

Measurement Invariance of the Beck Depression Inventory–Second Edition (BDI-II) Across Gender, Race, and Ethnicity in College Students

Mark A. Whisman, Charles M. Judd, Natalie T. Whiteford and Heather L. Gelhorn
Assessment published online 7 October 2012
DOI: 10.1177/1073191112460273

The online version of this article can be found at:
<http://asm.sagepub.com/content/early/2012/10/04/1073191112460273>

Published by:



<http://www.sagepublications.com>

Additional services and information for *Assessment* can be found at:

Email Alerts: <http://asm.sagepub.com/cgi/alerts>

Subscriptions: <http://asm.sagepub.com/subscriptions>


Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

>> [OnlineFirst Version of Record](#) - Oct 7, 2012

[What is This?](#)

Measurement Invariance of the Beck Depression Inventory–Second Edition (BDI-II) Across Gender, Race, and Ethnicity in College Students

Assessment
XX(X) 1–10
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191112460273
http://asm.sagepub.com


Mark A. Whisman¹, Charles M. Judd¹,
Natalie T. Whiteford¹, and Heather L. Gelhorn¹

Abstract

Measurement invariance of the Beck Depression Inventory–Second Edition (BDI-II) across gender, race, and ethnic groups was evaluated in a large sample of college students, using pooled data from 11 universities from diverse geographical regions in the United States ($N = 7,369$). Confirmatory factor analysis was used to test the fit of several possible factor structures, and the results from these analyses indicated that the BDI-II was most adequately represented by a hierarchical four-factor structure, composed of three first-order factors and one second-order factor. Results based on analyses of covariance structures indicated there was factorial invariance for this hierarchical four-factor structure across groups, suggesting that the BDI-II provides an assessment of severity of depressive symptoms that is equivalent across gender, race, and ethnicity in college students.

Keywords

measurement invariance, equivalence, bias, factor, depression

One of the most widely accepted measures for assessing the severity of depression is the Beck Depression Inventory–Second Edition (BDI-II; Beck, Steer, & Brown, 1996), which was developed to replace the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) and its successor, the BDI-IA (Beck, Rush, Shaw, & Emery, 1979; Beck & Steer, 1987). In developing the BDI-II, there were several substantive changes made to the BDI-IA: four items were eliminated and replaced by new items, items were changed to allow for increases as well as decreases in appetite and sleep, response options were reworded, item labels were added, and the time frame was changed from “past week” to “past two weeks” to be more consistent with the *Diagnostic and Statistical Manual of Mental Disorders*, fourth edition (*DSM-IV*). Because of these changes, the BDI-II is considered a “substantial revision of the original BDI” (Beck et al., 1996, p. 1). The BDI-II is viewed as one of the best existing measures of depressive symptoms (Joiner, Walker, Pettit, Perez, & Cukrowicz, 2005).

The BDI-II is frequently used to examine between-group differences in levels of depression or variables associated with depression. For example, female college students score higher on the BDI-II than male college students (e.g., Beck et al., 1996; Carmody, 2005; Osman et al., 1997), and some

correlates of the BDI-II differ between male and female college students (e.g., You, Merritt, & Conner, 2009). Similarly, there is some evidence of mean differences on the BDI-II across groups defined by race or ethnicity (e.g., Hambrick et al., 2010). Implicit in these studies is the assumption that the BDI-II is measuring the same construct the same way across groups of interest (i.e., that the scale is measurement equivalent). Although there is a large literature on the reliability and validity of the BDI-II (for a review, see Dozois & Covin, 2004), and although Hambrick et al. (2010) recently found little evidence for differential item functioning on BDI-II items between White and African American undergraduate students using item response theory (IRT), there is comparatively little research on measurement invariance of the scale. Measurement invariance or equivalence (or construct comparability) is defined as “the mathematical equality of corresponding

¹University of Colorado Boulder, Boulder, CO, USA

Corresponding Author:

Mark A. Whisman, Department of Psychology and Neuroscience,
University of Colorado Boulder, 345 UCB, Boulder, CO 80309-0345,
USA
Email: mark.whisman@colorado.edu

measurement parameters for a given factorially defined construct (i.e., the loadings and intercepts of a construct's multiple manifest indicators) across two or more groups" (Little, 1997, p. 55). Measurement invariance is important because if the BDI-II is not operating equivalently across groups, any between-group differences in means or correlates may be inaccurate and misleading. That is to say, without knowing whether the BDI-II is measurement invariant, it cannot be known whether differences across groups in means or correlates of the BDI-II are true differences or differences due to psychometric differences in item responses.

Few studies have directly evaluated the measurement invariance of the BDI-II across gender, race, or ethnicity.¹ We found only one study that examined the factorial invariance of the BDI-II, and results from this study suggested that the BDI-II demonstrated factorial invariance across Hong Kong and American high school students (Byrne & Stewart, 2006; Byrne, Stewart, Kennard, & Lee, 2007). Furthermore, factorial invariance was found for the BDI-IA on comparisons across English and French cultural groups in Canadian adolescents (Byrne & Baron, 1994), across Canadian, Swedish, and Bulgarian adolescents (Byrne & Campbell, 1999), and across gender in adolescents from Canada (Byrne, Baron, & Campbell, 1993, 1994), Sweden (Byrne, Baron, Larsson, & Melin, 1996), and Bulgaria (Byrne, Baron, & Balev, 1996).

If the assumption is that the BDI-II provides an assessment of depression that is equivalent across gender, race, and ethnicity, then the fact that the measurement invariance of the BDI-II has not been established across groups represents a potentially serious problem, because mean differences on the BDI-II between groups or group differences in patterns of correlations between variables and the BDI-II could be artifactual and misleading. The purpose of the current study, therefore, was to test for multigroup measurement invariance of the BDI-II across gender, race, and ethnic groups. Specifically, we used confirmatory factor analysis (CFA) and evaluated the extent to which the factor loadings and intercepts are equivalent across groups. Item factor loadings are analogous to item discrimination parameters in IRT, whereas item intercepts correspond to item difficulty parameters in IRT (Chan, 2000). Based on existing research on measurement invariance with the BDI-IA, we hypothesized that the BDI-II would demonstrate measurement invariance across groups.

Method

Participants

Participants included undergraduate students from 11 universities across the United States. We used data from multiple universities to have adequate numbers of minority students and to enhance generalizability of the results. To identify studies that included BDI-II data from minority

Table 1. Studies Providing Data for the Current Analyses

Study	University	<i>n</i>
Harari, Waehler, and Rogers (2005)	University of Akron	292
Klibert, Langhinrichsen-Rohling, and Saito (2005)	University of South Alabama	457
Lewandowski et al. (2006)	University of North Carolina at Greensboro	1,258
Norton (2005)	University of Houston	500
Okazaki (2002)	University of Illinois at Urbana-Champaign	398
Scarpa et al. (2002)	Virginia Polytechnic Institute and State University	500
Storch, Roberti, and Roth (2004)	University of Florida and Louisiana State University	401
Whisman, Perez, and Ramel (2000)	Yale University	490
Wiebe and Penley (2005)	University of Texas at El Paso	792
Unpublished data	University of Colorado Boulder	2,281

students, we first conducted a PsycINFO literature search for articles published through 2007 using the terms (a) "BDI-II" or "Beck Depression Inventory–Second Edition," (b) "student" or "undergraduate," and (c) "minority" or "African American" or "Black" or "Asian" or "Hispanic" or "Latino." In addition, we reasoned that studies that included large samples of undergraduates would be more likely to yield larger numbers of minority students, and because factor analytic studies tend to be based on relatively larger samples, we included factor analysis studies known to us that included the BDI-II, supplemented with a separate literature search using the same search terms for the BDI-II and the keywords "factor analysis." Because of potential problems with translated versions of the BDI-II, we limited our search to the English version of the measure. Finally, we limited the search to college samples within the United States. The search identified 16 studies. We then contacted the authors of these studies and requested the raw data on the BDI-II, as well as data on gender, race/ethnicity, and age (if available). Of the 16 requests, 9 authors provided data from their study; an additional 2 authors provided raw data on the BDI-II but did not have data on race/ethnicity. Thus, the overall response rate was 69%, and we obtained usable data from 56% of the requested studies. This response rate was better than some other studies based on reanalysis of existing data sets (e.g., Witcherts, Borsboom, Kats, & Molenarr, 2006). We also collected additional data on the BDI-II at the University of Colorado Boulder. Therefore, the study is based on combined data from 9 published studies and 1 unpublished sample, which represent data from 11 universities from diverse geographical regions in the United States. The studies that provided data and the number of participants per study are listed in Table 1. Informed consent was obtained from participants

in all samples, and the institutional review board at the University of Colorado Boulder approved the use of de-identified data for the current analyses. The final sample included 7,369 college students—4,790 women (65.0%) and 2,579 men. The self-reported race–ethnic distribution of the sample was 4,912 White (66.7%), 682 Black (9.3%), 646 Asian (8.8%), 953 Latino (12.9%), and 176 other (2.4%). The mean age of the sample was 20.0 years ($SD = 3.8$), although data on age were available from only 5,565 participants.

Measure

The BDI-II is a self-report questionnaire designed to measure severity of depressive symptoms in adolescents and adults (Beck et al., 1996). It consists of 21 items, rated on a 4-point scale ranging from 0 to 3, which are rated with respect to the “past two weeks, including today.” Items are summed to create a total score, ranging from 0 to 63, with higher scores reflecting greater severity. Prior research indicates that the measure has excellent internal consistency, differentiates between depressed and nondepressed individuals, and correlates highly with other measures of depressive symptoms and depression-related constructs (for a review, see Dozois & Covin, 2004). The total score demonstrated excellent internal consistency in this study ($\alpha = .90$).

Analyses

Testing factorial invariance was conducted within the framework of CFA modeling using procedures similar to those outlined by Byrne (Byrne, 2006; Byrne & Stewart, 2006) and by Chen, Sousa, and West (2005). These authors discuss how examination of measurement invariance should proceed in the context of a second-order factor model, which as we discuss below is the structure that we find and others have suggested for the BDI-II. The exact steps in these procedures vary between the two sets of authors and the steps that we follow represent an integration of the two approaches. All analyses were conducted with the EQS 6.1 (Bentler, 2005) program, which permits estimation based on the Satorra–Bentler (Satorra & Bentler, 1988) scaled χ^2 , permitting correct goodness-of-fit indices and standard errors for data that are nonnormally distributed. This approach treats responses as measured on a continuous scale, but previous explorations of factorial invariance of the BDI-II (e.g., Byrne et al., 2007; Byrne & Stewart, 2006) have also made this assumption.

We first conducted a series of baseline CFA models of several possible factor structures of the BDI-II, informed by past research on obtained factor structures in college student samples. These results were used to identify a well-fitting model to use in the analyses of factorial invariance. As already suggested, preliminary models converged on the same second-order latent factor structure obtained by Byrne

and colleagues (Byrne et al., 2007; Byrne & Stewart, 2006). Once this baseline model was shown to be consistent with the data, we then proceeded to test the equivalence of this model across subgroups, using a series of ordered steps based on integrating approaches outlined for second-order factor models by Byrne (2006) and by Chen et al. (2005).

Our first model specified simply configural invariance, meaning that the same factor structure was estimated simultaneously in both groups but no between-group constraints were placed on the parameter estimates (Model 1). Assuming this model is consistent with the data, we proceeded by imposing a series of more stringent between-group constraints to examine factorial invariance. Consistent with both Byrne (2006) and Chen et al. (2005), Model 2 imposed between-group equality constraints on the loadings of the measured variables on the first-order factors. Assuming these constraints remain consistent with the data, Model 3 is then estimated in which both first-order and second-order loadings are constrained to be equal across groups. This model specifies what is usually meant by measurement invariance, allowing differences in factor variances and error variances, but forcing measurement equivalence (equal loadings) across groups. Further constraints on the factor and item variances can then be specified to examine even more restrictive questions about the equality of variances and covariances in the two groups. Our Model 4 imposed equality constraints on the disturbance variances for two of the three first-order factors and then on the variance of the second-order factor. Given equal between-group loadings, if this model remains consistent with the data, then it suggests not only measurement invariance but also equivalent between-group variance in the latent factors or traits measured by the items. And finally, although we had no expectation that this model would remain consistent with the data, we estimated Model 5, which constrained disturbance variances of all indicator variables to be equal between groups. At this level, the model assumes that the full variance/covariance matrices of all indicator variables are identical in the two groups.

We used several indices for evaluating the model fit. First, because the BDI-II data were nonnormally distributed, we used the Satorra–Bentler scaled χ^2 (S-B χ^2 ; Satorra & Bentler, 1988) instead of the uncorrected maximum likelihood chi-square (χ^2). The S-B χ^2 incorporates a scaling correction for the χ^2 when distributional assumptions are violated. Similar to the χ^2 statistic, use of the S-B χ^2 is sensitive to sample size. Consequently, other goodness-of-fit statistics, developed and recommended in reporting results for analyses of measurement invariance, were also included. These included the comparative fit index (CFI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA) and its 90% confidence interval (90% CI; Hu & Bentler, 1999). CFI values range from 0 to 1.00, with values $>.94$ generally accepted as a good fit. SRMR values range from 0 to 1.00, with

values $<.08$ indicating a well-fitting model. The RMSEA is expressed per degree of freedom, which makes it sensitive to model complexity; values $<.05$ indicate acceptable fit. For both CFI and RMSEA, we report the robust versions of these measures (*CFI and *RMSEA).

The various models we tested can be seen as nested under each other, in the sense that as more between-group restrictions are included, the models are hierarchically nested. Nested models can be compared in pairs by calculating the differences in their overall χ^2 values and the related degrees of freedom; the χ^2 difference value ($\Delta\chi^2$) is distributed as χ^2 , with the degrees of freedom equal to the difference in degrees of freedom (Δdf). Historically, evidence in support of invariance has been based on this test: if the $\Delta\chi^2$ value is significant, it suggests that the constraints in the more restrictive model do not hold and therefore that the two models are not equivalent across groups. Similar comparisons can be made based on the S-B χ^2 , except that a correction to this difference value is needed because it is not distributed as χ^2 (Bentler, 2005). However, the use of the $\Delta\chi^2$ has come under criticism, because it is highly sensitive to sample size and therefore it is an impractical and unrealistic criterion on which to base evidence of invariance. Consequently, researchers have based decisions of invariance on alternative criteria.

Cheung and Rensvold (2002) examined the properties of 20 goodness-of-fit statistics within the context of invariance testing and identified three indexes, including the ΔCFI , as providing the best information for evaluating measurement invariance. Regarding ΔCFI , they suggest that this value should not exceed .01. More recently, Chen (2007) conducted simulation studies to examine the performance of various relative fit measures in examining measurement invariance in groups with large sample sizes. As a result of these studies, she made recommendations for particular measures of relative fit, and appropriate cutoff values, which have proven to be informative for examining measurement invariance in large samples. In particular, she recommends that measurement invariance in larger samples should be rejected when $\Delta CFI \geq .01$ and when $\Delta RMSEA \geq .015$. In the present work, we rely on these two measures of relative fit and we adopt Cheung and Rensvold (2002) and Chen's (2007) cutoff values for rejecting measurement invariance based on their practical approach, although we base our analyses on the robust versions of these measures (Δ^*CFI and Δ^*RMSEA).

Results

Descriptive Data

Scores on the BDI-II ranged from 0 to 61. The grand mean was 9.27 ($SD = 8.07$), the median was 7, and the mode was 0. Means, standard deviations, skewness, and kurtosis for

individual BDI-II items are presented in Table 2 for the full sample and means and standard deviations are presented separately by gender and race and ethnic groups.

Baseline Model

Multivariate normality was investigated through Mardia's (1970) multivariate kurtosis coefficient and normalized estimate of multivariate kurtosis. Mardia's multivariate kurtosis coefficient for the full sample was 270.55 and its normalized estimate was 373.62. Corresponding values for Mardia's multivariate kurtosis coefficient and its normalized estimate by group were 243.91 and 271.57 for women, 325.56 and 265.98 for men, 281.39 and 317.26 for Whites, 228.73 and 96.10 for Blacks, 215.47 and 88.10 for Asians, and 237.55 and 117.97 for Latinos. Mardia's normalized multivariate kurtosis estimates can be interpreted like z scores, and Bentler and Wu (2002) suggest that a normalized estimate >3 will lead to chi-square and standard error biases. The probability levels associated with the obtained normalized estimates were all $<.001$ and exceed Bentler and Wu's cutoff. The substantial multivariate kurtosis, which would be expected in a nonclinical sample, and which is consistent with prior research (e.g., Byrne & Stewart, 2006), support the use of the robust statistics for evaluating model fit.

We then conducted a CFA to evaluate the hierarchical four-factor structure supported by Byrne (Byrne et al., 2007; Byrne & Stewart, 2006) in Hong Kong and American adolescents. This model includes three first-order factors—Negative Attitude (NA; Items 1, 2, 3, 5, 6, 7, 8, 9, 10, 14), Performance Difficulty (PD; Items 4, 11, 12, 13, 17, 19), and Somatic Elements (SE; Items 15, 16, 18, 20). A single second-order factor was specified to account for the covariances among these three first-order factors, with all three second-order loadings estimated and the variance of the second-order factor fixed at 1. The loadings for Items 3, 12, and 16 were fixed to 1.00 for purposes of model identification and latent variable scaling.^{2,3} Finally, the residuals associated with NA and PD were constrained to be equal to address the issue of statistical identification at the higher level of the model, given only three first-order factors. Results indicated that this model provided a reasonable fit with the data (although the *CFI values were slightly below recommended levels), $S-B\chi^2_{(187)} = 2095.06$, $p < .001$; *CFI = .923; SRMR = .033; *RMSEA = .037, 90% CI = [.036, .039]. This model fit the data better than (a) a model with all the BDI-II items loading on a single latent factor ($S-B\chi^2_{(189)} = 3897.23$, $p < .001$; *CFI = .851; SRMR = .044; *RMSEA = .052, 90% CI = [.050, .053]), (b) a two-factor structure reported by Beck et al. (1996) for the college student sample data ($S-B\chi^2_{(188)} = 2605.66$, $p < .001$; *CFI = .903; SRMR = .036; *RMSEA = .042, 90% CI = [.040, .043]), (c) a two-factor model obtained by Dozois,

Table 2. Means, Standard Deviations, Skewness (S), and Kurtosis (K) for Individual Items, First-order Factors, and Total Score

	Full sample				Women		Men		White		Black		Asian		Latino	
	M	SD	S	K	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Item 1	0.4	0.6	1.5	2.9	0.4	0.6	0.3	0.5	0.4	0.6	0.4	0.6	0.5	0.6	0.4	0.6
Item 2	0.4	0.6	1.4	1.6	0.4	0.6	0.4	0.6	0.4	0.6	0.3	0.5	0.5	0.6	0.4	0.6
Item 3	0.4	0.6	1.5	1.6	0.4	0.6	0.4	0.6	0.4	0.6	0.3	0.6	0.6	0.7	0.5	0.7
Item 4	0.4	0.6	1.5	2.2	0.4	0.6	0.4	0.6	0.4	0.6	0.4	0.6	0.5	0.6	0.5	0.7
Item 5	0.4	0.6	1.4	2.2	0.4	0.6	0.4	0.6	0.4	0.6	0.4	0.6	0.5	0.6	0.5	0.6
Item 6	0.3	0.6	2.5	6.4	0.3	0.7	0.3	0.7	0.3	0.6	0.4	0.8	0.4	0.7	0.4	0.8
Item 7	0.4	0.7	1.6	1.8	0.5	0.7	0.4	0.7	0.4	0.7	0.4	0.7	0.6	0.8	0.5	0.8
Item 8	0.6	0.7	1.1	0.8	0.6	0.8	0.5	0.7	0.6	0.7	0.5	0.7	0.6	0.8	0.6	0.8
Item 9	0.2	0.4	2.4	6.7	0.2	0.4	0.2	0.4	0.2	0.4	0.1	0.4	0.3	0.5	0.2	0.5
Item 10	0.4	0.8	2.1	3.7	0.5	0.8	0.3	0.7	0.4	0.7	0.5	0.8	0.5	0.9	0.5	0.8
Item 11	0.5	0.7	1.4	2.1	0.5	0.7	0.5	0.7	0.5	0.7	0.5	0.7	0.6	0.7	0.5	0.7
Item 12	0.4	0.6	1.7	3.1	0.4	0.6	0.4	0.6	0.4	0.6	0.4	0.6	0.5	0.7	0.4	0.6
Item 13	0.4	0.7	1.8	3.0	0.5	0.8	0.4	0.6	0.4	0.7	0.4	0.7	0.6	0.8	0.6	0.8
Item 14	0.3	0.6	2.3	4.9	0.3	0.7	0.2	0.6	0.3	0.6	0.3	0.7	0.4	0.7	0.3	0.6
Item 15	0.5	0.6	1.0	0.8	0.6	0.7	0.5	0.6	0.5	0.6	0.5	0.6	0.7	0.7	0.7	0.7
Item 16	0.8	0.8	0.7	0.1	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	1.0	0.7	1.0	0.8
Item 17	0.5	0.6	1.4	1.8	0.5	0.7	0.4	0.6	0.4	0.6	0.5	0.7	0.4	0.7	0.5	0.7
Item 18	0.6	0.7	1.2	1.4	0.6	0.8	0.5	0.7	0.5	0.7	0.6	0.7	0.7	0.8	0.8	0.8
Item 19	0.6	0.7	1.1	0.5	0.6	0.8	0.6	0.7	0.5	0.7	0.5	0.7	0.8	0.8	0.7	0.8
Item 20	0.6	0.6	1.0	0.9	0.6	0.7	0.5	0.6	0.5	0.6	0.5	0.6	0.7	0.7	0.7	0.7
Item 21	0.2	0.6	2.7	7.2	0.3	0.6	0.2	0.5	0.2	0.5	0.3	0.6	0.2	0.5	0.2	0.6
Factor 1 (NA)	3.8	4.2	1.6	3.0	4.0	4.2	3.4	4.0	3.6	4.0	3.4	4.0	4.8	4.6	4.1	4.4
Factor 2 (PD)	3.0	3.0	1.4	2.3	3.1	3.0	2.7	2.9	2.8	2.9	3.0	2.9	3.5	3.1	3.4	3.1
Factor 3 (SE)	2.5	2.1	0.9	0.8	2.6	2.1	2.3	2.0	2.3	2.0	2.3	2.0	3.1	2.1	3.1	2.2
Total Score	9.3	8.1	1.4	2.6	9.8	8.2	8.4	7.8	8.7	7.8	8.8	7.6	11.4	8.7	10.7	8.6

Note. NA = Negative Attitude; PD = Performance Difficulty; SE = Somatic Elements.

Dobson, and Ahnberg (1998) ($S-B\chi^2_{(188)} = 2724.61$, $p < .001$; *CFI = .898; SRMR = .038; *RMSEA = .043, 90% CI = [.041, .044]), or (d) a three-factor model reported by Osman et al. (1997) ($S-B\chi^2_{(184)} = 4685.43$, $p < .001$; *CFI = .819; SRMR = .144; *RMSEA = .058, 90% CI = [.056, .059]). Because this model fit the data better than the other tested models, and because this model has been supported in other analyses, we used it in our analyses evaluating factorial invariance. A diagrammatic representation of the model is depicted in Figure 1. Good internal consistency was obtained for the NA ($\alpha = .85$), PD ($\alpha = .77$), and SE ($\alpha = .73$) first-order factors.

Before evaluating factorial invariance between groups, we first estimated the model separately in each group (defined first by gender and subsequently by race and ethnicity). Results from these analyses (available from the first author on request) indicated that in every case, the hierarchical four-factor model provided a reasonably good fit with the data (again, with slightly lower than recommended values for *CFI). Means and standard deviations for the three first-order factors based on the sums of the items on each factor, assuming equal weighting, are presented in

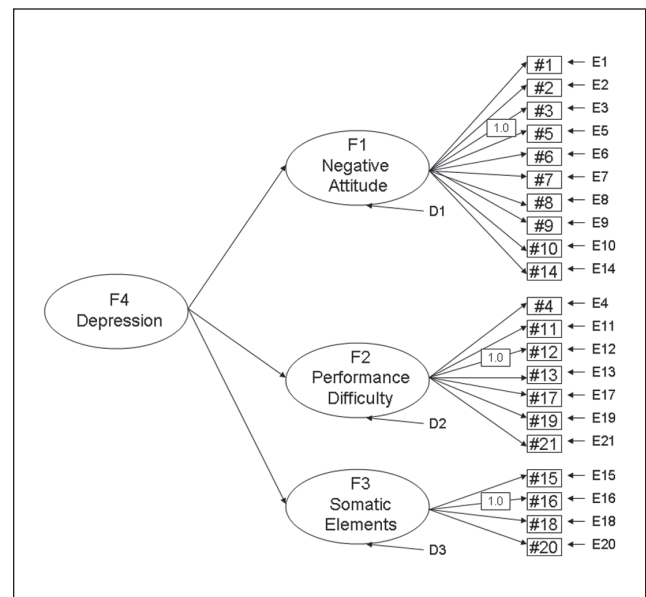


Figure 1. Hypothesized hierarchical model of the factorial structure of the Beck Depression Inventory–Second Edition (BDI-II)

Table 3. Goodness-of-Fit Statistics for Tests for Invariance of Beck Depression Inventory–II Hierarchical Structure for Men and Women

Model and constraints	S-B χ^2	df	*CFI	SRMR	*RMSEA	*RMSEA 90% CI	Δ *CFI	Δ *RMSEA
1. None	2199.94	374	.924	.034	.036	[.035, .038]	—	—
2. First-order loadings	2263.22	392	.922	.040	.036	[.035, .037]	-.002	.000
3. First- and second-order loadings	2267.21	395	.922	.042	.036	[.034, .037]	.000	.000
4. First- and second-order loadings; factor variances	2288.65	397	.921	.044	.036	[.035, .037]	-.001	.000
5. First- and second-order loadings; factor and error variances	2686.35	418	.909	.053	.038	[.037, .040]	-.012	.002

Note. *CFI = robust version of comparative fit index; SRMR = standardized root mean square residual; S-B χ^2 = Satorra–Bentler scaled χ^2 ; *RMSEA = robust version of root mean square error of approximation; CI = confidence interval.

Table 2 for the full sample and separately by gender and race and ethnic group; factor loadings and other estimated parameters for this model for the full sample and separately by gender and race and ethnic group are available from the first author.

Gender Comparisons

In Table 3, we present the results from the five models that examine factor invariances between men and women in our data. As described above, Model 1 imposes the same configural second-order factor structure on the data from both genders but forces no between-group constraints. It provides a good fit to the data and thus becomes the baseline against which the following four models are compared. Models 2 and 3 constrain between-group equalities in the model loadings. Both these models compare quite favorably with the baseline model without these constraints. The resulting Δ *CFI and Δ *RMSEA are negligible and certainly fall well below the recommended cutoff values for rejecting invariance. Therefore, we conclude that measurement invariance of the factor loadings is supported between genders.

Given equal loadings, Model 4 imposes equality constraints on the variances of the second-order factor and variances of the disturbances of the first-order factor. The comparison of this model with the baseline model yields very small changes in the fit indexes, suggesting that there is no evidence for unequal variability in the latent factors or traits responsible for covariation among the BDI-II items. Finally, Model 5 additionally constrains the variances of the disturbances of the indicators to be equal between groups, thus forcing equality of the two variance/covariance matrices between the genders in the context of this second-order factor model. Here, we have evidence of a relative lack-of-fit, as the increment to *CFI of .012 exceeds the recommended cutoff value for the conclusion of invariance. Hence, although the models suggest both measurement invariance (factor-loading invariance) and equality of the variances of the latent factors, we conclude that complete equivalence of the variance/covariance matrices must

be rejected because of unequal variances at the level of the individual item disturbances.

Race and Ethnic Comparisons

Table 4 presents the results for the same multigroup models tested across groups defined by race or ethnicity. In each case, we compared one minority group—Blacks, Asians, or Latinos—with the majority (i.e., White) group. For each comparison, (a) the first three multigroup models provide a good fit with the data; and (b) for comparisons between Model 1 (the configural model) and Models 2 and 3, the Δ *CFI never exceeded .001 and the Δ *RMSEA never exceeded .001, which fall far below the recommended cutoffs for rejecting invariance of .01 and .015, respectively. Therefore, we conclude that there is measurement invariance between Whites and minority groups, with equivalent first- and second-order factor loadings in the context of the four-factor hierarchical structure that forms the basis of these comparisons.

When we further impose variance constraints on the groups (i.e., Models 4 and 5), we reach the further conclusion that the fit of the model does not deteriorate when we assume equal variance in the latent factors between Whites and minority groups (i.e., comparing Model 4 with the baseline model). Additionally, Model 5, in which equal items disturbance variances are imposed between groups, continues to fit the data as well as the configural model; this holds true for the comparisons between Whites and each minority group. Thus, the conclusion is not only that the BDI-II shows measurement invariance across Whites and minority groups, but additionally, within the confines of this second-order factor model, that there are equivalent variances and covariances between Whites and minority groups in the variances of the underlying latent factors and additionally in the variances of the individual items.

Discussion

Consistent with Byrne and colleagues' (Byrne et al., 2007; Byrne & Stewart, 2006) research with high school students,

Table 4. Goodness-of-Fit Statistics for Tests for Invariance of Beck Depression Inventory–II Hierarchical Structure for Race and Ethnic Groups

Model and constraints	S-B χ^2	df	*CFI	SRMR	*RMSEA	*RMSEA 90% CI	Δ *CFI	Δ *RMSEA
Whites and Blacks								
1. None	1948.01	374	.913	.040	.039	[.037, .040]	—	—
2. First-order loadings	1979.79	392	.912	.050	.038	[.036, .040]	-.001	-.001
3. First- and second-order loadings	1991.96	395	.911	.056	.038	[.036, .040]	-.001	.000
4. First- and second-order loadings; factor variances	1987.99	397	.912	.055	.038	[.036, .040]	.001	.000
5. First- and second-order loadings; factor and error variances	1955.18	418	.911	.063	.036	[.035, .038]	-.001	-.002
Whites and Asians								
1. None	1950.22	374	.922	.040	.039	[.037, .041]	—	—
2. First-order loadings	1970.41	392	.921	.047	.038	[.036, .040]	-.001	-.001
3. First- and second-order loadings	1966.29	395	.922	.063	.038	[.036, .040]	.001	.000
4. First- and second-order loadings; factor variances	1967.33	397	.922	.064	.038	[.036, .039]	.000	.000
5. First- and second-order loadings; factor and error variances	1815.55	418	.918	.073	.035	[.033, .036]	-.004	-.003
Whites and Latinos								
1. None	1919.95	374	.923	.036	.038	[.036, .039]	—	—
2. First-order loadings	1945.65	392	.923	.044	.037	[.035, .038]	.000	-.001
3. First- and second-order loadings	1939.30	395	.923	.054	.037	[.035, .038]	.000	.000
4. First- and second-order loadings; factor variances	1936.44	397	.923	.054	.036	[.035, .038]	.000	-.001
5. First- and second-order loadings; factor and error variances	1923.54	418	.916	.062	.035	[.033, .037]	-.007	-.001

Note. *CFI = robust version of comparative fit index; SRMR = standardized root mean square residual; S-B χ^2 = Satorra–Bentler scaled χ^2 ; *RMSEA = robust version of root mean square error of approximation; CI = confidence interval.

we found evidence that the factor structure of the BDI-II in college students is best represented by a hierarchical structure, consisting of three first-order factors and one second-order factor. To the best of our knowledge, this is the first study that has tested the hierarchical four-factor model in college students. Osman et al. (1997) evaluated and found support for a related three-factor structure for the BDI-II in college students (which was replicated by Carmody, 2005; Vanheule, Desmet, Groenvynck, Rosseel, & Fontaine, 2008), but that model differed from the current model with respect to item placement on factors. Their model also differed from ours in that the three first-order factors were simply allowed to covary rather than arguing that their covariances could be accounted for by a single higher order factor. The present hierarchical model seems both more parsimonious and is also supported by work other than our own (Byrne et al., 2007; Byrne & Stewart, 2006). The conclusion that the best model includes a second-order factor measuring a general factor supports the common practice of using a single composite score to represent a general or global evaluation of severity of depressive symptoms.

With respect to factorial invariance, we found evidence for measurement invariance in the context of the hierarchical four-factor structure of the BDI-II between women and men and between Whites and racial (Blacks, Asians) or ethnic (Latinos) minority groups. Specifically, across models

in which there was increasingly restricted parameterization on the variance/covariance matrices of the indicators, there was consistent evidence that the hierarchical four-factor structure provided good fit with the data. Furthermore, the Δ *CFI and Δ *RMSEA values for comparisons between Model 1 (i.e., the configural model) and Models 2 and 3 were all negligible. Taken together, these results provide strong evidence for measurement invariance for the BDI-II across gender, race, and ethnicity, which lead us to conclude that the BDI-II does not measure different hypothetical traits for one group than another. Because factorial invariance was obtained for analyses constraining factor structure and loadings, these results suggest that it is appropriate to compare processes (i.e., correlates of depressive symptoms) across groups.

Turning next to the results obtained for analyses constraining variances (Models 4 and 5), results suggest that variances of the latent first-order and second-order factors between women and men or between Whites and racial (Blacks, Asians) or ethnic (Latinos) minority groups are equivalent. Specifically, the Δ *CFI and Δ *RMSEA values for comparisons between Model 1 (i.e., the configural model) and Model 4 were all negligible. Furthermore, the Δ *CFI and Δ *RMSEA values for comparisons between Model 1 and Model 5 suggest that Model 5 did not provide as good of fit when groups were defined in terms of gender,

but provided as good of fit when groups were defined in terms of race or ethnicity. Thus, there was evidence for unequal variances at the level of individual items in comparisons between men and women, but no evidence for unequal variances at the level of individual items in comparisons between Whites and racial and ethnic minority groups. Taken together, these results provide particularly strong evidence for measurement invariance for the BDI-II across gender, and even stronger evidence for measurement invariance across groups defined by race or ethnicity.

This study is based only on college students, and elevated scores on the measure, without diagnostic information, should be interpreted as measuring dysphoria or nonspecific negative affectivity rather than clinical depression (Coyne, 1994; Kendall, Hollon, Beck, Hammen, & Ingram, 1987). Although the results are important for studies using student samples, it is unclear whether similar results would be obtained in clinical samples or nonclinical samples of people at different ages. It is also worth noting that race and ethnicity were defined by self-report, and that there is considerable diversity within these broad categories with respect to acculturation; family, cultural, and religious background; country of origin; and other related factors. The findings, therefore, do not address measurement invariance of the BDI-II for subgroups that exist within larger categories defined by race or ethnicity, particularly as college students may represent the most acculturated members of their communities. As such, the use of college students may underestimate the impact of values and practices of a minority group on BDI-II responses.

In sum, the results provide evidence for measurement invariance of the BDI-II across groups defined by gender, race, or ethnicity. Specifically, evidence was obtained for invariance of the factor structure and loadings of the BDI-II across groups, which supports continued investigation in comparisons of process (i.e., correlates of depressive symptoms) across groups.

Acknowledgments

We would like to thank the researchers who provided us with the raw data from their published studies to include in the current analyses.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article:

This work was supported by a grant from the National Alliance for Research on Schizophrenia and Depression.

Notes

1. Studies evaluating the measurement equivalence of the BDI-II across groups other than demographic groups (e.g., across different types of patient groups) are not included in this review.
2. The choice of these items to scale the factors is consistent with Byrne and Stewart's (2006) choice and is also consistent with exploratory factor analyses reported by them and replicated by us, with the current data, in which these were the three highest loading items in a three-factor solution.
3. Item 21 was not included in the Byrne et al. (2007) study because it was considered objectionable by Hong Kong school principals; we included this item and allowed it to load on the Performance Difficulty factor.

References

- Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. New York, NY: Guilford Press.
- Beck, A. T., & Steer, R. A. (1987). *Manual for the Beck Depression Inventory*. San Antonio, TX: Psychological Corporation.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory-II manual*. San Antonio, TX: Psychological Corporation.
- Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, 4, 561-571. doi:10.1001/archpsyc.1961.01710120031004
- Bentler, P. M. (2005). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Wu, E. J. C. (2002). *EQS for Windows user's guide*. Encino, CA: Multivariate Software.
- Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Byrne, B. M., & Baron, P. (1994). Measuring adolescent depression: Tests of equivalent factorial structure for English and French versions of the Beck Depression Inventory. *Applied Psychology: An International Review*, 43, 33-47. doi:10.1111/j.1464-0597.1994.tb00808.x
- Byrne, B. M., Baron, P., & Balev, J. (1996). The Beck Depression Inventory: Testing for its factorial validity and invariance across gender for Bulgarian adolescents. *Personality and Individual Differences*, 21, 641-651. doi:10.1016/0191-8869(96)00134-1
- Byrne, B. M., Baron, P., & Campbell, T. L. (1993). Measuring adolescent depression: Factorial validity and invariance of the Beck Depression Inventory across gender. *Journal of Research on Adolescence*, 3, 127-143. doi:10.1207/s15327795jra0302_2
- Byrne, B. M., Baron, P., & Campbell, T. L. (1994). The Beck Depression Inventory [French version]: Testing for gender-invariant factorial structure for nonclinical adolescents. *Journal of Adolescent Research*, 9, 166-179. doi:10.1177/074355489492003

- Byrne, B. M., Baron, P., Larsson, B., & Melin, L. (1996). Measuring depression for Swedish nonclinical adolescents: Factorial validity and equivalence of the Beck Depression Inventory across gender. *Scandinavian Journal of Psychology, 37*, 37-45. doi:10.1111/j.1467-9450.1996.tb00637.x
- Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure: A look beneath the surface. *Journal of Cross-Cultural Psychology, 30*, 555-574. doi:10.1177/0022022199030005001
- Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*, 287-321. doi:10.1207/s15328007sem1302_7
- Byrne, B. M., Stewart, S. M., Kennard, B. D., & Lee, P. W. H. (2007). The Beck Depression Inventory-II: Testing for measurement equivalence and factor mean differences across Hong Kong and American adolescents. *International Journal of Testing, 7*, 293-309. doi:10.1080/15305050701438058
- Carmody, D. P. (2005). Psychometric characteristics of the Beck Depression Inventory-II with college students of diverse ethnicity. *International Journal of Psychiatry in Clinical Practice, 9*, 22-28. doi:10.1080/13651500510014800
- Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption-Innovation Inventory using multi-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35*, 169-199. doi:10.1207/S15327906MBR3502_2
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*, 464-504. doi:10.1080/10705510701301834
- Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12*, 471-492. doi:10.1207/s15328007sem1203_7
- Cheung, C. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. doi:10.1207/S15328007SEM0902_5
- Coyne, J. C. (1994). Self-reported distress: Analog or ersatz depression? *Psychological Bulletin, 116*, 29-45. doi:10.1037/0033-2909.116.1.29
- Dozois, D. J. A., & Covin, R. (2004). The Beck Depression Inventory-II (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS). In M. J. Hilsenroth & D. L. Segal (Eds.), *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment* (pp. 50-69). Hoboken, NJ: John Wiley.
- Dozois, D. J. A., Dobson, K. S., & Ahnberg, J. L. (1998). A psychometric evaluation of the Beck Depression Inventory-II. *Psychological Assessment, 10*, 83-89. doi:10.1037/1040-3590.10.2.83
- Hambrick, J. P., Rodebaugh, T. L., Balsis, S., Woods, C. M., Mendez, J. L., & Heimberg, R. G. (2010). Cross-ethnic measurement equivalence of measures of depression, social anxiety, and worry. *Assessment, 17*, 155-171. doi:10.1177/1073191109350158
- Harari, M. J., Waehler, C. A., & Rogers, J. R. (2005). An empirical investigation of a theoretically based measure of perceived wellness. *Journal of Counseling Psychology, 52*, 93-103. doi:10.1037/0022-0167.52.1.93
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi:10.1080/10705519909540118
- Joiner, T. E., Walker, R. L., Pettit, J. W., Perez, M., & Cukrowicz, K. C. (2005). Evidence-based assessment of depression in adults. *Psychological Assessment, 17*, 267-277. doi:10.1037/1040-3590.17.3.267
- Kendall, P. C., Hollon, S. D., Beck, A. T., Hammen, C. L., & Ingram, R. E. (1987). Issues and recommendations regarding use of the Beck Depression Inventory. *Cognitive Therapy and Research, 11*, 289-299. doi:10.1007/BF01186280
- Klibert, J. J., Langhinrichsen-Rohling, J., & Saito, M. (2005). Adaptive and maladaptive aspects of self-oriented versus socially prescribed perfectionism. *Journal of College Student Development, 46*, 141-156. doi:10.1353/csd.2005.0017
- Lewandowski, K. E., Barrantes-Vidal, N., Nelson-Gray, R. O., Clancy, C., Kepley, H. O., & Kwapil, T. R. (2006). Anxiety and depression symptoms in psychometrically identified schizotypy. *Schizophrenia Research, 83*, 225-235. doi:10.1016/j.schres.2005.11.024
- Little, T. D. (1997). Mean and covariance structures (MACS) analysis of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76. doi:10.1207/s15327906mbr3201_3
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519-530. doi:10.1093/biomet/57.3.519
- Norton, P. J. (2005). A psychometric analysis of the Intolerance of Uncertainty Scale among four racial groups. *Anxiety Disorders, 19*, 699-707. doi:10.1016/j.janxdis.2004.08.002
- Okazaki, S. (2002). Self-other agreement on affective distress scales in Asian Americans and White Americans. *Journal of Counseling Psychology, 49*, 428-437. doi:10.1037/0022-0167.49.4.428
- Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Gutierrez, P. M., & Chiros, C. E. (1997). Factor structure and psychometric characteristics of the Beck Depression Inventory-II. *Journal of Psychopathology and Behavioral Assessment, 19*, 359-376. doi:10.1007/BF02229026
- Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi square statistics in covariance structure analysis. In *American Statistical Association 1988 proceedings of the business and economic sections* (pp. 308-313). Alexandria, VA: American Statistical Association.
- Scarpa, A., Fikretoglu, D., Bowser, F., Hurley, J. D., Pappert, C. A., Romero, N., & Van Voorhees, E. (2002). Community violence exposure in university students: A replication and extension. *Journal of Interpersonal Violence, 17*, 253-272. doi:10.1177/0886260502017003002
- Storch, E. A., Roberti, J. W., & Roth, D. A. (2004). Factor structure, concurrent validity, and internal consistency of the Beck

- Depression Inventory—Second Edition in a sample of college students. *Depression and Anxiety*, *19*, 187-189. doi:10.1002/da.20002
- Vanheule, S., Desmet, M., Groenvynck, H., Rosseel, Y., & Fontaine, J. (2008). The factor structure of the Beck Depression Inventory-II: An evaluation. *Assessment*, *15*, 177-187. doi:10.1177/1073191107311261
- Whisman, M. A., Perez, J. E., & Ramel, W. (2000). Factor structure of the Beck Depression Inventory—Second Edition (BDI-II) in a student sample. *Journal of Clinical Psychology*, *56*, 545-551. doi:10.1002/(SICI)1097-4679(200004)56:4<545::AID-JCLP7>3.0.CO;2-U
- Wiebe, J. S., & Penley, J. A. (2005). A psychometric comparison of the Beck Depression Inventory-II in English and Spanish. *Psychological Assessment*, *17*, 481-485. doi:10.1037/1040-3590.17.4.481
- Witcherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726-728. doi:10.1037/0003-066X.61.7.726
- You, S., Merritt, R. D., & Conner, K. R. (2009). Do gender differences in the role of dysfunctional attitudes in depressive symptoms depend on depression history? *Personality and Individual Differences*, *46*, 218-223. doi:10.1016/j.paid.2008.10.002