Published in final edited form as:

Psychol Assess. 2010 December ; 22(4): 912–922. doi:10.1037/a0020704.

Factor Structure and Psychometric Properties of the Brief Symptom Inventory–18 in Women: A MACS Approach to Testing for Invariance Across Racial/Ethnic Groups

Margit Wiesner,

University of Houston

Vincent Chen,

University of Texas Health Science Center

Michael Windle,

Emory University

Marc N. Elliott,

RAND Corporation, Santa Monica, California

Jo Anne Grunbaum,

Centers for Disease Control and Prevention, Atlanta, Georgia

David E. Kanouse, and

RAND Corporation, Santa Monica, California

Mark A. Schuster

Children's Hospital Boston/Harvard Medical School

Abstract

This study used data from 3 sites to examine the invariance and psychometric characteristics of the Brief Symptom Inventory–18 across Black, Hispanic, and White mothers of 5th graders ($N = 4,711$; $M = 38.07$ years of age, $SD = 7.16$). Internal consistencies were satisfactory for all subscale scores of the instrument regardless of ethnic group membership. Mean and covariance structures analysis indicated that the hypothesized 3-factor structure of the instrument was not robust across ethnic groups. It provided a reasonable approximation to the data for Black and White women but not for Hispanic women. Tests for differential item functioning (DIF) were therefore conducted for only Black and White women. Analyses revealed no more than trivial instances of nonuniform DIF but more substantial evidence of uniform DIF for 3 of the 18 items. After having established partial strong factorial invariance of the instrument, latent factor means were found to be significantly higher for Black than for White women on all 3 subscales (somatization, depression,

© 2010 American Psychological Association

Correspondence concerning this article should be addressed to Margit Wiesner, Department of Educational Psychology, University of Houston, 491 Farish Hall, Houston, TX 77204-5029. mfwiesner@uh.edu.

Margit Wiesner, Department of Educational Psychology, University of Houston; Vincent Chen, Division of Behavioral Sciences and Division of Biostatistics, University of Texas Health Science Center; Michael Windle, Department of Behavioral Sciences and Health Education, Emory University; Marc N. Elliott and David E. Kanouse, RAND Health, RAND Corporation, Santa Monica, California; Jo Anne Grunbaum, Prevention Research Centers Program, Division of Adult and Community Health, Centers for Disease Control and Prevention, Atlanta, Georgia; Mark A. Schuster, Department of Medicine, Children's Hospital Boston/Harvard Medical School.

anxiety). In conclusion, the instrument may be used for mean comparisons between Black and White women.

Keywords

BSI-18; mean and covariance structures analysis; race/ethnicity; distress; adults

The Brief Symptom Inventory–18 (BSI-18; Derogatis, 2000) was introduced in recent years as an abbreviated version of the Brief Symptom Inventory (BSI; Derogatis, 1993). The BSI consists of 53 items and assesses levels of psychological distress on nine dimensions. The much shorter BSI-18 was developed as a screening tool for identifying psychological distress on the basis of just the three most prevalent psychiatric syndromes, namely depression, anxiety, and somatization. It may be used with medical and community populations. Information about the validity and reliability of the BSI-18 is still limited because it is a relatively new instrument (Derogatis & Fitzpatrick, 2004). Notably lacking are studies of its psychometric characteristics and the invariance of its posited factor structure across racial/ethnic groups in the United States. This is especially unfortunate because of the ongoing demographic shifts toward a more ethnically diverse society (see U.S. Census Bureau, 2002). Rapidly growing ethnic minority groups such as Hispanics may differ from Whites in a number of ways, including cultural background, language use, experience of discrimination, and acculturative stress. Any of these factors may affect the response patterns in instruments designed to screen for psychological distress. Therefore, one cannot assume a priori that the BSI-18 is equally applicable to all racial/ethnic groups in the United States.

In order to adequately assess the levels of distress and, by extension, mental health needs of Hispanics and other racial/ethnic groups, one must understand the extent to which the psychometric properties and theoretical constructs (factors) of screening instruments for psychological distress generalize across differing racial/ethnic groups. The purpose of the current cross-sectional study was to apply a mean and covariance structures analysis (MACS) to the BSI-18 in order to address this limitation of the literature, using data gathered from a large community sample of White, Hispanic, and Black women.

Prior Research on the Factor Structure of the BSI-18

Psychometric characteristics and the factor structure of the BSI-18 have been examined in just a handful of studies. Derogatis (2000) applied principal components analysis to data from his normative community sample of 1,134 men and women, whose racial/ethnic composition and socioeconomic status were not described. The somatization and depression dimensions of the BSI-18 were replicated almost perfectly; however, the third hypothesized anxiety dimension split into a traditional anxiety component and a marginal (three items with an eigenvalue of exactly one) panic component. He cautioned that the “panic-related items may just as likely coalesce with general symptoms of anxiety” (p. 15) in other studies. Zabora et al. (2001) repeated these principal components analyses for 1,543 cancer patients (again, the racial/ethnic composition of the sample was not described). Although they also retained a four-component solution, the loading pattern was sometimes quite different from

that in the prior study. Moreover, support for a fourth component was far weaker (i.e., a single-item [suicidal thoughts] component with an eigenvalue of 0.98). Recklitis and colleagues (2006) noted that the weak support for an independent fourth dimension of the BSI-18, along with the inconsistency in its meaning across studies, suggests that it may be a product of overextraction.

Another study examined the factor structure of the BSI-18 for a sample of 100 Central American immigrants (Asner-Self, Schreiber, & Marotta, 2006). Principal components analysis revealed a strong general factor of psychological distress but offered relatively weak support for the hypothesized three dimensions. Two additional studies provided considerably stronger empirical evidence by pursuing a cross-validation approach, employing confirmatory factor analysis (CFA) techniques, and directly testing the fit of several different model specifications. Prelow, Weaver, Swenson, and Bowman (2005) used data from 1,115 Latinas to examine the factor structure of the BSI-18. Exploratory principal axis factor analysis revealed a single general factor for a randomly chosen derivation subsample. For the cross-validation subsample, CFA showed that both a single-factor model (with five estimated covariances among item residuals) and the hypothesized three-factor model fit the data reasonably well. Lack of discriminant validity among the three factors and considerations of model parsimony led the authors to choose the one-factor specification as the best model. Finally, Recklitis et al. (2006) reported findings for 8,945 cancer survivors who were predominantly European Americans. For the randomly chosen subsample, exploratory principal factor analysis supported the hypothesized three-dimensional structure quite well. For the cross-validation subsample, CFA indicated a good fit to the data for both the hypothesized three-factor model and an alternative four-factor model. The addition of a second-order factor to each model made little difference in terms of model fit. A single-factor model showed poorer fit to the data. The three-factor model was chosen as the best model. It was most consistent with the hypothesized factor structure of the BSI-18 and tenable for both genders according to subsequent invariance tests.

In sum, the few available empirical studies on the factor structure of the BSI-18 have yielded mixed results, only partially supporting its proposed three-factor structure. Some tentative evidence has suggested the existence of a fourth independent factor, but its practical significance appears to be limited, and it may not replicate reliably across large independent samples. In addition, questions remain about the discriminant validity among the three posited dimensions. The existing studies are further hampered by a number of methodological features. Three studies used data from a special population (cancer patients/survivors, Central American immigrants), which generates concerns about the generalizability of the findings to other segments of the general population. Three studies used principal components analysis, which analyzes total variance (i.e., variability due to unique item variance is retained) and is therefore well suited for data reduction purposes (Thompson, 2004) but does not examine the latent factor structure of a given instrument. None of these studies tested the invariance of the posited factor structure across different racial/ethnic groups. This has been recognized as a limitation in the literature (see e.g., Recklitis et al., 2006). In fact, Prelow et al. (2005) stressed the need to conduct stringent invariance tests of the BSI-18 via MACS analysis in order to determine whether the

theoretical constructs and the latent factor mean levels are similar across different racial/ethnic groups. Further empirical work on these issues is clearly needed.

Study Aims

The main purpose of the current study was to examine whether the theoretical constructs of the BSI-18 are equivalent across different racial/ethnic groups from a large U.S. community sample. A secondary goal was to provide information on the internal consistency of the instrument for the given racial/ethnic groups. Several authors have argued that Sörbom's (1974) MACS is especially well suited to address the primary question because it allows for (a) simultaneous model fitting of a hypothesized factorial structure in two or more groups; (b) tests of the cross-group equivalence of all reliable measurement parameters (i.e., both intercepts and factor loadings) while correcting for measurement error; and (c) tests of between-groups differences in primary construct moments, such as construct mean levels and covariances, via nested model comparisons in multigroup analyses (see e.g., Byrne & Stewart, 2006; Little, 1997). Note that tests of factorial invariance within traditional CFA models are capable of testing only "weak" forms of invariance (Meredith, 1993) because they rely on the analysis of covariance structures. Thus, they can be viewed as a special case of the more general MACS approach to measurement equivalence, which additionally includes means on observed and latent variables in the analysis and therefore can test "strong" forms of invariance (Meredith, 1993). This is an important extension because equality tests of latent factor means require evidence of strong factorial invariance.

It is also useful to consider the evidence of measurement invariance within a more integrative framework that relates the interpretation of findings from MACS to the interpretation of results from item response theory (IRT). When testing for strong factorial invariance in IRT modeling, analysts seek to identify differential item functioning (DIF), focusing especially on two parameters, item difficulty and item discrimination, both of which describe the link between a test item and its underlying latent factor (Widaman & Reise, 1997). Item difficulty can be described as a location parameter that relates the latent trait factor and the mean item response (i.e., probability of endorsing any particular item response option). Item discrimination refers to the extent to which the item is able to distinguish between individuals differing in levels of the latent trait factor. "Uniform DIF" exists when only the item difficulty parameter differs across groups (i.e., there is no interaction between latent trait factor level and group membership). "Nonuniform DIF" exists when the item discrimination parameter differs across groups (i.e., there is an interaction between latent trait factor level and group membership; cf. Chan, 2000). The IRT framework fits a nonlinear model, whereas the MACS framework fits a linear model (Meade & Lautenschlager, 2004). Within the MACS framework, the item intercept corresponds to the item difficulty parameter, and the item factor loading corresponds to the item discrimination parameter (Chan, 2000; Ferrando, 1996). The higher the intercept value, the stronger the endorsement of the item in question. The higher the factor loading value, the less ambiguous an item is perceived to be. Therefore, lack of invariance in only item intercepts across racial/ethnic groups (called "uniform DIF") is far less of a problem than is lack of invariance in item factor loadings across racial/ethnic groups (called "nonuniform DIF"), because the latter suggests systematic differences in the definition of theoretical

constructs (factors) among groups (see Cooke, Kosson, & Michie, 2001). The findings from the current study are interpreted within this integrative framework.

We expected that, on the basis of the available preliminary findings on the factor structure of the BSI-18, the originally proposed three-factor structure would provide a reasonable approximation to the data from each racial/ethnic group. Beyond this, the existing empirical literature on the BSI-18 offers little guidance with regard to the existence of measurement invariance and the pattern of equivalence in primary construct moments (i.e., latent factor means). Therefore, the present study should be viewed as exploratory.

Nevertheless, there are reasons to expect some degree of DIF as well as differences in latent factor mean levels in the current study. For instance, it has been noted by others that Hispanics, especially if they are less educated and less acculturated, have a tendency to endorse the extreme ends of scales (see e.g., Elliott, Haviland, Kanouse, Hambarsoomians, & Hays, 2009; Hui & Triandis, 1989; Marin, Gamba, & Marin, 1992; Weech-Maldonado, Elliott, Oluwole, Schiller, & Hays, 2008). In some studies they have also been found to not clearly separate mind from body symptoms of distress (see e.g., Guarnaccia, Angel, & Worobey, 1989), which might diminish the distinctiveness of the depression and somatization dimensions of the BSI-18 among Hispanics (Prelow et al., 2005). Finally, racial/ethnic groups may differ in levels of psychological distress due to differential exposure to adversity, including the experience of acculturative stress and discrimination, limited access to physical and mental health care services, low educational attainment, and socioeconomic disadvantage (see e.g., Bratter & Eschbach, 2005; Hovey & Magaña, 2000; McLoyd, 1990; McVeigh et al., 2006; Schulz et al., 2000). Hence, it appears likely that one might observe some degree of uniform (or even nonuniform) DIF, as well as differences in latent factor mean levels, in our sample, especially for the Hispanic group. However, because of the lack of more precise prior work, we were unable to formulate specific hypotheses on which items (or racial/ethnic groups) would exhibit DIF.

Method

Participants

Healthy Passages, funded by the Centers for Disease Control and Prevention (CDC), is a longitudinal study of a cohort of 5,147 fifth graders and their parents that explores health behaviors, outcomes, and related risk and protective factors using a multilevel approach. Healthy Passages provides a comprehensive assessment of adolescent health and behavior using data from multiple sources (child, parent, school). Psychometrically sound standardized measures were used when feasible. Qualitative (i.e., focus groups, cognitive interviews) and quantitative studies were conducted during study development, pretesting, and pilot testing in order to evaluate the appropriateness of survey language, translation, field procedures, and language-specific study materials. The background, project history, and conceptual framework of this study are described elsewhere (Windle et al., 2004).

Participants were recruited from public schools in each of these three geographic areas: (a) 25 contiguous public school districts in Los Angeles County, California; (b) 10 contiguous public school districts in and around Birmingham, Alabama; and (c) the largest public

school district in Houston, Texas. Eligible schools had an enrollment of at least 25 fifth graders, representing over 99% of students enrolled in regular classrooms in the three areas. A cluster probability sampling procedure was used to recruit students from each site. Public schools within the three study site communities were randomly selected with probabilities proportionate to a weighted measure of the scarcity of a school's students relative to race/ethnicity targets. Within sampled schools, all fifth-grade students (English- and Spanish-speaking) enrolled in regular academic classrooms of sampled schools were invited to participate. Of the 11,532 students enrolled in the 118 randomly selected schools, 6,663 of their primary caregivers (PCGs) who either agreed to be contacted about the study or who were unsure were invited to participate; 77.2% ($N = 5,147$) of them completed an interview.

In the current study, baseline data collected during the period 2004–2006 from female PCGs were used ($N = 4,711$). Of those, 33.9% ($n = 1,595$) were Hispanic, 25.6% ($n = 1,205$) White, 35.0% ($n = 1,650$) Black, and 5.5% ($n = 261$) other. The last group contained all participants who did not fit into any of the three main racial or ethnic target groups. Because there were so few participants in the “other” category, they were not analyzed as a distinct racial/ethnic group in this study. About 46.7% of the households contained both biological parents of the participating student. The average age of the women was 38.07 years ($SD = 7.16$), 56.9% were currently married, 8.3% were living with a partner, 65.6% were working part- or full-time, 30.9% had not graduated from high school, 20.7% had a GED or high school degree but had not attended college, and 48.5% had some years of college education.

Procedures

All three Healthy Passages research sites used standardized data collection materials and protocols, including training manuals, field manuals, and validation procedures. Institutional review boards at each study site and the CDC approved the study. Materials about the study and the Permission to Contact Form were distributed to eligible students in their classrooms. Students were asked to take home and share these materials with their PCGs. PCGs agreeing to learn more about the study were contacted by project staff to schedule a home visit. Alternative locations were available for PCGs who preferred to meet with field interviewers at a location other than their homes. After obtaining informed consent from the PCG and assent from the child, anthropometric measurements were taken for both the child and the PCG. Next, one interviewer conducted the child interview, and the other interviewer conducted the PCG interview. The PCG and child interviews consisted of a computer-assisted personal interview (CAPI) component followed by an audio computer-assisted self-interview (A-CASI) segment. The PCG and the child completed their interviews separately in private spaces. English and Spanish versions of the PCG and child CAPI/A-CASI interviews were available. On average, it took about 3 hr for the field interviewers to complete everything, including consent procedures, anthropometrics, CAPI, and A-CASI with the PCG and the child. Additional assessments included a school staff survey, school records data, teacher surveys, census tract data, and neighborhood observations. PCGs were reimbursed \$50, and children were given a \$20 gift card from a national chain store as reimbursement for their time completing the interview. Participating schools also received monetary reimbursement.

Measures

The current study used only a subset of the measures from the first wave of Healthy Passages. The BSI-18 was administered during the A-CASI with the PCG. Information on sociodemographic characteristics was mostly gathered during the CAPI with the PCG.

The BSI-18—The BSI-18 (Derogatis, 2000) is a self-reported screening inventory designed to assess participants' level of psychological distress on three dimensions: somatization, depression, and anxiety. The 18 items are divided equally across the three dimensions and were presented with the standard instructions asking participants to rate how much they have been "distressed or bothered" in the past 7 days, including today, by the given symptom, using a 5-point Likert scale ranging from 0 (*not at all*) to 4 (*extremely*). Each item contributes to only one subscale, which is scored by summing the scores on each of the six subscale items. The three raw subscale scores range from 0 to 24. The global severity index (GSI) of distress represents the sum across the three subscales. The raw GSI score ranges from 0 to 72, with higher scores indicating higher levels of psychological distress (Derogatis, 2000). The internal consistency estimates reported by Derogatis (2000) for the normative community sample of 1,134 adults are quite acceptable (.74 for somatization, .79 for anxiety, .84 for depression, and .89 for global severity index scores). Information on the test-retest reliability of the BSI-18 scores is not available from the scale author. Concurrent validity with the Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1994) is high, with correlations ranging from .91 to .96 on both subscale and GSI scores for the normative community sample.

Statistical Analysis

Our MACS model testing procedure was based on the strategies outlined by Byrne and Stewart (2006), Chan (2000), and Little (1997). First, four alternative CFA models with mean structures were estimated for each racial/ethnic group separately in order to identify the best fitting factor structure for each group. This included (a) a null model (i.e., only variances and intercepts of the items were freely estimated), (b) a single-factor model, (c) the hypothesized three-factor model (in which Items 2, 5, 8, 11, 14, and 17 loaded on the depression factor; Items 3, 6, 9, 12, 15, and 18 on the anxiety factor; and Items 1, 4, 7, 10, 13, and 16 on the somatization factor, with the three factors being specified as intercorrelated), and (d) an alternative four-factor model (i.e., the depression and somatization factors were specified as before, but Items 3, 6, and 15 loaded on an anxiety factor and Items 9, 12, and 18 on a panic-related factor). The best fitting factor structure was thereafter used in all simultaneous multigroup models. For each group-specific model, the first factor loading of each congeneric set of items (e.g., Items 1, 2, and 3 for the hypothesized three-factor model) was fixed to the value 1 for latent variable scaling purposes, and latent factor means were constrained to zero for model identification purposes.

Second, a simultaneous multigroup model was estimated that had the same number of factors and the same factor loading patterns across racial/ethnic groups (configural invariance model; Meredith, 1993). This model without any group-equality constraints served as our baseline model in subsequent hierarchical model comparisons. Third, group-

equality constraints were imposed on all factor loadings (weak factorial invariance model; Meredith, 1993). If this assumption did not appear tenable on the basis of an inspection of model fit relative to the baseline model, then the size of the modification index (MI) associated with each factor loading was used to flag items exhibiting nonuniform DIF (see details later). Only those group-equality constraints that were tenable were retained in subsequent nested models. Fourth, additional group-equality constraints were imposed on all indicator intercepts (strong factorial invariance model; Meredith, 1993). If this assumption did not appear tenable on the basis of an inspection of model fit relative to the baseline model, then the size of the MI associated with each indicator intercept was used to flag items exhibiting uniform DIF (see details later). Only those group-equality constraints that were tenable were retained in subsequent nested models. Fifth, an omnibus test was conducted to test the equality of the latent factor means. Throughout the entire simultaneous multigroup model-testing series, the first factor loading of each congeneric set of items (i.e., Items 1, 2, and 3 for the three-factor model) was fixed to the value 1 for latent variable scaling purposes. White PCGs were used as the reference group in all simultaneous multigroup models. This implies that the latent factor means for this group were fixed to zero for scaling purposes, with the latent factor means of all other racial/ethnic groups being compared with this reference group (except in the baseline model and the models that tested for weak factorial invariance; for those, all latent factor means had to be fixed to zero for identification purposes).

Analyses were conducted using the statistical software program Mplus 5.2 (Muthén & Muthén, 1998–2008). All analyses were performed with design weights (to account for differential probabilities of selection of students according to their school) and a cluster variable (to account for clustering of students within schools). All models were tested using the robust maximum likelihood (MLR) estimator, which corrects for both nonnormality (several of the BSI-18 items showed evidence of moderate non-normality on the basis of their skewness and kurtosis) and dependence due to the clustering of students within schools. Specifically, MLR uses the pseudomaximum likelihood asymptotic covariance matrix and a scaled test statistic (MLR χ^2) that is asymptotically equivalent to the Yuan–Bentler T2* test statistic (Asparouhov, 2005).

Evaluation of the tested models was based on multiple criteria that considered statistical, practical, and substantive fit. Values for the (corrected) chi-square statistic were reported for comparison purposes but not used for hypothesis testing because this statistic is known to be an overly sensitive index of model fit under conditions with large numbers of constraints, especially with large samples (see e.g., Marsh, Balla, & McDonald, 1988). With large samples, it tends to reject models with small discrepancies between the sample and fitted means and covariance matrices that are of little theoretical or practical relevance (see e.g., Bentler & Bonnett, 1980). Following the recommendations of Hu and Bentler (1999), we used the comparative fit index (CFI; Bentler, 1990), the root-mean-square error of approximation (RMSEA; Steiger, 1990), and the standardized root-mean-square residual (SRMR; Bentler, 1995). The cutoff values recommended by Hu and Bentler (1999) for the selected model fit indices were derived from Monte Carlo simulation work with the maximum likelihood estimator; relatively little is known about the extent to which these rules can be generalized to other estimation methods, such as the MLR estimator that was

used in the current study (see also the concerns raised by Marsh, Hau, & Wen, 2004). This should be kept in mind when evaluating the fit of the estimated models.

The CFI ranges in value from zero to one; values greater than .90 and .95 typically reflect acceptable and good model fit, respectively, of a target model relative to the null model (Bentler & Bonnett, 1980; Hu & Bentler, 1999). The RMSEA is a measure of a model's approximate fit in the population. Values less than .05 indicate good fit, and values as high as .08 represent acceptable errors of approximation in the population (Browne & Cudeck, 1993; Steiger, 1990). Finally, the SRMR is the average standardized residual value derived from fitting the hypothesized variance–covariance matrix to that of the observed data. It ranges from zero to one, with a value less than .08 indicating good model fit (Hu & Bentler, 1999).

Two strategies were applied during the hierarchical model comparisons. First, nested models were compared using the scaled chi-square difference test (i.e., the corrected MLR χ^2 as described by Muthén & Muthén, 2008; see also Satorra, 2000; Satorra & Bentler, 2001). The df for these model comparisons were computed as the degrees of freedom for the restricted model minus the degrees of freedom for the less restrictive comparison model. If the scaled chi-square difference value was statistically significant, it suggested that the group-equality constraints imposed in the more restrictive model were not tenable. In the latter case, we adopted the iterative strategy outlined by Chan (2000). Specifically, the size of the MI was used to flag DIF (or lack of invariance) for the given model parameter. If the largest MI was higher than the critical value (for $df = 1$), then the group-equality constraint in question was removed and the model was refitted to identify the largest MI associated with the remaining parameters upon which group-equality constraints were still imposed. This iterative procedure continued until the largest MI was no longer significant. To control for multiple testing, a stringent value of $p < .001$ was used during these post hoc examinations of the MIs. Any individual invariance constraints that passed this test were retained in subsequent hierarchical models. It has been noted in the literature that the assumption of measurement equivalence may be tenable for situations with (limited) partial invariance in factor loadings and indicator intercepts (see e.g., Byrne, Shavelson, & Muthén, 1989; Chan, 2000). However, several scholars have argued that the chi-square difference test is quite sensitive to sample size (see e.g., Cheung & Rensvold, 2002; Little, 1997), thereby rendering it an impractical criterion when testing invariance constraints with large samples. For this reason, we additionally inspected the practical fit of models with group-equality constraints (for a discussion, see Byrne & Stewart, 2006; Little, 1997). If models with invariance constraints exhibited both adequate fit to the data and a negligible difference in model fit relative to the baseline model, then the invariance assumption was deemed tenable. On the basis of the simulation study from Cheung and Rensvold (2002), a CFI that exceeded the value .01 was interpreted as evidence that the invariance constraints in question were not tenable.

Results

Descriptive Findings for the BSI-18

For each racial/ethnic group, Table 1 shows the means and standard deviations for each BSI-18 item, as well as each item's Pearson correlation with its corresponding subscale.

Mean raw item scores ranged from 0.04 (Item 17, suicidal thoughts) to 1.13 (Item 6, being tense) on a 5-point Likert scale where 0 corresponds to *not at all* and 1 corresponds to *a little bit*. These means differed significantly among the racial/ethnic groups ($p < .001$ for each item), with White women almost without exception showing the lowest endorsement levels and Hispanic women showing the highest endorsement levels. The same pattern of racial/ethnic group differences was observed for the raw GSI score. Correlations of each item with its respective subscale ranged from .51 to .96. The internal consistency of the three subscale scores was examined for each racial/ethnic group separately by using Cronbach's alpha (see Table 1). Coefficients ranged from .75 to .91 and are similar to those reported by Derogatis (2000) for a community sample (see the BSI-18 subsection under Method, Measures). The observed correlations between the raw GSI and each of the three sub-scales were quite high (not shown) and ranged from .80 to .98.

Preliminary Single-Group Analyses

As described previously, four alternative models were estimated separately by racial/ethnic group in preliminary analyses in order to identify the best fitting factor structure for each group. Information regarding fit for each model is shown in Table 2. For the three racial/ethnic groups, both the three-factor and the four-factor models exhibited acceptable to good fit to the data.¹ On the basis of the criteria of model fit, model parsimony, small normalized residuals, and adequacy and interpretability of parameter estimates (e.g., for Black women, a few estimated factor intercorrelations of $>.90$ indicated substantial redundancy among several of the specified factors in the four-factor model but not in the three-factor model), the hypothesized three-factor model was chosen as the final model for the simultaneous multigroup MACS models. Note that for Hispanic women the results for both the three- and four-factor models indicated linear dependency among some of the factors (i.e., factor correlations equal to or above 1.00), resulting in inadmissible model solutions. Similar problems (i.e., high or out-of-bounds standardized higher order factor loadings) were encountered when second-order confirmatory factor models (with a second-order general distress factor) were estimated post hoc for this group. Therefore, the simultaneous multigroup MACS analyses were conducted only for Black and White women.

Multigroup MACS Models

Next, the multigroup MACS models described earlier were tested. Information regarding the fit of each multigroup model and findings for model comparisons are shown in Table 3. The baseline configural invariance model with three factors also indicated acceptable to good fit to the data when estimated for Black and White women simultaneously (see Model 1 in Table 3). Further, all normalized residuals were below 2.58 for White women. For Black

¹The model fit indices showed a small discrepancy in their assessment of fit to the data. In contrast to RMSEA and SRMR, values for the CFI were not as close to cutoff values for good model fit as advocated by some scientists (see e.g., Hu & Bentler, 1999), especially for the Black women. Incremental fit indices such as the CFI compare the fit of the target model with the fit of the null model, and if the correlations between the manifest variables are fairly low, then there are relatively small amounts of covariance to explain and consequently there is less room for incremental fit indices to support a target model, even if it fits the data (Marsh et al., 1988). Sample correlations between the BSI-18 items were typically below .50 (range: .20 to .68 for White women; .18 to .68 for Black women). Thus, small item intercorrelations might have contributed to the small discrepancy between the model fit indices. Given that the more stringent CFI cutoff value proposed by Hu and Bentler (1999) may not be applicable to MLR estimation, we followed past practice (see e.g., Chan, 2000) for this variant of MACS models and used the lower cutoff (i.e., CFI value close to .90) as a criterion for acceptable model fit.

women, just four out of 189 normalized residuals were above 2.58 ($p < .01$). The standardized factor loadings were significant for each item ($p < .001$ for each) and ranged from .44 to .81 across the two groups. The communalities, the proportion of variance explained in each indicator by its respective latent factor, were generally of moderate size. For the somatization indicators, the average item communality was .42, ranging from .26 to .59. For the depression indicators, the average item communality was .52, ranging from .19 to .66. For the anxiety indicators, the average item communality was .47, ranging from .32 to .60. In an exploratory post hoc analysis, potential differences in the fit of the baseline model among the three study sites were examined for each racial/ethnic group separately. For White women, the simultaneous configural invariance model revealed few instances of local misfit for the Birmingham ($n = 562$) and Los Angeles ($n = 285$) study sites but slightly more instances of local misfit for the Houston ($n = 358$) study site (i.e., four out of 189 normalized residuals were above 2.58 at $p < .01$ for Houston, one for Los Angeles, and none for Birmingham). For Black women, the simultaneous configural invariance model showed few instances of local misfit for the Houston ($n = 412$) and Los Angeles ($n = 371$) study sites but slightly more instances of local misfit for the Birmingham ($n = 842$) study site (i.e., 10 out of 189 normalized residuals were above 2.58 at $p < .01$ for Birmingham, one for Los Angeles, and two for Houston). Although statistical power for these follow-up analyses was somewhat limited, findings showed little evidence of major differences in local (mis)fit between the three study sites for the two racial/ethnic groups.

Next, the assumption of weak factorial invariance across the two racial/ethnic groups was tested. The scaled chi-square difference test for Model 2a indicated that the assumption of full factor loading invariance was not tenable for the data (see Table 3). Using the iterative strategy from Chan (2000), we thereafter relaxed the invariance assumption for the factor loadings of Items 6 and 18 from the anxiety subscale, and the scaled chi-square difference test revealed that the resulting Model 2b did not show significantly worse fit compared with the baseline model (see Table 3). However, from a practical point of view, the differences in the freely estimated factor loadings between the two racial/ethnic groups were minor: The standardized factor loading was 0.56 (White) versus 0.59 (Black) for Item 6 and 0.70 (White) versus 0.73 (Black) for Item 18.² Furthermore, the difference in practical fit (CFI) between Model 2a and the baseline model was minimal (see Table 3). For these reasons, the assumption of full factor loading invariance was retained in all subsequent models.

Next, the assumption of full strong factorial invariance across the two racial/ethnic groups was tested. Both the scaled chi-square difference test and the difference in practical fit (CFI) between Model 3a and the baseline model indicated that this assumption was not tenable (see Table 3). Using the iterative strategy from Chan (2000), we thereafter relaxed the invariance assumption for the intercepts of Items 3, 6, and 18 from the anxiety subscale.

²The point estimate, standard error, and 95% confidence interval (CI) of the factor loading difference under the partial weak factorial invariance assumption (Model 2b) were calculated for the two items as proposed by Meade and Bauer (2007), using the unstandardized estimates: $\lambda = 0.41$, $SE = 0.49$, 95% CI = -0.55 to 1.36 for Item 6 and $\lambda = 0.22$, $SE = 0.35$, 95% CI = -0.47 to 0.90 for Item 18. The item communalities for White and Black women, respectively, were .31 and .35 for Item 6 and .49 and .53 for Item 18. The simulation study from Meade and Bauer revealed that, under conditions of equal group sizes, statistical power to detect nonuniform differential item functioning is high for large samples ($n = 400$ per group), high-factor overdetermination (about six items per factor), and relatively high item communalities. Effects of unequal group sizes were not evaluated in their simulation study.

Although the scaled chi-square difference test for the resulting Model 3b still indicated significantly worse fit of this model specification compared with the baseline model (see Table 3), no modification indices larger than 10.83 were observed for the remaining invariant item intercepts, and the CFI between the two models was negligible. Therefore, Model 3b was retained in subsequent analyses. The freely estimated intercepts were 0.45 ($SE = 0.03$; White) and 0.25 ($SE = 0.04$; Black) for Item 3; 1.14 ($SE = 0.03$; White) and 0.61 ($SE = 0.04$; Black) for Item 6; and 0.25 ($SE = 0.02$; White) and 0.12 ($SE = 0.03$; Black) for Item 18.

Finally, after having established partial strong factorial invariance, we assessed the equality of the latent factor means via an omnibus test (see Model 4 in Table 3), but it was found to be untenable for the two racial/ethnic groups on the basis of both the scaled chi-square difference test and the difference in practical fit (CFI). The freely estimated latent factor means from the final model indicated that Black women had significantly higher means on all three factors relative to White women.

Findings of a post hoc sensitivity analysis are shown in Table 4 and indicate that the magnitude of the latent factor mean difference for the anxiety factor was most affected by the chosen specification of strong factorial invariance. Specifically, the magnitude of the difference for the anxiety factor was smaller under the assumption of full strong factorial invariance relative to assuming partial strong factorial invariance. Standardized effect sizes indicated that the difference was relatively trivial: The mean for the anxiety factor was 0.24 factor standard deviations higher for Black women under full strong factorial invariance, whereas it was 0.39 factor standard deviations higher for Black women under partial strong factorial invariance. Cohen (1988) characterizes standardized effect sizes of approximately 0.50 as “medium” effects and those below 0.20 as “small” effects. Beyond this, the pattern of latent factor mean differences appeared to be fairly robust regardless of the chosen model specification of strong factorial invariance.

Discussion

This study applied MACS modeling to data from a large U.S. community sample to examine the invariance of the factor structure of the BSI-18 across Black, Hispanic, and White mothers of fifth graders. Findings indicated that the hypothesized three-factor structure of the BSI-18 was a reasonable representation of the data for Black and White women, whereas all tested multifactor solutions for Hispanic women exhibited substantial redundancy among several of the factors and inadmissible parameter estimates. This means that the underlying factor structure posited by the scale authors of the BSI-18 was supported for Black and White women but not for Hispanic women. Invariance testing was therefore performed for only Black and White women and revealed uniform DIF for three of the 18 items. Black women had significantly higher means on the three latent factors compared with White women. The internal consistency estimates were satisfactory for all racial/ethnic groups.

To the best of our knowledge, this is the first study that applied a rigorous invariance testing strategy in the form of a MACS analysis to the BSI-18 in order to examine the comparability

of its hypothesized factor structure across different racial/ethnic groups. The findings of this study have two key implications.

First, the factor structure of the BSI-18 does not appear to be equally robust across all racial/ethnic groups and is characterized by a substantial lack of distinctiveness between the three hypothesized factors (somatization, depression, anxiety) for Hispanic women. In this general sense, our findings echo prior research (Asner-Self et al., 2006; Prelow et al., 2005) on the BSI-18 with Hispanic samples. Note that about half of the Hispanic women came from the Houston site, and the other half from the Los Angeles site. Therefore, it is unlikely that this finding is an artifact of unique circumstances in one study site. It is possible that the factor structure of the BSI-18 is not robust for Hispanic adults because culturally relevant symptoms of psychological distress are only partially captured by the instrument (Prelow et al., 2005). Another possibility is that there are important differences in the factor structure of the BSI-18 between Latinas differing in country of origin and that they are obscured if Latinas are treated as a homogenous subgroup. However, subgroup analyses were not feasible for the current data set because the participating Hispanic women were predominantly Mexican American. This merits attention in further research. A final possibility is that the traditional CFA or MACS approach in which each item loads on one factor with no cross-loadings is not optimal for examinations of the BSI-18.

New methodological work on exploratory structural equation modeling (ESEM) techniques (Asparouhov & Muthén, 2009) aims to provide a means to evaluate the suitability of a cross-loading model specification for the given data and to incorporate the exploratory factor analysis (EFA) approach in which items load on all factors into the latent variable modeling framework. In one of the first ESEM applications to date, Marsh et al. (2009) demonstrated that the magnitude of the factor intercorrelations dropped markedly for the measure in question after the modeling approach was changed from the traditional CFA specification to the EFA specification. These results suggest that allowing for (potentially minor) cross-loadings for all items of the BSI-18 could be helpful in dealing with numerical estimation problems caused by linear dependency among latent factors, such as those observed in the current study for Hispanic women. In the future, researchers should investigate whether the ESEM approach is better suited for examining the factor structure of the BSI-18 than is the traditional CFA or MACS approach. However, without further work of this nature, we tentatively concur with others who have emphasized the utility of the BSI-18 as a measure of general psychological distress but cautioned against its application as a diagnostic tool for anxiety, depressive, and factious disorders (see Asner-Self et al., 2006; Prelow et al., 2005). The findings of this study suggest that usage of subscale scores may be especially problematic for Hispanic women.

The second key implication of this study relates to the existence of differential item functioning. The absence of more than trivial nonuniform DIF in this study was reassuring because it suggests that there are no systematic differences in the operational definition of the hypothesized factors of the BSI-18 between Black and White women. In other words, the constructs of this instrument are quite comparable across the two groups, and it is not necessary to remove items from the instrument in order to conduct group comparisons. The few indications of uniform DIF that were found in the current study applied to only one of

the three subscales. Specifically, levels of endorsement were lower for three items from the anxiety subscale among Black compared with White women. Note that only one of these items (Item 18) captured panic-related symptoms. Therefore, it is unlikely that the uniform DIF observed in this study was related to a fourth, panic factor that has been occasionally found in prior research (see e.g., Derogatis, 2000). However, uniform DIF poses few problems in the use of the BSI-18 across these two groups as a measure of levels of psychological distress as long as the interest is in conducting mean comparisons (see later discussion of results for sensitivity analysis). As one anonymous reviewer rightfully pointed out, uniform DIF is considerably more problematic when the BSI-18 is used as a screening tool and the same cutoff value on the GSI score and the anxiety subscale score is used for Black and White women. To the best of our knowledge, cutoff scores have not yet been fully established for the BSI-18 (see Derogatis, 2000). Nevertheless, a few qualifications are in order regarding our conclusion.

A problem of invariance testing within the MACS approach is the underlying premise that the items chosen as reference indicators do not themselves exhibit nonuniform DIF. If this assumption is not true, then DIF estimates may be distorted during testing (Chan, 2000; Yoon & Millsap, 2007). Although the proportion of Black women was somewhat larger than it was for White women, which somewhat decreases the statistical power of the invariance tests for a fixed total sample size (Kaplan & George, 1995), the total sample sizes are large enough that the issue of power is unlikely to be of concern. Kaplan and George (1995) also noted that if the null hypothesis of equal latent factor means is rejected under conditions of unequal sample size, then one can be fairly confident that the hypothesis is false, given that power has no bearing on Type I error rates. Our measurement invariance testing strategy did not include tests for strict factorial invariance (Meredith, 1993). Many scholars regard strict factorial invariance as an overly restrictive assumption and an optional part of sound testing for factorial invariance (see e.g., Brown, 2006; Byrne & Stewart, 2006; Little, 1997; Raju et al., 2002). Finally, our strategy of identifying differential item functioning via large modification indices was essentially exploratory in nature. Given the lack of solid prior research on this issue for the BSI-18, this was a defensible and appropriate procedure for the current study, but we concur with Chan (2000) that a more confirmatory approach to DIF testing is desirable if a priori hypotheses on the expected pattern of differential item functioning can be formulated for the instrument in question. For all these reasons, cross-validation of our findings is warranted.

Taken together, the results from this study therefore imply that mean comparisons on the BSI-18 are meaningful, at least for women from these two racial/ethnic groups, and that the finding of a discrepancy in levels of psychological distress—as indexed by the latent factor means for somatization, depression, and anxiety—reflects true differences between Black and White women. Although the effect size was fairly modest in each case, the group differences in the three latent variable means appeared to be quite robust, according to the sensitivity analysis. Further, the direction of the latent variable mean differences between Black and White women was consistent with findings from large national survey data (see e.g., Bratter & Eschbach, 2005) but has not been consistently indicated in the literature (see e.g., Nuru-Jeter, Williams, & LaVeist, 2008). Sometimes, the direction is even reversed after factors such as social class, acculturation, marital status, income, or poverty concentration

are taken into account (see e.g., Schulz et al., 2000). Although the current study was not able to illuminate the causal processes behind the observed group difference in levels of psychological distress, previous research has suggested that they are likely complex in nature and the result of joint effects of multiple factors (see e.g., Nuru-Jeter et al., 2008; Ulbrich, Warheit, & Zimmerman, 1989). This issue merits further research.

Although findings of the current study offer some intriguing insights, they should be viewed in the context of study limitations. First, the BSI-18 is a self-report measure, and, as with all self-report instruments, responses may have been influenced by problems such as informant bias and shared measurement variance. Information from clinical rating scales or other sources of independent diagnostic data was not available for this sample but would be useful in future research for the external validation of the subscales and/or examinations of the criterion-related validity of the BSI-18. Second, the limited group size of some racial/ethnic groups within study sites and numerical estimation difficulties precluded systematic testing of weak and strong factorial invariance of the BSI-18 among the three study sites. Even though our exploratory post hoc inspection did not reveal major differences in local fit for the three-factor model among the study sites, more systematic tests would have been preferable. Third, this was a study of female caregivers. As such, the findings may not generalize to men or to women of different ages or circumstances.

Nevertheless, this study also had several strengths, including the relatively large sample size and data from three racial/ethnic groups, which facilitated our efforts to address an important unresolved question in the literature in a fairly rigorous manner. Further research on this issue should include similar analyses for men and other racial/ethnic groups, such as Asian Americans. Longitudinal invariance testing of the BSI-18 across different time intervals is also warranted.

Acknowledgments

The Healthy Passages study is funded by the Centers for Disease Control and Prevention (Cooperative Agreements U48DP000046, U48DP000057, and U48DP000056). The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

- Asner-Self KK, Schreiber JB, Marotta SA. A cross-cultural analysis of the Brief Symptom Inventory-18. *Cultural Diversity and Ethnic Minority Psychology*. 2006; 12:367–375. [PubMed: 16719583]
- Asparouhov T. Sampling weights in latent variable modeling. *Structural Equation Modeling*. 2005; 12:411–434.
- Asparouhov T, Muthén B. Exploratory structural equation modeling. *Structural Equation Modeling*. 2009; 16:397–438.
- Bentler PM. Comparative fit indexes in structural models. *Psychological Bulletin*. 1990; 107:238–246. [PubMed: 2320703]
- Bentler, PM. EQS structural equations program manual. Encino, CA: Multivariate Software; 1995.
- Bentler PM, Bonett DG. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*. 1980; 88:588–606.
- Bratter JL, Eschbach K. Race/ethnic differences in nonspecific psychological distress: Evidence from the National Health Interview Survey. *Social Science Quarterly*. 2005; 86:620–644.

- Brown, TA. Confirmatory factor analysis for applied research. New York, NY: Guilford Press; 2006.
- Browne, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. Testing structural equation models. Newbury Park, CA: Sage; 1993. p. 136-162.
- Byrne BM, Shavelson RJ, Muthén B. Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*. 1989; 105:456–466.
- Byrne BM, Stewart SM. The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*. 2006; 13:287–321.
- Chan D. Detection of differential item functioning on the Kirton Adaptation-Innovation Inventory using multiple-group mean and covariance structure analysis. *Multivariate Behavioral Research*. 2000; 35:169–199.
- Cheung GW, Rensvold RB. Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*. 2002; 9:233–255.
- Cohen, J. Statistical power analysis for the behavioral sciences. 2. Hillsdale, NJ: Erlbaum; 1988.
- Cooke DJ, Kosson DS, Michie C. Psychopathy and ethnicity: Structural, item, and test generalizability of the Psychopathy Checklist-Revised (PCL-R) in Caucasian and African American participants. *Psychological Assessment*. 2001; 13:531–542. [PubMed: 11793896]
- Derogatis, LR. Brief Symptom Inventory (BSI): Administration, scoring, and procedures manual. 3. Minneapolis, MN: NCS Pearson; 1993.
- Derogatis, LR. Symptom Checklist-90-R (SCL-90-R): Administration, scoring, and procedures manual. 3. Minneapolis, MN: NCS Pearson; 1994.
- Derogatis, LR. Brief Symptom Inventory (BSI)-18: Administration, scoring, and procedures manual. Minneapolis, MN: NCS Pearson; 2000.
- Derogatis, LR.; Fitzpatrick, M. The SCL-90-R, the Brief Symptom Inventory (BSI), and the BIS-18. In: Maruish, ME., editor. The Use of Psychological Testing for Treatment Planning and Outcomes Assessment: Vol. 3. Instruments for adults. Mahwah, NJ: Erlbaum; 2004. p. 1-41.
- Elliott MN, Haviland A, Kanouse DE, Hambarsoomians K, Hays RD. Adjusting for subgroup differences in extreme response tendency when rating health care: Impact on disparity estimates. *Health Services Research*. 2009; 44:542–561. [PubMed: 19040424]
- Ferrando PJ. Calibration of invariant item parameters in a continuous item response model using the extended LISREL measurement submodel. *Multivariate Behavioral Research*. 1996; 31:419–439.
- Guarnaccia PJ, Angel R, Worobey JL. The factor structure of the CES-D in the Hispanic Health and Nutrition Survey: The influence of ethnicity, gender, and language. *Social Science and Medicine*. 1989; 29:85–94. [PubMed: 2740931]
- Hancock GR. Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*. 2001; 66:373–388.
- Hovey JD, Magaña C. Acculturative stress, anxiety, and depression among Mexican immigrant farmworkers in the midwest United States. *Journal of Immigrant Health*. 2000; 2:119–131. [PubMed: 16228745]
- Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*. 1999; 6:1–55.
- Hui CH, Triandis HC. Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*. 1989; 20:296–309.
- Kaplan D, George R. A study of power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling*. 1995; 2:101–118.
- Little TD. Means and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*. 1997; 32:53–76.
- Marin G, Gamba RJ, Marin BV. Extreme response style and acquiescence among Hispanics: The role of acculturation and education. *Journal of Cross-Cultural Psychology*. 1992; 23:498–509.
- Marsh HW, Balla JR, McDonald RP. Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*. 1988; 103:391–410.

- Marsh HW, Hau KT, Wen Z. In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's findings. *Structural Equation Modeling*. 2004; 11:320–341.
- Marsh HW, Muthén B, Asparouhov T, Lüdtke O, Robitzsch A, Morin AJS, Trautwein U. Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*. 2009; 16:439–476.
- McLoyd VC. The impact of economic hardship on Black families and children: Psychological distress, parenting, and socioemotional development. *Child Development*. 1990; 61:311–346. [PubMed: 2188806]
- McVeigh KH, Galea S, Thorpe LE, Mautsby C, Henning K, Sederer LI. The epidemiology of nonspecific psychological distress in New York City, 2002 and 2003. *Journal of Urban Health*. 2006; 83:394–405. [PubMed: 16739043]
- Meade AW, Bauer DJ. Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling*. 2007; 14:611–635.
- Meade AW, Lautenschlager GJ. A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*. 2004; 7:361–388.
- Meredith W. Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*. 1993; 58:525–543.
- Muthén, LK.; Muthén, BO. Mplus user's guide. Los Angeles, CA: Muthén & Muthén; 1998–2008.
- Muthén, LK.; Muthén, BO. Chi-square difference testing using the Satorra-Bentler scaled chi-square. 2008. Retrieved from <http://www.statmodel.com/chidiff.shtml>
- Nuru-Jeter A, Williams CT, LaVeist TA. A methodological note on modeling the effects of race: The case of psychological distress. *Stress and Health*. 2008; 24:335–420.
- Prelow HM, Weaver SR, Swenson RR, Bowman MA. A preliminary investigation of the validity and reliability of the Brief Symptom Inventory-18 in economically disadvantaged Latina American mothers. *Journal of Community Psychology*. 2005; 33:139–155.
- Raju NS, Laffitte LJ, Byrne BM. Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*. 2002; 87:517–529. [PubMed: 12090609]
- Recklitis CJ, Parsons SK, Shih MC, Mertens A, Robison LL, Zeltzer L. Factor structure of the Brief Symptom-Inventory-18 in adult survivors of childhood cancer: Results from the Childhood Cancer Survivor Study. *Psychological Assessment*. 2006; 18:22–32. [PubMed: 16594809]
- Satorra, A. Scaled and adjusted restricted tests in multi-sample analysis of moment structures. In: Heijmans, RDH.; Pollock, DSG.; Satorra, A., editors. *Innovations in multivariate statistical analysis: A Festschrift for Heinz Neudecker*. London, England: Kluwer Academic; 2000. p. 233–247.
- Satorra A, Bentler PM. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*. 2001; 66:507–514.
- Schulz A, Williams D, Israel B, Becker A, Parker E, James SA, Jackson J. Unfair treatment, neighborhood effects, and mental health in the Detroit metropolitan area. *Journal of Health and Social Behavior*. 2000; 41:314–332. [PubMed: 11011507]
- Sörbom D. A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology*. 1974; 27:229–239.
- Steiger JH. Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*. 1990; 25:173–180.
- Thompson, B. *Exploratory and confirmatory factor analysis*. Washington, DC: American Psychological Association; 2004.
- Thompson, MS.; Green, SB. Evaluating between-group differences in latent variable means. In: Hancock, GR.; Mueller, RO., editors. *Structural equation modeling: A second course*. Greenwich, CT: Information Age; 2006. p. 119–169.
- Ulrich PM, Warheit GJ, Zimmerman RS. Race, socioeconomic status, and psychological distress: An examination of differential vulnerability. *Journal of Health and Social Behavior*. 1989; 30:131–146. [PubMed: 2723379]

- U.S. Census Bureau. Demographic trends in the 20th century (Census 2000 Special Reports, Series CENSR-4). Washington, DC: U.S. Department of Commerce; 2002.
- Weech-Maldonado RW, Elliott MN, Oluwole T, Schiller C, Hays RD. Survey response style and differential use of CAHPS rating scales by Hispanics. *Medical Care*. 2008; 46:963–968. [PubMed: 18725851]
- Widaman, KF.; Reise, SP. Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In: Bryant, KJ.; Windle, M.; West, SG., editors. *The science of prevention*. Washington, DC: American Psychological Association; 1997. p. 281-324.
- Windle M, Grunbaum JA, Elliott M, Tortolero SR, Berry S, Gilliland J, Schuster M. Healthy Passages: A multilevel, multimethod longitudinal study of adolescent health. *American Journal of Preventive Medicine*. 2004; 27:164–172. [PubMed: 15261905]
- Yoon M, Millsap RE. Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*. 2007; 14:435–463.
- Zabora J, BrintzenhofeSzoc K, Jacobsen P, Curbow B, Piantadosi S, Hooker C, Derogatis L. A new psychosocial screening instrument for use with cancer patients. *Psychosomatics*. 2001; 42:241–246. [PubMed: 11351113]

Table 1

Summary Statistics for the BSI-18 Items (Raw Scores) by Subscale and Racial/Ethnic Group (N = 4,450)

'Subscale/GSI, item no., and item	White (n = 1,205)				Hispanic (n = 1,595)				Black (n = 1,650)			
	M	SD	α	r^2	M	SD	α	r^2	M	SD	α	r^2
Depression Subscale			.85				.88				.84	
2. No interest	0.39	0.77		.92	0.84	0.96		.83	0.65	1.02		.72
												42.02 ***
5. Lonely	0.45	0.84		.96	0.78	0.99		.85	0.70	1.12		.75
												17.92 ***
8. Blue	0.58	0.86		.90	1.02	1.04		.90	0.69	1.07		.84
												28.48 ***
11. Worthlessness	0.27	0.69		.91	0.71	1.01		.92	0.39	0.91		.86
												27.19 ***
14. Hopelessness	0.34	0.75		.94	1.00	1.06		.84	0.55	1.03		.88
												61.43 ***
17. Suicidal thoughts	0.04	0.29		.82	0.42	0.94		.85	0.12	0.58		.64
												50.13 ***
Anxiety Subscale			.79				.88				.84	
3. Nervousness	0.45	0.81		.91	0.79	0.99		.88	0.48	0.96		.73
												27.32 ***
6. Tense	1.13	1.02		.60	0.83	0.96		.61	0.84	1.12		.53
												35.64 ***
9. Scared	0.16	0.51		.84	0.64	0.88		.95	0.40	0.89		.87
												59.65 ***
12. Panic episodes	0.16	0.55		.88	0.85	1.09		.87	0.35	0.92		.79
												60.71 ***
15. Restlessness	0.26	0.65		.87	0.61	0.85		.88	0.44	0.92		.79
												36.61 ***
18. Fearful	0.25	0.61		.86	0.52	0.83		.88	0.30	0.73		.74
												26.10 ***
Somatization Subscale			.75				.91				.82	
1. Faintness	0.24	0.59		.66	0.62	0.97		.82	0.37	0.82		.51
												43.99 ***
4. Chest pains	0.15	0.53		.87	0.69	1.06		.88	0.44	0.94		.71
												75.19 ***
7. Nausea	0.39	0.76		.78	0.69	0.89		.94	0.57	1.00		.59
												23.00 ***
10. Short of breath	0.16	0.58		.93	0.73	1.06		.94	0.37	0.91		.74
												51.85 ***
13. Numb/tingling	0.29	0.75		.94	0.81	1.03		.91	0.60	1.09		.78
												41.10 ***
16. Body weakness	0.29	0.69		.85	0.78	0.99		.90	0.60	1.05		.80
												62.85 ***
GSI			.90				.96				.92	
Raw score	6.01	7.96			13.32	13.39			8.88	11.48		
												47.02 ***

Note. Summary statistics (except Cronbach's alpha) adjusted for sample weights and clusters, using statistical software package STATA 10. BSI-18 = Brief Symptom Inventory-18; GSI = global severity index.

^aItem-subscale correlation.

 $p < .001$.

Table 2

Model fit of Preliminary Models Estimated for Each Racial/Ethnic Group Separately

Model and group	MLR χ^2	df	CFI	RMSEA [90% CI]	SRMR
Null model					
White	4,289.85	153		.150 [0.146, 0.154]	.356
Black	5,435.60	153		.146 [0.142, 0.149]	.382
Hispanic	9,697.59	153		.202 [0.198, 0.205]	.502
One-factor model					
White	687.87	135	.866	.058 [0.054, 0.063]	.061
Black	1,051.73	135	.826	.065 [0.061, 0.068]	.060
Hispanic	1,100.67	135	.899	.068 [0.065, 0.072]	.045
Three-factor model					
White	376.91	132	.941	.039 [0.035, 0.044]	.047
Black	704.05	132	.892	.052 [0.048, 0.055]	.048
Hispanic	939.09	132	.915	.063 [0.059, 0.067]	.043
Four-factor model					
White	329.22	129	.952	.036 [0.031, 0.041]	.044
Black	647.11	129	.902	.050 [0.046, 0.054]	.046
Hispanic	931.56	129	.916	.064 [0.060, 0.068]	.043

Note. All models were adjusted for sample weights and clusters. MLR = robust maximum likelihood; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root-mean-square residual.

Table 3

Testing for Invariance of the Three-Factor Model Across White (n = 1,205) and Black (n = 1,650) Women

Model and type	MLR χ^2	df	CFI	RMSEA [90% CI]	SRMR	Model comparison	CFI	MLR χ^2	df	p
Model 1: Configural invariance (baseline model)										
Full	1,076.47	264	.915	.047 [0.044, 0.050]	.048					
Model 2: Weak factorial invariance (invariance of factor loadings)										
Full (Model 2a)	1,096.13	279	.914	.045 [0.043, 0.048]	.056	2a vs. 1	.001	34.95	15	<.01
Partial (Model 2b)	1,075.74	277	.916	.045 [0.042, 0.048]	.054	2b vs. 1	.001	20.34	13	>.05
Model 3: Strong factorial invariance (invariance of factor loadings and intercepts)										
Full (Model 3a)	1,267.37	294	.898	.048 [0.046, 0.051]	.060	3a vs. 1	.017	190.45	30	<.001
Partial (Model 3b)	1,150.43	291	.910	.046 [0.043, 0.048]	.057	3b vs. 1	.005	77.91	27	<.001
Model 4: Partial strong factorial invariance and full latent factor means invariance										
Full	1,209.13	294	.904	.047 [0.044, 0.050]	.071	4 vs. 3b	.011	84.84	3	<.001

Note. All models were adjusted for sample weights and clusters. MLR = robust maximum likelihood; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; CI = confidence interval; SRMR = standardized root-mean-square residual.

Table 4

Sensitivity Analysis for Latent Construct Mean Differences Under Full Versus Partial Strong Factorial Invariance

Factorial invariance and factor	Estimated latent factor mean for Black women	SE	99% CI	Standardized effect size
Full strong factorial invariance (Model 3a)				
Somatization	0.16***	0.02	[0.12, 0.21]	0.45
Depression	0.16***	0.04	[0.07, 0.25]	0.27
Anxiety	0.14***	0.03	[0.06, 0.21]	0.24
Partial strong factorial invariance (Model 3b)				
Somatization	0.16***	0.02	[0.12, 0.21]	0.45
Depression	0.16***	0.04	[0.07, 0.25]	0.27
Anxiety	0.23***	0.04	[0.14, 0.32]	0.39

Note. The reference group was White women. The method for computing the standardized effect size was similar to that used in Hancock (2001), using the formula shown in M. S. Thompson and Green (2006) with a pooled estimate of the standard deviation because the factor variances were not constrained to be equal across groups. CI = confidence interval.

 $p < .001$.