discrepancy per model degree of freedom

and approaches zero as model fit improves. Browne and Cudek (1993) recommended rejecting models with RMSEA values greater than 0.1; Hu and Bentler (1998) suggested that values close to 0.06 or less indicate adequate model fit. The CFI provides a goodness-of-fit measure based on a normed sample index with a range between 0 and 1 (Gerbing and Anderson, 1993). CFI values greater than 0.95 generally indicate adequate model fit (Bentler, 1990; Muthén, 1989). The model χ^2 is a test of the fit of the model to the data, with p -values greater than 0.05 indicating good fit. This test of model fit is generally recognized as over-powered in large sample sizes and very sensitive to departures from multivariate normality, however it is an indicator of the goodness or badness of fit in that large values imply good fit and low values imply poor fit (Jöreskog and Sörbom, 1982). Nested model comparisons can be approached with a χ^2 difference test, but depending upon the scale of the dependent variables and parameter estimator option used, special procedures may be required (Muthén and Muthén, 1998–2008). We describe the model building strategy below.

The full information maximum likelihood multivariate logistic regression approach offers other advantages. The most important for this study include the computation of standard errors for standardized regression parameters and the ability to express regression parameters as odds ratios (ORs). Following Cole and colleagues (2000), ORs may be more interpretable measures of effect for DIF statistics, and describe the proportional difference in log odds of responding to a symptom question in an indicated direction across the latent trait spectrum.

2.3.2. Effects of covariates: model building—We evaluated the effects of background variables on the factors using the multiple indicators and multiple causes (MIMIC) model with a bi-factor measurement model. This is essentially a bi-factor model with covariates, which can influence the general and specific factors. This model includes only the so-called indirect effects (Jöreskog and Goldberger, 1975): regressions of the latent factors on background variables. Consistent with a previously published study on examining DIF in the bi-factor model (Yang and Jones, 2008), we will refer to indirect effects as general effects, which are the relationships between the covariates and observed dependent variables (latent factor indicators) mediated by latent factors. We examined the total general effects in a unidimensional MIMIC model (Model 1) and in a bi-factor MIMIC model (Model 2).

The modeling framework for this and the following stage was multivariate probit using the Mplus WLSMV estimator. Building upon the bi-factor MIMIC model (Model 2), a forward stepwise model building procedure was used to identify significant direct effects between covariates and observed dependent variables. Using WLSMV Mplus provides modification indices (MI), model derivatives scaled as chi-square for the direct effects: regressions of CES-D items on age, sex, and race (regressions constrained to zero in the baseline model). The MI can be used to identify direct effects that if freely estimated would significantly improve model fit. Our algorithm for forward stepwise identification of items with DIF involves identifying the largest MI, relaxing the model constraint forcing this parameter to zero in the baseline model, and estimating a new set of parameters. All items that are not regressed on covariates are anchor items, and the anchor item set changes with each iteration, each time including a smaller set of items as more and more DIF effects are identified. To evaluate the significance of the direct effect, we performed a robust chi-square model difference test using the Mplus DIFFTEST (difference testing) function suitable for testing nested models with the WLSMV estimator (Muthén and Muthén, 1998–2008; Yang and Jones, 2007). The stepwise procedure continued until no significant direct effects were found, based on the DIFFTEST procedure.

As a third stage, the final bi-factor model with total effects (includes both general and DIF effects) due to the background variables was then re-estimated with the MLR estimator and Monte Carlo Integration to obtain standardized effects with standard errors and regression parameters suitable for expressing as ORs. A rule of thumb to determine if DIF is “relatively

large” and meaningful measurement bias suggested by Cole and colleagues (2000) is any proportional OR > 2.0 and <0.5. The rationale for this modeling strategy is that proportional odds ratios are more understandable to a public health research audience.

In the fourth stage, for each item with DIF, we examined the total effects from the covariate through the general and specific factors to the item using unstandardized estimates. This is a multiple mediator model, the magnitude of analysis is to show the mediating effect of the latent variable, by taking the absolute values of the negative effects to calculate the proportion of the magnitude and source of DIF is mediated by either the general or specific factors (Alwin and Hauser, 1975; Davis, 1985; MacKinnon, 2000; MacKinnon et al., 2000). This is a new approach to examine DIF effects using MODEL INDIRECT function in Mplus. We also obtained estimates for the total general effects (total indirect effects) and DIF effects (direct effects), from the covariate to the item with DIF; the general or indirect effect from the covariate through the general factor to the item with DIF; and the specific or indirect effect from the covariate through the specific factor to the item with DIF.

3. Results

The results from the EFA ($\chi^2 = 53.67$, degrees of freedom (df) = 20, $p < 0.001$; CFI = 0.99; RMSEA = 0.02) and CFA (Table 3) inform us of the following four factors were present in this sample and administration of the CES-D: *depressed affect*, *low positive affect*, *somatic complaints*, and *interpersonal difficulties*. These four factors are consistent with the findings from previous studies (Blazer et al., 1998; Hertzog et al., 1990; Iwata and Roberts, 1996; Knight et al., 1997; Radloff, 1977). The *depressed affect* factor consists of the following five items: “blues,” “depressed,” “lonely,” “cry,” and “sad.” Five items make up the *low positive affect* factor, which are: “good,” “hopeful,” “happy,” “enjoy life,” and “failure.” The *somatic complaints* factor comprises the largest number of items, which are the following eight depressive symptoms: “bothered,” “eating,” “mind,” “effort,” “sleep,” “talked,” “not get going,” and “fearful.” The two items left are “unfriendly” and “dislike,” which belong to the *interpersonal difficulties* factor.

After estimating the bi-factor model (Fig. 3 and Table 3), we found that the factor loadings for each item were greater on the general depressive factor than on each specific factor, except for the items in the *interpersonal difficulties* factor. Therefore, the *interpersonal difficulties* factor is considered a nuisance factor, sharing partial variation with the general depressive factor, but greater variation independent of the general depressive factor. Both “people were unfriendly” and “people dislike me” items were perhaps captured perceived interpersonal conflict rather than a subdomain of depression.

Table 4 shows similar fully standardized parameter estimates for the general depressive factor across the following three models: unidimensional MIMIC model (Model 1), bi-factor MIMIC model with general effects (Model 2), and bi-factor MIMIC model with total effects (Model 3). As for the subdomains, the standardized parameter estimates were also very close for the *depressed affect* factor, *interpersonal difficulties* factor, and the *lack of positive affect* factor. However, the *somatic complaints* factor showed the largest differences when we included the DIF effects. The two items within the *somatic complaints* factor that remained significant after adjusting for age, sex, and race/ethnicity were “bothered” (fully standardized parameter estimate = 0.38, standard error (SE) = 0.01, $p < .01$) and “sleep” (fully standardized parameter estimate = 0.22, SE = 0.08, $p < .05$). The “bothered” and “sleep” items in Model 3 were, in respective order, a 7% and 21% effect size decrease from Model 2.

The standardized parameter estimates for the general and specific effects on the general and specific depressive factors, respectively, are shown in Table 5. Beginning with the differences

between the unidimensional MIMIC model (Model 1) and the bi-factor MIMIC model with only general effects (Model 2), when we include the specific factors in the model, general effects due to age and sex are no longer significant. There are specific effects found in Model 2 that reflect significant effects due to age, sex, and race/ethnicity. For example, the *depressed affect* factor for women is 0.78 standard deviation (S.D.) units higher than men. Upon including the DIF effects, there was a severe underestimation of overall depression in women when compared with men between Model 2 (fully standardized parameter estimate = -0.02 , $SE = 0.10$, n.s.) and Model 3 (fully standardized parameter estimates = 0.27 , $SE = 0.08$, $p < .01$). In Model 3, women were 0.32 S.D. units lower than men for endorsing the *interpersonal difficulties* factor while blacks were 0.71 S.D. units higher than whites on the *interpersonal difficulties* factor.

The final bi-factor MIMIC model with DIF effects (Model 3) showed measurement noninvariance in items attributable to age, sex, and race/ethnicity on both the item and specific factor levels of the CES-D bi-factor model. Compared with whites, older blacks or African Americans are more likely to endorse the *interpersonal* items, as well as “eating,” “hopeful,” and “effort” items. Older women are more likely to admit that they “feel like crying,” “alked less than usual,” and “everything was an effort,” but less likely to have “felt that life is a failure,” when compared with men. Adults over the age of 75, when compared with their younger counterparts, are more prone to endorse the following items: “mind,” “hopeful,” and “failure.”

Upon examining the odds ratios of the direct effects, older women had a higher proportional odds than men for endorsing the “crying” item (odds ratio [OR] = 1.86 [95% confidence interval [CI]: $1.17, 2.96$) and lower proportional odds than men for endorsing the “failure” item (OR = 0.63 CI: $0.42, 0.94$). Those who were 75 years and older also had a lower proportional odds for endorsing the “failure” item (OR = 0.65 , 95% CI: $0.43, 0.97$). However, applying the rule of meaningful significant DIF by Cole and colleagues (2000), we find that the measurement bias found in both the “crying” and “failure” items attributable to women and the “failure” item attributable to those aged 75 and older are trivial after controlling for the underlying depressive symptoms experienced by women, which were higher than for men.

We suspect that the effects of measurement bias are trivial because Cole’s rule for significant DIF in the unidimensional context might be too high in the bi-factor model. That is because Cole’s rule was developed in a unidimensional framework, and in the bi-factor framework what is DIF is decomposed into a portion of DIF shared among a set of items that share residual covariates (the specific factor) and unique DIF. We therefore calculated the magnitudes of the general and specific effects through the general and specific factors, respectively based on the standardized parameter estimates for categorical data. The total DIF effect from age to “failure” is 72% of the total effect. That is, a majority of the variance in the “failure” item is due to an independent effect of age rather than general depression. We found no specific effect of age through the *low positive affect* factor to the “failure” item. The proportion of the DIF effects from sex to “crying” and “failure” are 54% and 51%, respectively, of the total effect. The specific effects from sex through the *depressed affect* and *low positive affect* factors were much lower at 2% of the total effects for each item. This shows that the majority of the DIF effect for the “failure” item comes from age. Of the total effect, the proportion of the DIF effect of sex directly on the “failure” item (51%) was slightly lower through the general effect, which is mediated by the general depressive factor (47%). The DIF effect of sex directly on “crying” was 54% of the total effect, which was slightly higher than the general effect (44%).

4. Discussion

This study examined measurement noninvariance in a diverse older population that informs us about underlying differences in the commonly used CES-D measure. The differences in

depressive symptom endorsement between whites and blacks, men and women, and those younger than age 75 and age 75 and older, contribute to our understanding of mental health disparities among older adults. Blacks, compared with whites, were more likely to endorse items loading on the *interpersonal difficulties* factor, which include “people dislike me” and “people are unfriendly towards me.” Women were less likely to endorse items loading on the *interpersonal difficulties* factor and the “failure” item than men; and consistent with previous findings, women had a higher proportional odds than men for endorsing the “crying” item (Callahan and Wolinsky, 1994). Those 75 years and older were less likely to endorse the “failure” item than those aged 65–74. However, the measurement noninvariance found in both the “crying” and “failure” items attributable to women and those aged 75 and older were trivial after controlling for the underlying depressive symptoms based on our magnitude and mediation analyses.

Previous work conducted on the bi-factor model (Reise et al., 2007; Gibbons and Hedeker, 1992; Gibbons et al., 2007) leaves room for further research on examining DIF in psychiatric scales that are not unidimensional. We studied a widely used depression screening scale to advance this method. The bi-factor model yields a measure of overall severity on an underlying latent dimension within a multidimensional IRT model, and also includes specific mental health domains that may be relevant to the patient. The examination of DIF in a bi-factor framework provides additional information about DIF detected in a unidimensional modeling framework. We can determine if a DIF effect reflects a unique relationship between a covariate and an individual item, or if the DIF effect is better explained as a relationship between a covariate and a sub-set of items that share covariance after accounting for a single underlying general factor (the specific factor). Specific factors may capture important phenomenological features of the syndrome (e.g., somatic symptoms), or methods artifacts (e.g., positively worded depression items). Researchers interested in measurement bias are cautioned to infer bias based on the results of statistical procedures to detect DIF in conjunction with thoughtful analysis with content experts (Camilli and Shepard, 1994). That advice continues to apply in the context of bias and DIF in the bi-factor model case.

In sum, we have extended the bi-factor model to examine measurement noninvariance showing the importance of this method for assessing the validity of a widely used depression instrument in a diverse older population. We found, as have others, that *interpersonal* items are biased against blacks (Cole et al., 2000; Yang and Jones, 2007). Moreover, a substantial fraction of variance in “failure” and “crying” items are due to the DIF effects of background variables rather than general or specific aspects of depression. For psychiatric scales that are multidimensional, the detection of DIF in a bi-factor item response theory model will help refine and improve the accuracy of computerized adaptive testing (CAT) to automatically tailor and streamline assessments to a diverse older population based on their responses and sociodemographic characteristics as demonstrated for psychiatric scales in adult populations by Gibbons and colleagues (Gibbons et al., 2008, 2009).

There are some limitations of our study that deserve comment. First, the sample is somewhat dated and reflects a narrow geographical region, despite being a randomly selected community-dwelling sample of older adults. Second, the data used in the study were collected in 1982, and the cohort of older adults at that time were born during the “Lost Generation” (1883–1900) and “GI Generation” (1901–1919) (Strauss and Howe, 1991). These generations of older adults may have different historical and cultural experiences than later generations of older adults, which may affect their interpretation of CES-D items and cause measurement bias. Third, a possible disadvantage of the single group MIMIC is that it does not detect non-uniform DIF, but will only detect non-uniform DIF as (uniform) DIF if the result of the DIF is to shift the conditional probability of an indicated response across groups.

The overall issues of measurement and health disparities are at the forefront of research agendas for the Institute of Medicine (IOM), NIH, and the Department of Health and Human Services (DHHS). The IOM has recommended that the NIH continue evaluating health measures to shed light on the inappropriate and unfair measures of health status among Americans (Institute of Medicine, 2006). The US Department of Health and Human Services (2000) has also noted that differences are found in the detection and treatment of depression by age, sex, and race/ethnicity. Examining health measures for DIF due to sociodemographic characteristics inform researchers and clinicians regarding instruments that are appropriate, fair, and culture-free before embarking on data collection, health assessments, or treatments that may result in costly and biased outcomes. A plausible explanation for the variation in the incidence and prevalence rates for depression in the population is influenced by how depressive symptoms are perceived, expressed, and reported in these groups of older adults (Williams et al., 2007).

Based on the National Comorbidity Survey Replication (NCS-R) study, Kessler and colleagues (2003) found that between 1990 and 2002, only 484,700 to 525,400 Americans received depression treatment among the 13.1 to 14.2 million Americans (aged 18 and older) who suffer from 12-month major depressive disorder (MDD). Furthermore, of those who received treatment in this nationally representative sample, only 104,696 to 113,487 received adequate treatment for depression. Other studies, including the Epidemiological Catchment Area (ECA) (Robins and Regier, 1991) study (aged 18 and older) and the National Comorbidity Survey (NCS) (Kessler et al., 1994) (aged 18–54), found even higher prevalence estimates and lower incidence for treatment of MDD. Furthermore, the cumulative lifetime prevalence for MDD increased linearly with a steeper slope in successively more recent cohorts, defined by the following four age groups: 18–29, 30–44, 45–59, and 60 and older. Based on these findings, it is expected that with each cohort experiencing greater longevity and diversity in the US, depressive symptoms will become more serious and highly prevalent in the aging population. Therefore, there is increased importance for accurate and adaptive depression screening among diverse older Americans (Beyer, 2007; Kessler et al., 2003) to provide adequate and improved treatment for depression.

Acknowledgment

The authors are grateful for the research support of Sarah McGarry and Kitty Wong for their respective preparation of the literature review, analytic support, and proof-reading of the manuscript. Special thanks go to Dr. Jeanne A. Teresi for her comments, reviews, and editing of the manuscript. Additional thanks for the thoughtful comments made by two anonymous reviewers.

Role of funding source

This research was made possible through the National Institutes of Health (NIH)/National Institute on Aging (NIA) 5-T32 AG023480 award, NIH/NIA 5 R01 AG025308-02, NIH/NIA 5 R01 AG025308-02 Diversity Supplement, NIH/NIA P60AG008812, Harvard Medical School Livingston Fellowship, and the NARSAD Young Investigator Award.

References

- Alwin DF, Hauser RM. The decomposition of effects in path analysis. *American Sociological Review* 1975;40:37–47.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders, fourth edition (DSM-IV) text revision. Washington, DC: American Psychiatric Association; 1994.
- Barreto RM, Segal SP. Use of mental health services by Asian Americans. *Psychiatric Services* 2005;56:746–748. [PubMed: 15939954]
- Bentler, P.; Chou, C. Practical issues in structural equation modeling. In: Long, J., editor. *Common problems/proper solutions: avoiding error in quantitative research*. Newbury Park, CA: Sage; 1988.
- Bentler PM. Comparative fit indexes in structural models. *Psychological Bulletin* 1990;107:238–246. [PubMed: 2320703]

- Beyer JL. Managing depression in geriatric populations. *Annals of Clinical Psychiatry* 2007;19:221–238. [PubMed: 18058280]
- Blazer D, Landerman L, Hays J, Simonsick E, Saunders W. Symptoms of depression among community-dwelling elderly African American and white older adults. *Psychological Medicine* 1998;28:1311–1320. [PubMed: 9854272]
- Brown, TA. Confirmatory factor analysis for applied research. New York: Guilford Publications; 2006.
- Browne, M.; Cudek, R. Alternative ways of assessing model fit. In: Bollen, K.; Long, J., editors. *Testing structural equation models*. Thousand Oaks, CA: Sage; 1993. p. 136-162.
- Callahan CM, Wolinsky FD. The effect of gender and race on the measurement properties of the CES-D in older adults. *Medical Care* 1994;32:341–356. [PubMed: 8139299]
- Camilli, G.; Shepard, LA. *Methods for identifying biased test items*. Newbury Park, California: Sage Publishers; 1994.
- Chan KS, Orlando M, Ghosh-Dastidar B, Duan N, Sherbourne CD. The interview mode effect on the Center for Epidemiological Studies Depression (CES-D) scale: an item response theory analysis. *Medical Care* 2004;42:281–289. [PubMed: 15076828]
- Chen FF, Sousa KH, West SG. Testing measurement invariance of second-order factor models. *Structural Equation Modeling* 2005;12:471–492.
- Chen FF, West SG, Sousa KH. A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research* 2006;41:189–225.
- Cole SR, Kawachi I, Maller SJ, Berkman LF. Test of item-response bias in the CES-D scale. Experience from the New Haven EPESE Study. *Journal of Clinical Epidemiology* 2000;53:285–289. [PubMed: 10760639]
- Cook BL, McGuire T, Miranda J. Measuring Trends in Health Disparities, 2000–2004. *Psychiatric Services* 2007;58:1533–1540. [PubMed: 18048553]
- Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. *Statistical Medicine* 2004;23:241–256.
- Davis, JA. The logic of causal order. In: GE, SA., editor. *SAGE Quantitative Psychology Series Paper 55*. CA: Newbury Park: 1985. p. 56-58.
- Dobalian A, Rivers PA. Racial and ethnic disparities in the use of mental health services. *Journal of Behavioral Health Services & Research* 2008;35:128–141. [PubMed: 18074230]
- Edelen MO, Reeve BB. Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research* 2007;16:5–18. [PubMed: 17375372]
- Finch H, French BF. Anomalous type I error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement* 2008;68:742–759.
- Foley K, Reed P, Mutran E, DeVellis R. Measurement adequacy of the CES-D among a sample of older African Americans. *Psychiatry Research* 2002;109:61–69. [PubMed: 11850052]
- Gallo JJ, Anthony JC, Muthén BO. Age differences in the symptoms of depression: a latent trait analysis. *Journals of Gerontology, Psychological Sciences* 1994;49:251–264.
- Gallo JJ, Rabins PV. Depression without sadness: alternative presentations of depression in late life. *American Family Physician* 1999;60:820–826. [PubMed: 10498109]
- Gerbing, DW.; Anderson, JC. Monte carlo evaluations of goodness-of-fit indices. In: Bollen, KA.; Long, JS., editors. *Wiley series in probability and mathematical statistics*. New York: John Wiley and Sons; 1993. p. 294-316.
- Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, et al. Full-Information Item Bifactor Analysis of Graded Response Data. *Applied Psychological Measurement* 2007;31:4–19.
- Gibbons RD, Hedeker D. Full-information item bi-factor analysis. *Psychometrika* 1992;57:423–436.
- Gibbons RD, Rush AJ, Immekus JC. On the psychometric validity of the domains of the PDSQ: an illustration of the bi-factor item response theory model. *Journal of Psychiatric Research* 2009;43:401–410. [PubMed: 18554611]
- Gibbons RD, Weiss DJ, Kupfer DJ, Frank E, Fagioli A, Grochocinski VJ, et al. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services* 2008;59:361–368. [PubMed: 18378832]

- Gregorich SE. Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care* 2006;44:S78–S94. [PubMed: 17060839]
- Gustafsson J, Balke G. General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research* 1993;28:407–434.
- Hauser, RM.; Goldberger, AS. The treatment of unobservable variables in path analysis. In: Costner, HL., editor. *Sociological Methodology*. San Francisco: Jossey-Bass; 1971. p. 81-117.
- Hertzog C, Van Alstine J, Usala P, Hultsch D, Dison R. Measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D) in older populations. *Psychological Assessment* 1990;2:64–72.
- Hu L, Bentler P. Fit indices in covariance structure analysis: sensitivity to underparameterized model misspecifications. *Psychological Methods* 1998;4:424–453.
- Huguet N, Kaplan MS, Feeny D. Socioeconomic status and health-related quality of life among elderly people: results from the Joint Canada/United States Survey of Health. *Social Science and Medicine* 2008;66:803–810. [PubMed: 18155337]
- Institute of Medicine. Board on health sciences policy. Washington, DC: The National Academies; 2006. Examining the health disparities research plan of the national institutes of health: unfinished business.
- Iwata N, Roberts RE. Age differences among Japanese on the Center for Epidemiologic Studies Depression Scale: an ethnocultural perspective on somatization. *Social Science and Medicine* 1996;43:967–974. [PubMed: 8888466]
- Jones RN. Racial bias in the assessment of cognitive functioning of older adults. *Aging and Mental Health* 2003;7:83–102. [PubMed: 12745387]
- Jöreskog K, Goldberger A. Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* 1975;10:631–639.
- Jöreskog KG, Sörbom D. Recent Developments in Structural Equation Modeling. *Journal of Marketing Research* 1982;19:404–416.
- Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association* 2003;289:3095–3105. [PubMed: 12813115]
- Kessler RC, McGonagle KA, Zhao S, Nelson CB, Hughes M, Eshleman S, et al. Results from the National Comorbidity Survey. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. *Archives of General Psychiatry* 1994;51:8–19. [PubMed: 8279933]
- Kleinman, A. *Rethinking psychiatry*. New York: The Free Press; 1988.
- Knight RG, Williams S, McGee R, Olanoff S. Psychometric properties of the Centre for Epidemiologic Studies Depression Scale (CES-D) in a sample of women in middle life. *Behaviour Research and Therapy* 1997;35:373–380. [PubMed: 9134792]
- Lipton RB, Katz MJ, Kuslansky G, Sliwinski MJ, Stewart WF, Verghese J, et al. Screening for dementia by telephone using the memory impairment screen. *Journal of the American Geriatric Society* 2003;51:1382–1390.
- Lord F. A theory of test scores. *Psychometric Monographs* 1952;7:84.
- MacKinnon, DP. Contrasts in multiple mediator models. In: Rose, J.; Chassin, L.; Presson, CC.; Sherman, SJ., editors. *Multivariate applications in substance use research: new methods for new questions*. Mahwah, NJ: Erlbaum; 2000. p. 141-160.
- MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding, and suppression effect. *Prevention Science* 2000;1:173–181. [PubMed: 11523746]
- Manly JJ. Deconstructing race and ethnicity: implications for measurement of health outcomes. *Medical Care* 2006;44:S10–S16. [PubMed: 17060816]
- Martin LG, Schoeni RF, Freedman VA, Andreski P. Feeling better? Trends in general health status. *Journals of Gerontology, Psychological Sciences* 2007;62:S11–S21.
- Mazor KM, Hambleton RK, Clauser BE. Multidimensional DIF analyses: the effects of matching on unidimensional subtest scores. *Applied Psychological Measurement* 1998;22:357–367.
- McHorney CA. Ten recommendations for advancing patient-centered outcomes measurement for older persons. *Annals of Internal Medicine* 2003;139:403–409. [PubMed: 12965966]

- Mojtabai R, Olfson M. Major depression in community-dwelling middle-aged and older adults: prevalence and 2- and 4-year follow-up symptoms. *Psychological Medicine* 2004;34:623. [PubMed: 15099417]
- Mui A, Burnette D, Chen L. Cross-cultural assessment of geriatric depression: a Review of the CES-D and the GDS. *Journal of Mental Health and Aging* 2001;7
- Muthén, B. The development of heavy drinking and alcohol related problems from ages 18 to 37 in a US national sample. Los Angeles: Graduate School of Education and Information Studies; 1998.
- Muthén, B.; Khoo, S-T.; Francis, D. Multi-stage analysis of sequential developmental processes to study reading progress: new methodological developments using general growth mixture modeling. Los Angeles: UCLA Graduate School of Education and Information Studies; 1998.
- Muthén B, Lehman J. Multiple group IRT modeling: applications to item bias analysis. *Journal of Educational Statistics* 1985;10:133–142.
- Muthén BO. Latent variable modeling in heterogeneous populations. *Meetings of Psychometric Society* (1989, Los Angeles, California and Leuven, Belgium). *Psychometrika* 1989;54:557–585.
- Muthén, LK.; Muthén, BO. Mplus Version 5.1. Los Angeles: Muthén and Muthén;; 1998–2008.
- National Center for Health Statistics. Hyattsville. Maryland: Public Health Service; 2001. Healthy people 2000 final review.
- Radloff L. The CES-D scale: a self-report depression scale for research in the general population. *Applied Psychological Measurement* 1977;1:385–401.
- Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute of Educational Research. 1960
- Reise S, Morizot J, Hays RD. The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research* 2007;16:19–31. [PubMed: 17479357]
- Robins, LN.; Regier, DA. Psychiatric disorders in America: the epidemiologic catchment area study. New York, NY: The Free Press; 1991.
- Ross C, Mirowsky J. Components of depressed mood in married men and women: the Center for Epidemiologic Studies Depression Scale. *American Journal of Epidemiology* 1984;119:997–1004. [PubMed: 6731436]
- Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement* 1969;34:100.
- Satorra, A. Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In: Heijmans, DDH.; Pollock, DSG.; Satorra, A., editors. *Innovations in multivariate statistical analysis: A festschrift for Heinz Neudecker*. Dordrecht: Kluwer Academic Publisher; 2000. p. 233–247.
- Schmid J, Leiman J. The development of hierarchical factor solutions. *Psychometrika* 1957;5:141–152.
- Stahl SM, Hahn AA. The National Institute on Aging's Resource Centers for Minority Aging Research. Contributions to measurement in research on ethnically and racially diverse populations. *Medical Care* 2006;44:S1–S2. [PubMed: 17060815]
- Steffick, D. Documentation of affective functioning measures in the Health and Retirement Study. Ann Arbor, MI: Survey Research Center, University of Michigan; 2000. available at <<http://www.umich.edu/hrswww/docs/userg/index.html>>
- Stommel M, Given BA, Given CW, Kalaian HA, Schulz R, McCorkle R. Gender bias in the measurement properties of the Center for Epidemiologic Studies Depression Scale (CES-D). *Psychiatry Research* 1993;49:239–250. [PubMed: 8177918]
- Strauss, W.; Howe, N. Generations: the history of America's Future, 1584 to 2069. New York: Perennial; 1991.
- Taylor, JO.; Wallace, RB.; Ostfeld, AM.; Blazer, DG.; Fake, N. Established populations for epidemiologic studies of the elderly. 1981–1993: (East Boston, Massachusetts, Iowa and Washington Counties, Iowa, New Haven, Connecticut, and North Central North Carolina). [Computer File]. 3rd ICPSR version. Bethesda, MD: National Institute on Aging (producer), 1997. Ann Arbor, Michigan: Interuniversity Consortium for Political and Social Research (distributor); 1998.
- Teresi JA. Different approaches to differential item functioning in health applications. Advantages, disadvantages and some neglected topics. *Medical Care* 2006a;44:S152–S170. [PubMed: 17060822]

- Teresi JA. Overview of quantitative measurement methods. Equivalence, invariance, and differential item functioning in health applications. *Medical Care* 2006b;44:S39–S49. [PubMed: 17060834]
- Teresi JA, Stewart AL, Morales LS, Stahl SM. Measurement in a multi-ethnic society. Overview to the special issue. *Medical Care* 2006;44:S3–S4. [PubMed: 17060831]
- US Department of Health and Human Services. Healthy people 2010: understanding and improving health. Vol. 2nd ed. Washington, DC: US Government Printing Office;
- Williams CD, Taylor TR, Makambi K, Harrell J, Palmer JR, Rosenberg L, et al. CES-D four-factor structure is confirmed, but not invariant, in a large cohort of African American women. *Psychiatry Research* 2007;150:173–180. [PubMed: 17291596]
- Yang FM, Jones RN. Center for Epidemiologic Studies-Depression Scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *Journal of Clinical Epidemiology* 2007;60:1195–1200. [PubMed: 17938063]
- Yang FM, Jones RN. Measurement differences in depression: chronic health-related and sociodemographic effects in older Americans. *Psychosomatic Medicine* 2008;70:993–1004. [PubMed: 18981269]
- Zimmerman M, Mattia JI. A self-report scale to help make psychiatric diagnoses: the Psychiatric Diagnostic Screening Questionnaire. *Archives of General Psychiatry* 2001;58:787–794. [PubMed: 11483146]

Fig. 1.

This is a depiction of a generic unidimensional scale with a covariate, which is shown as the box marked with an “x.” This is called a multiple indicator, multiple causes (MIMIC) model with one general construct (g) made up of five factor indicators (y_1 – y_5). There is a DIF effect found in the direct effect from the covariate (x) to the third item (y_3) in the scale. The general effect is the indirect effect between x and y_3 mediated by the general latent factor (g).

Fig. 2.

The bi-factor multiple indicator, multiple causes (MIMIC) model adds one specific construct (s), measured by three items of the scale (y_1 – y_3), to the generic unidimensional MIMIC model shown in Fig. 1. The general effect is the indirect effect between the covariate (x) and item y_3 mediated by the general latent factor (g). The DIF effect is the direct effect from x to y_3 . The unidimensional DIF effect is decomposed in the bi-factor model into a specific effect and a (residual) DIF effect. The specific effect is the pathway from x to s to y_3 .

Fig. 3.

This is a depiction of the bi-factor item response theory for the 20-item Center for Epidemiological Studies-Depression (CES-D) scale. This schematic conceptual representation uses the structural equation modeling path diagram to show the latent factors (depicted by oval shapes) that include the general depression factor and the specific factors of depressed affect, interpersonal difficulties, low positive affect, and somatic complaints. The latent factor indicators are depicted by rectangles, which are the 20 CES-D items asked of each subject.

Table 1Demographic characteristics of the New Haven Established Populations for Epidemiologic Studies ($N = 2,733$).

Demographic Characteristics		<i>N</i>	%
Age (years)	65–69	962	(34.2)
	70–74	768	(27.3)
	75–79	491	(17.5)
	80 to 84	335	(11.9)
	85+	256	(9.1)
Sex	Male	1,169	(41.6)
	Female	1,643	(58.4)
Race/ethnicity	White	2,283	(81.2)
	Black	529	(18.8)
Education (grade)	<8th	812	(29.8)
	8th–11th	1,045	(38.4)
	12th	494	(18.1)
	>12th	374	(13.7)
Marital Status	Married	1,046	(41.5)
	Not Married	1,475	(58.5)

Table 2

Item response frequencies as percentages, means, standard deviations, and item-scale correlations for the New Haven EPESE CES-D items (N = 2733).

Item	Response frequencies (as percentages)					Mean	Standard deviation	Item-scale correlation
	Rarely or none of the time	Some of the time	Much of the time	Most or all of the time				
(1) Bothered by things	68.8	19.2	2.5	5.8	1.43	0.81	0.52	
(2) Poor appetite	75.3	12.9	2.6	6.3	1.38	0.83	0.48	
(3) Could not shake the blues	74.6	14.4	3.1	4.2	1.34	0.74	0.64	
(4) As good as others	80.5	3.8	5.5	4.0	1.29	0.76	0.37	
(5) Trouble keeping mind on task	69.2	19.7	3.0	4.6	1.41	0.77	0.48	
(6) Felt depressed	61.4	25.8	3.6	5.8	1.52	0.83	0.72	
(7) Everything an effort	65.5	18	3.6	8.9	1.54	0.94	0.60	
(8) Felt hopeful	51.4	6.2	14.3	18.4	2.00	1.24	0.48	
(9) Life a failure	83.0	8.3	1.3	2.8	1.20	0.61	0.48	
(10) Felt fearful	77.5	14.1	1.4	3.3	1.28	0.66	0.50	
(11) Restless sleep	63.1	19	3.3	11.1	1.61	1.00	0.53	
(12) Happy	63.5	9.6	15.6	7.1	1.65	1.00	0.64	
(13) Talked less	74.8	12.3	3.0	5.4	1.36	0.79	0.44	
(14) Felt lonely	64.0	21.2	3.6	7.7	1.53	0.90	0.63	
(15) People unfriendly	81.1	9.8	1.5	3.3	1.24	0.65	0.36	
(16) Enjoyed life	72.7	7.2	10.3	5.6	1.47	0.90	0.62	
(17) Crying spells	82.8	10	1.6	2.1	1.20	0.57	0.49	
(18) Felt sad	62.9	26.4	3.4	3.6	1.46	0.74	0.67	
(19) People disliked me	85.8	7.2	0.7	1.6	1.14	0.49	0.35	
(20) Could not get going	72.2	16.6	2.9	4.1	1.36	0.74	0.55	
Total raw score Mean & S.D.	28 (SD 8.28)							
Skewness	1.62							
Kurtosis	6.24							
Coefficient alpha	0.86							

Table 3
Model fit statistics for unidimensional, second-order (hierarchical), and bi-factor models with and without covariates ($n = 2733$).

Model	Limited information multivariate probit (WLSMV) fit statistics			Full information multivariate logit (MLR) fit statistics			
	χ^2	df	CFI	RMSEA	Log-likelihood	Scaling Factor	Free Parameters
Models without covariates							
1 Unidimensional model	184.3	18	0.947	0.058	-35217	1.682	80
2 Second order (hierarchical) model	55.9	18	0.988	0.028	-34841	1.686	83
3 Bi-factor model	52.0	19	0.989	0.025	-34804	1.754	99
Models with covariates							
4 Unidimensional model without DIF effects	154.5	19	0.950	0.051	-35177	1.680	83
5 Unidimensional model with DIF effects ^a	151.5	19	0.951	0.051	-35123	1.701	86
6 Bi-factor model without specific & DIF effects	61.8	20	0.985	0.028	-34767	1.768	102
7 Bi-factor model with specific but without DIF effects	49.9	19	0.989	0.024	-34700	1.765	114
8 Bi-factor model with specific and DIF effects	48.6	20	0.989	0.023	-34695	1.748	117

χ^2 model chi-square; df, degrees of freedom; CFI, comparative fit index; RMSEA, root mean squared error of approximation; Scaling factor is due to Satorra (2000); AIC, Akaike information criterion; BIC, sample size adjusted Bayesian information criterion.

^aReported in Yang and Jones (2007).

Table 4

Parameter estimates^a and mean differences for the latent depression measurement model for older adults in the New Haven Established Population Epidemiological Studies on the Elderly ($N = 2733$).

General	Model 1		Model 2		Model 3	
	Unidimensional model		Bi-factor model with group effects		Bi-factor model with DIF effects	
	Standardized Parameter Estimate	SE	Standardized Parameter Estimate	SE	Standardized Parameter Estimate	SE
Bothered	0.60	*** (0.03)	0.56	*** (0.03)	0.57	*** (0.04)
Eating	0.56	*** (0.03)	0.55	*** (0.04)	0.57	*** (0.05)
Blues	0.80	*** (0.02)	0.76	*** (0.03)	0.76	*** (0.02)
Good	0.49	*** (0.04)	0.44	*** (0.04)	0.45	*** (0.04)
Mind	0.56	*** (0.02)	0.53	*** (0.03)	0.57	*** (0.02)
Depressed	0.85	*** (0.01)	0.82	*** (0.02)	0.81	*** (0.01)
Effort	0.67	*** (0.02)	0.65	*** (0.03)	0.68	*** (0.04)
Hopeful	0.46	*** (0.03)	0.44	*** (0.03)	0.44	*** (0.03)
Failure	0.67	*** (0.03)	0.70	*** (0.03)	0.70	*** (0.03)
Fearful	0.65	*** (0.03)	0.66	*** (0.04)	0.69	*** (0.04)
Sleep	0.57	*** (0.03)	0.55	*** (0.03)	0.57	*** (0.04)
Happy	0.68	*** (0.03)	0.65	*** (0.03)	0.65	*** (0.04)
Talked	0.54	*** (0.03)	0.54	*** (0.03)	0.56	*** (0.03)
Lonely	0.70	*** (0.02)	0.65	*** (0.03)	0.67	*** (0.03)
Unfriendly	0.45	*** (0.03)	0.48	*** (0.03)	0.49	*** (0.03)
Enjoy	0.70	*** (0.02)	0.66	*** (0.02)	0.66	*** (0.02)
Cry	0.72	*** (0.02)	0.63	*** (0.03)	0.63	*** (0.03)
Sad	0.79	*** (0.02)	0.75	*** (0.03)	0.75	*** (0.03)
Dislike	0.56	*** (0.03)	0.60	*** (0.04)	0.60	*** (0.04)
Not get going	0.66	*** (0.03)	0.65	*** (0.02)	0.67	*** (0.03)
<i>Factor 1: Depressed Affect</i>						
Blues			0.34	*** (0.02)	0.31	*** (0.01)
Depressed			0.31	*** (0.04)	0.30	*** (0.04)

General	Model 1	Model 2		Model 3		
	Unidimensional model	Bi-factor model with group effects		Bi-factor model with DIF effects		
	Standardized Parameter Estimate	SE	Standardized Parameter Estimate	SE	Standardized Parameter Estimate	SE
Lonely			0.35	*** (0.05)	0.30	*** (0.06)
Cry			0.49	*** (0.06)	0.43	*** (0.06)
Sad			0.36	*** (0.04)	0.36	*** (0.06)
Factor 2: Interpersonal Difficulties						
Unfriendly			0.63	*** (0.02)	0.64	*** (0.02)
Dislike			0.58	*** (0.02)	0.59	*** (0.02)
Factor 3: Lack of Positive affect						
Good			0.44	*** (0.01)	0.43	*** (0.01)
Hopeful			0.34	*** (0.05)	0.34	*** (0.04)
Happy			0.47	*** (0.04)	0.46	*** (0.05)
Enjoy			0.53	*** (0.03)	0.53	*** (0.03)
Failure			0.12	* (0.05)	0.15	* (0.06)
Factor 4: Somatic complaints						
Bothered			0.42	*** (0.01)	0.39	*** (0.01)
Eating			0.26	*** (0.06)	0.19	(0.11)
Mind			0.25	*** (0.06)	0.11	(0.10)
Effort			0.33	*** (0.07)	0.25	(0.17)
Sleep			0.28	*** (0.06)	0.22	* (0.08)
Talked			0.14	* (0.07)	0.06	(0.13)
Notgoing			0.33	** (0.10)	0.27	(0.21)
Fearful			0.15	* (0.07)	−0.02	(0.11)

*
p < .05.

**
p < .01.

p < .001.

ρ Fully standardized measurement slopes (factor loadings) with regard to mean and variance depressive symptoms regressed on latent depressive factors.

Table 5

Parameter estimates^a and mean differences for latent depression structural model and sociodemographic group effects in the New Haven Established Population Epidemiological Studies on the Elderly ($N = 2733$).

Factors	Model 1		Model 2		Model 3	
	Unidimensional model		Bi-factor model with group effects		Bi-factor model with DIF effects	
	Standardized parameter estimate	SE	Standardized parameter estimate	SE	Standardized parameter estimate	SE
General						
Aged 75 years and older	0.19	(0.07)**	-0.06	(0.14)	0.16	(0.13)
Female	0.32	(0.05)***	-0.02	(0.10)	0.27	(0.08)**
Black	-0.01	(0.07)	-0.01	(0.22)	0.08	(0.15)
<i>Factor 1: Depressed Affect</i>						
Aged 75 years and older			0.42	(0.25)	0.02	(0.24)
Female			0.78	(0.19)***	0.22	(0.16)
Black			-0.24	(0.41)	-0.41	(0.30)
<i>Factor 2: Interpersonal Difficulties</i>						
Aged 75 years and older			0.21	(0.16)	0.03	(0.15)
Female			-0.09	(0.11)	-0.32	(0.09)***
Black			0.78	(0.21)***	0.71	(0.16)***
<i>Factor 3: Lack of Positive Affect</i>						
Age (centered at 75 years old)			0.22	(0.16)	-0.03	(0.15)
Female			0.40	(0.12)**	0.09	(0.10)
Black			-0.10	(0.29)	-0.19	(0.22)

Factors	Model 1	Model 2	Model 3
	Unidimensional model	Bi-factor model with group effects	Bi-factor model with DIF effects
	Standardized parameter estimate	Standardized parameter estimate	Standardized parameter estimate
	SE	SE	SE
<i>Factor 4: Somatic Complaints</i>			
Aged 75 years and older		0.60 (0.25)*	0.24 (0.31)
Female		0.53 (0.19)	0.01 (0.20)
Black		-0.02 (0.39)*	-0.23 (0.40)

* p < .05.
** p < .01.
*** p < .001.

^a Standardized regression coefficients with respect to the mean and variance of latent variables.