Psychology
2012. Vol.3, Special Issue, 762-774
Published Online September 2012 in SciRes (http://www.SciRP.org/journal/psych)
http://dx.doi.org/10.4236/psych.2012.329116

Scalar Equivalence in Self-Rated Depressive Symptomatology as Measured by the Beck Depression Inventory-II: Do Racial and Gender Differences in College Students Exist?

Lisa M. Hooper[1,3], Lixin Qu[1], Cindy A. Crusto[2], Lauren E. Huffman[1]
[1]The University of Alabama, Tuscaloosa, USA
[2]Yale School of Medicine, The Consultation Center, New Haven, USA
[3]Hooper Research Lab, Tuscaloosa, USA
Email: lhooper@bamaed.ua.edu

Received June 9th, 2012; revised July 12th, 2012; accepted August 11th, 2012

Using item response theory and confirmatory factor analysis, the current investigation examined the equivalence in responses derived from the widely used 21-item Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996) among 1229 college students (mean = 21.15, *SD* = 6.19) in the United States. Results from differential item functioning analyses indicated that the items endorsed by Black American and White American college students were slightly different. However, items endorsed by female and male college students were almost invariant. The results of the study found partial support for using the BDI-II in college student populations. Directions for future culturally tailored assessment and research are proffered.

Keywords: Differential Item Functioning (DIF); Depression; Depressive Symptoms; Race; Gender; Scalar Equivalence; Item Response Theory (IRT); Confirmatory Factor Analysis; American College Students

Introduction

In 2007, depression-related suicide was the third leading cause of death for adolescents and emerging adults (ages 12 to 24) (Centers for Disease Control and Prevention, 2010). Depression is one of the most significant, disabling, and deleterious mental health disorders in all populations, including college students (American College Health Association, 2009; Blanco et al., 2008; Hankin, 2002; World Health Organization, 2002). For emerging adults, depression is the most common clinical disorder, with prevalence rates estimated to approach 11% (American College Health Association, 2009; Blanco et al., 2008). Importantly, findings from the National Comorbidity Study-Revised (NCS-R; Kessler et al. 2003) suggested that many adults who had a reported history of a depressive episode in the previous year failed to receive adequate treatment (i.e., guideline-concordant care such as pharmacotherapy or psychotherapy; see American Psychiatric Association, 2000b; Gonzalez, Vega, Williams, Tarraf, West, & Neighbors, 2010) for their depression. Therefore, there is an urgent need to better understand why depression continues to be undertreated and often undetected, including in college student populations (Carmody, 2005; Kadison, 2004; Kisch, Leino, & Silverman, 2005; Tjia, Givens, & Shea, 2005).

One factor that may account for the undertreatment of depression is the lack of detection of depressive symptoms in individuals (American College Health Association, 2009; Carmody, 2005; Hooper, 2010; Leino & Kisch, 2005). An obvious first step in uncovering factors that may impede the effective treatment of depression is clarifying the barriers that affect providers' ability to detect depressive signs and symptoms and their competency to make an accurate diagnosis. Toward this end, instruments or assessment tools that produce reliable and valid scores are paramount (Boughton & Street, 2007). An additional consideration is the extent to which instruments are culturally, linguistically, and clinically sensitive (Anderson & Mayes, 2010; Manly, 2006). With the increasing focus on racial and cultural diversity in the human helping disciplines (see Chao & Otsuki-Clutter, 2011; Day, 1996; McHorney & Fleischman, 2006) discussions on the extent to which assessment, diagnosis, and treatment methods are culturally responsive and relevant are important and timely. McHorney and Fleischman (2006) suggested, "If items in outcome measures are biased, detection rates can be biased (overestimated or underestimated), leading to over- and under-detection and over- and under-treatment" (p. s205).

In this article, we first provide a brief overview of the importance of measures that compose scores that reliably and validly assess for depressive symptoms. Then we review the empirical literature on depression and depressive symptoms in college students, including the implications of race and gender for the presentation, detection, and screening of depression. Next we describe the research design and summarize the results of the current investigation conducted to test the extent to which the Beck Depression Inventory-II (BDI-II; Beck, Steer, & Brown, 1996) demonstrates scalar equivalence for female and male college students and for Black American and White American college students. We describe two rigorous differential item functioning (DIF) analytic procedures—item response theory and confirmatory factor analysis—that were employed to detect if the endorsements of depressive symptoms among these four groups are biased or equivalent. We conclude with the implications of the findings and directions for culturally tailored assessment and future research.

Background Literature

The Healthy People 2020 Initiative characterizes major depression as a national priority (US Department of Health and Human Services, n.d.). Empirical studies have suggested that the manifestation and characterization of depressive symptoms may be influenced by demographic, familial, and ecological factors (e.g., race, gender, neighborhoods, and discrimination) (Anderson & Mayes, 2010; Gregorich, 2006; Iwata & Buka, 2002). Therefore, another step in optimally treating depression and accurately detecting depressive signs and symptoms involves ensuring that racially and culturally diverse individuals are included in research studies that examine diagnosis and treatment methods and measures for medical conditions and mental health disorders such as depression (see Manly, 2006). The mandate from the National Institutes of Health (NIH) and the recently established National Institute of Minority Health and Health Disparities explicitly underlines this proposition: the inclusion of racial and cultural minority participants in research studies to inform practice, which includes reliable assessments and accurate diagnoses (National Institutes of Health, 2002; National Institute of Minority Health, 2010). Other mental health services and intervention researchers have also underscored the importance of inclusion of vulnerable and racial minority populations in the development of measures and controlled clinical trials (Anderson & Mayes, 2010; Manly, 2006; McHorney & Fleischman, 2006; National Institute of Mental Health, 2010; National Institutes of Health, 2002; Paniagua, 1994; Sperry, 2010).

Depression in College Students

Some scholars have asserted that depression and anxiety disorders are the leading clinical issues with which college students must contend and with which health care providers (e.g., college counseling center staff) must be prepared and competent to face (Carmody, 2005; Kisch et al., 2005). Depression care guidelines and published recommendations have relevance for the current investigation about one element of depression care: the assessment of depressive symptoms. With regard to the criticality of assessment in depression care, empirical research consistently has found an association between depressive symptoms and suicide behavior as well as other negative sequelae (e.g., anxiety symptomatology, disordered eating behaviors and attitudes, alcohol and drug use, and interpersonal violence) in college student populations (Centers for Disease Control and Prevention, 2010; Kisch et al., 2005; Wilcox et al., 2010). Taken together, these are significant clinical issues and functional consequences that are often evinced in college and university populations (American College Health Association, 2009; Arria et al., 2009; Blanco et al., 2008; Centers for Disease Control and Prevention, 2010; Eisenberg, Gollust, Golberstein, & Hefner, 2007; Furr, Westefeld, McConnell, & Jenkins, 2001; Kisch et al., 2005; Nolen-Hoeksema, 1990; Wilcox et al., 2010). As previously mentioned, suicide is also currently the third leading cause of death among adolescents and emerging adults (Centers for Disease Control and Prevention, 2010). Because of the life-threatening element (i.e., suicidal ideation and suicidality) highly associated with depression, scores that are derived from measures that reliably and validly capture depression are paramount.

Toward this end, reliable and valid assessments (i.e., scores) to capture depressive symptoms are needed, but assessments that demonstrate cultural and linguistic equivalence also are needed (Anderson & Mayes, 2010; Eisenberg et al., 2007; Manly, 2006). Iwata and Buka (2002) stated, "Specific response patterns and psychometric properties of assessment instruments across ethnic/cultural populations require further investigation" (p. 2243). Consistent with most psychometricians' suggestions (see Borsboom, 2006), we believe the ideal scenario is that widely used assessments such as the BDI-II (Beck et al., 1996) should be equivalent (i.e., absent from item- and scale-level biases) across cultural and ecological factors, such as race, gender, geographical regions, socioeconomic statuses, and so forth.

Depression and Gender

A commonly reported claim is that depression is more prevalent in females than males (Beck et al., 1996; Hankin, 2002; Kessler et al., 1994; Nolen-Hoeksema, 1990; World Health Organization, 2002), although this commonly recounted assertion is based primarily on cross-sectional studies. The results that have accumulated from epidemiological studies offer some support for gender-related differences in depression and depressive symptoms, and they add to the results derived from cross-sectional studies. Epidemiological studies have suggested that gender differences in depression and depressive symptoms emerge during adolescence (see Hankin, 2002; Hankin & Abramson, 2001; Nolen-Hoeksema, 1990; Nolen-Hoeksema & Girgus, 1994; Rao & Chen, 2009). Results from studies composed of college students have suggested the relation among gender and depression and depressive symptoms is inconsistent (see Gladstone & Koenig, 1994; Nolen-Hoeksema & Girgus, 1994; Silverstein, 1999; Steer & Clark, 1997). Therefore, a more accurate refrain may be that gender-related differences in depression and depressive symptoms are equivocal, in particular among college-aged populations.

Indeed, gender-related variances in depressed mood and symptoms are unclear and not well understood—not only during the developmental stage of emerging adulthood but also across the entire lifespan (Eaton et al., 2011; Hooper, 2010; Rao & Chen, 2009). Moreover, the commonly reported assertion that women have higher levels of depressive symptoms and greater prevalence rates of major depressive disorder is not consistently found in the empirical literature. For example, Steer and Clark (1997) found that male college student-respondents reported levels of depressive symptoms similar to those reported by female college student-respondents. In another example, Silverstein (1999) suggested that gender differences disappear when anxiety and somatic symptoms are statistically controlled for. In other words, when somatic and anxiety symptoms are statistically controlled for gender differences related to depression are nonexistent. Silverstein found "large gender differences in the prevalence of anxious somatic depression among samples of high school students, college students, and adults" (p. 480). He concluded, however, that there are no gender differences in pure depression. This conclusion is buttressed by findings in other empirical studies (see Gladstone & Koenig, 1994; Nolen-Hoeksema & Girgus, 1994).

In contrast, in a study regarding the psychometric properties of the BDI-II in a college student sample, Carmody (2005) found statistically significant gender differences. Specifically, Carmody reported that female participants had higher mean scores for depressive symptoms than their male counterparts. Moreover, the college student participants' scores reported in Carmody's study were comparable to those self-rated scores of

college students found in the Beck et al.'s (1996) validation study. Carmody also explored differences in respondents' item-level scores derived from the BDI-II based on ethnicity and gender, finding differences at the item level based on gender (i.e., BDI-II items 1, 15, 10, and 20). Osman and colleagues (1997) also found statistically significant differences in gender-based comparisons of depressive symptoms: Female college students reported higher levels on six items (BDI-II items 1, 7, 10, 17, 20, and 21) than male college students reported. With regard to BDI-II total scale score comparisons in Osman et al.'s study, there were significant difference between males (mean = 9.41) and females (mean = 11.88) as well.

Relevant to the current investigation, few studies have investigated the psychometric properties (i.e., scalar equivalence for females and males) of the measures that are often used to assess for gender differences. The current investigation allows for cross-gender comparisons at the item level. Determining whether the often-reported gender differences in depression and depressive symptomatology are true and real, rather simply reflecting differences in items or measurement bias (i.e., DIF) evidenced on the BDI-II, has implications for assessment, diagnosis, and treatment on college campuses and in the broader clinical community. Clarifying the role of gender in the manifestations of depression has important implications for optimal depression care and management (assessment, diagnosis, and treatment). It may be that gender-focused and gender-tailored treatment for depression may be more efficacious and effective than current treatment practices (see Anderson & Mayes, 2010; Eaton et al., 2011; van de Vijver & Tanzer, 2006). Hankin and Abramson (2001) and Rao and Chen (2009) indicated that the gender differences seen in some studies are consistent and that they consistently occur across race and ethnic groups, although Culbertson (1997) suggested that the converse is true, that is, gender differences vary based on cultural factors such as racial and ethnic group membership.

Importantly, some of the disagreement in the literature related to the differential effects of gender on depression and depressive symptoms may be explained by measurement issues (Anderson & Mayes, 2010). Boughton and Street (2007) contended that gender differences related to depression may be a result of the questions or items that appear on depression screening tools. Specifically, they asserted, "These questions may reflect too narrow of a definition of depression that fails to include symptoms associated with depression in men" (p. 194). They concluded that some self-rated assessment tools may overestimate depressive symptoms in women and underestimate depressive symptoms in men, leading to inaccurate depression care (assessment, diagnosis, and treatment recommendations) and prevalence rates.

Cochran and Rabinowitz (2003) advocated for gender-sensitive assessment and intervention strategies of depression given that the correlates, symptom presentation, and course of depression can be and often are different for men and women. For instance, compared to women, depression in men is more likely to be related to issues such as gender-role conflict, and the symptom presentation is more likely to be related to issues such as aggression, physical and sexual risk-taking, chronic anger, interpersonal conflict, work-related conflict, substance use and abuse, and criminal behavior (Kilmartin, 2005). Given the masculine-related presentation of depressive symptoms that may exist, is it essential that instruments used to assess depression are sensitive to these gender-related differences and accurately identify and classify depression in males and females (see Fields & Cochran, 2011).

Depression and Race

Compared to what is known about the relation between depression (and depressive symptoms) and gender, even less is known about possible differences in the manifestations of depression and depressive symptoms based on race (Coyne & Marcus, 2006; George & Lynch, 2003). The *Unequal Treatment* report of the Institute of Medicine (2002) outlined numerous factors that may relate to presentation of mental health symptoms, misdiagnosis of mental health disorders, and differential treatments and services, including depression care, based on race. Specifically, racial and cultural factors have long been conjectured to relate to the establishment of accurate diagnoses (e.g., depression, schizophrenia, bipolar disorder, and eating disorders). However, as contended by George and Lynch, "The existence, nature, and strength of race differences in mental health remain unclear after several decades of research" (p. 353). In addition, George and Lynch suggested, "Despite a voluminous research base, the basic question of whether blacks and whites differ in levels of depression and psychological distress remains unclear" (p. 353). A careful review of the empirical literature reveals the lack of consistency in race-focused empirical studies. Therefore, similar to gender differences, the most accurate refrain for the variance in depression and depressive symptoms based on race may be that race-related variances in depression and depressive symptoms are equivocal.

The lack of clarity and definitiveness about real differences in depression and depressive symptoms based on race have been described in the literature. George and Lynch (2003) suggested that some of the lack of clarity in the literature results because researchers have drawn conclusions based on the combined differential effects both of a diagnosis of depression and of depressive symptoms. When a total scale score from a given measure (e.g., BDI-II, Center for Epidemiologic Studies Depression Scale [CES-D; Radloff, 1977], and so forth) is used to make comparisons, the results may be different from those obtained from comparisons based on the items of the measure (see Teresi, Ramirez, Lai, & Silver, 2008).

Some empirical evidence derived from adult-focused studies has shown differential depressive symptoms as well as a differential probability of being diagnosed with depression may be based on race (Coyne & Marcus, 2006; Dunlop, Song, Lyons, Manheim, & Chang, 2003). For example, Leino and Kisch (2005), in their study of college students, reported that Black American students were less likely to be diagnosed with depression than their White American counterparts. Moreover, cross-sectional and epidemiological studies have—for the most part—demonstrated racial and ethnic differences in depression and depressive symptoms in adult samples (Kessler et al., 1994).

With relevance to the current study, few studies have examined DIF in depression measures. Most studies that have been conducted have focused on the CES-D (see Teresi et al., 2008, for a comprehensive review). Fewer studies have examined race-related endorsement patterns based on items of the BDI-II (Beck et al., 1996). Only one study was located that has examined DIF in college student populations: Carmody (2005), who examined the psychometric properties of the BDI-II with a sample of racially diverse college students. He found variance

in items endorsed by racially diverse American students. Specifically, DIF was evidenced on three items (BDI-II items 11, 14, and 17). White American students had higher scores on item 11 (agitation) and item 14 (worthlessness) than did Asian American students. White American students also had higher scores on item 17 (irritability) than did Latino American students. It is noteworthy that Carmody's study of college students found no differences in the BDI-II total score based on racial groups. Carmody concluded that the lack of differences in depressive symptom profiles based on race (only three items resulted in DIF) may be related to the commonality of the college experience, or else "college school culture" may have superseded or attenuated any racial or ethnic differences relative to the depressive symptomatology.

The BDI-II has been used with a range of populations, including racially diverse college students (Carmody, 2005; Storch, Roberti, & Roth, 2004; Whisman, Perez, & Ramel, 2000; Wilcox et al., 2010). Despite considerable research supporting differences in depression and depressive symptoms based on race, some evidence indicates there are no differences among racial and ethnic groups as well. Given the dearth of DIF studies, more research focused on scalar equivalence is clearly needed.

The Current Investigation

Our brief review of the empirical literature suggests that cultural differences in the presentation, manifestation, and endorsement of select depressive symptoms often—but not always—vary by race and gender in many populations (Boughton & Street, 2007; Eaton et al., 2011). However, the accumulated results for racial and gender differences in the specific population of college students remain most unclear.

As previously mentioned, one of the instruments most commonly used to screen for a probable diagnosis of depression and depressive symptomatology is the BDI-II (Beck et al., 1996). In spite of its wide use with a range of diverse populations, including college students, few studies have examined the extent to which the inventory demonstrates scalar equivalence in college student populations as well as other populations. Some researchers have suggested that the factor structure of the BDI-II differs based on the type of sample (e.g., clinical vs. nonclinical) (Carmody, 2005; Storch et al., 2004; Teresi et al., 2008). Likewise, it is assumed that the factor structure of the BDI-II may vary based on the cultural background of the sample (Black Americans vs. White Americans). Supporting the need for the current investigation, Santor, Zuroff, Cervantes, Palacios, and Ramsay (1995) stated, "How individuals endorse items on a depression inventory may vary across *items* on a single measure of depression, across *measures* of depression, as well as across levels of *depressive severity*" (p. 131; emphasis added). Moreover, the validity of the findings derived from the BDI-II is only as good as the validity of the scores that are derived from the BDI-II and the items that compose it (Harachi, Choi, Abbott, Catalano, Bliesner, 2006; McHorney & Fleischman, 2006; van de Vijver & Tanzer, 2004). Given that establishing validity is an ongoing process, studies that add to the accumulating evidence in the literature on the possible biases and equivalence of the BDI-II scores and items are important and needed (Schmidt & Hunter, 2003). The current investigation fills a gap in and contributes to the depression literature by examining DIF among the 21 BDI-II items. More specifically, we used item response theory (IRT) modeling and confirmatory factor analyses (CFA) to assess DIF among the BDI-II items.

Based on the gaps in the literature and the methodological benefits of IRT, we established two research questions to guide the current investigation: 1) To what extent does the BDI-II (Beck et al., 1996) provide equivalent scalar measurement for depressive symptoms in Black American and White American college students? And 2) to what extent does the BDI-II (Beck et al., 1996) provide equivalent scalar measurement for depressive symptoms in female and male college students?

Method

Participants and Procedure

Participants were a convenience cross-sectional sample of 1229 students from a large state university in the southeastern region of the United States. The sample included 145 Black American students and 1031 White American students. Gender samples were nearly equivalent; the study sample included 684 female participants and 545 male participants. Mean age in years for the sample was 21.15 (SD = 6.19). Year of school was almost evenly distributed among freshman, sophomore, junior, and senior levels (see **Table 1**). Participants reported low levels of depressive symptomatology. BDI-II mean scores based on self-reported race were 7.7 (SD = 7.5) and 8.7 (SD = 8.3) for Black American and White American students, respectively. BDI-II mean scores were 8.6 (SD = 8.2) and 8.5 (SD = 8.5) for females and males, respectively.

Following approval from our Institutional Review Board, we recruited participants in undergraduate-level classrooms and then later by email. Study invitations were sent to students through university email lists and individual class emails. We administered the electronic survey packet online using a web-based survey protocol. Before beginning the survey, participants viewed and electronically signed the study's informed consent form. The online survey included a demographic information survey and the BDI-II (Beck et al., 1996). The BDI-II and the demographic data sheet were presented in English. Extra course credit was provided both as an incentive and as compensation for time and effort involved in participating in the study.

Measures

Demographic Information. A researcher-designed demographic information sheet was created for the investigation. Questions inquired about the participant's year in school, academic discipline, religious affiliation, age, gender, and racial and ethnic background.

Beck Depression Inventory-II. We used the BDI-II (Beck et al., 1996) to assess each participant's level of depressive sympoms during the preceding 14 days. The BDI-II consists of 21 self-rated questions that assess for depressive symptomatology consistent with the criteria for major depressive disorder delineated in the *Diagnostic and Statistical Manual of Mental Disorders*-IV (4th ed., text rev.; *DSM-IV-TR*; American Psychiatric Association, 2000a). Participants are asked to select the option that best corresponds to the way they have been feeling during the past two weeks. Responses are self-rated on a four-point Likert-type scale: 0 (*absence of symptoms*) to 3 (*severe presence of symptoms*).

The BDI-II is scored by summing the participant's response for each of the 21 BDI-II items (Beck et al., 1996). Scores range from 0 to 63; higher scores reflect greater severity of

Table 1.
Demographics of study samples by gender and race for beck depression Inventory-I.

Demographic characteristic	Gender (n = 1229)		Race[b] (n = 1176)	
	Female (n = 684)	Male (n = 545)	Black American (n = 145)	White American (n = 1031)
	No. of students (%)/ mean (*SD*)	No. of students (%)/ mean (*SD*)	No. of students (%)/ mean (*SD*)	No. of students (%)/ mean (*SD*)
Age, years	20.9 (3.8)	20.6 (3.0)	22.6 (6.3)	20.5 (2.6)
Gender, female			63 (43%)	454 (44%)
Race				
Black American	82 (12%)	63 (12%)		
White American	571 (83%)	454 (83%)		
School year[a]				
Freshman	140 (21%)	81 (15%)	15 (10%)	188 (18%)
Sophomore	200 (30%)	195 (36%)	55 (38%)	330 (32%)
Junior	208 (31%)	162 (30%)	38 (26%)	315 (31%)
Senior	115 (17%)	89 (17%)	31 (22%)	170 (17%)
BDI-II mean score	8.6 (8.2)	8.5 (8.5)	7.7 (7.5)	8.7 (8.3)
BDI-II score (0 to 12)	526 (77%)	414 (76%)	116 (80%)	785 (76%)
BDI-II score (13 to 19)	84 (12%)	68 (12%)	12 (8%)	133 (13%)
BDI-II score (20 to 63)	74 (11%)	63 (12%)	17 (12%)	113 (11%)

Note: [a]Five participants failed to report year of school; [b]Fifty-three participants failed to report race.

depressive symptomatology and a greater probability of a clinical diagnosis of major depression. Beck and colleagues reported that scores of 16 or greater point to a probable diagnosis of depression. Beck and colleagues also suggested the following descriptions and interpretations related to severity: scores of 0 to 13 reflect minimal severity; 14 to 19 reflect mild severity; 20 to 28 reflect moderate severity; and scores of 29 or greater reflect severe symptomatology.

With regard to reliability, scores from the BDI-II have been shown to have sound internal stability. Studies using the BDI-II have reported alpha coefficients ranging from .77 to .92 (Carmody, 2005; Dozois, Dobson, & Ahnberg, 1998; Hirsch, Webb, & Jeglic, in press; Hooper & Doehler, 2011; Osman et al., 1997; Whisman et al., 2000). For comparison, the original validation study—composed in part of college student participants—reported a Cronbach's alpha value of .93 (Beck et al., 1996).

In terms of construct validity, although the BDI-II (or the earlier versions) cannot confirm a diagnosis of depression, the scores can point to probable depression (see Beck et al, 1996; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961). Findings from Osman and colleagues (1997) suggested that BDI-II scores yield sound convergent, construct, and discriminant validity. Research conducted by Dozois and colleagues (1998) and Storch and colleagues (2004) provided evidence for construct validity based on the relations between scores on the BDI-II, the State-Trait Anxiety Inventory-Depression (STAI-D; Spielberger,1983; Spielberger, Gorsuch, & Lushene, 1972), and the State-Trait Anxiety Inventory-Anxiety (STAI-A; Spielberger, 1983; Spielberger et al., 1972) factors scores. Dozois and colleagues provided guidance on recommended cutoff scores in college student populations: scores 0 to 12 indicate not depressed; scores 13 to 19 indicate dysphoria; and scores 20 to 63 indicate clinically depressed.

In the current investigation, Cronbach's alpha values resulting from the 21-item BDI-II reflected sound reliability of scale scores in all four samples. Consistent with stability coefficients in other studies (Beck et al., 1996; Carmody, 2005; Dozois et al., 1998; Hooper & Doehler, 2011; Whisman et al., 2000), Cronbach's alpha values were α = .90 for the Black American study sample and α = .92 for the White American study sample. The Cronbach's alpha value for females was α = .92; for males it was α = .92.

Missing Data

We examined missing data for all items on the BDI-II. All analyses in the current investigation included subjects with nonmissing values for all 21 items on the BDI-II. Therefore, only observed values were used; no imputation was performed. Of the 1229 participants, 53 participants failed to report their race and thus were excluded from the analyses. We used responses from all 1229 participants for gender-related analyses (see **Table 1**).

Data Analysis Plan

To examine research questions 1 and 2 we used item response theory (IRT) modeling to assess differential item functioning (DIF) among the BDI-II items. In addition to the recommendations put forward by numerous scholars (see Hambleton, 2006; Stark, Chernyshenko, & Drasgow, 2006; Hays, Morales, & Reise, 2000; Samejima, 1969), several factors influenced our rationale for the data analysis plan. For example, since the BDI-II is a polytomous instrument (items with more than two response options) in which items are measured on a four-point Likert-type scale, we considered a graded response model (GRM) as an appropriate method of IRT parameter estimation (Samejima, 1969). The GRM estimates for each item (i) a slope parameter (α_i) and a threshold parameter for each between-category threshold (β_{ij}). For the four-item Likert-type scale specifically, there are three between-category threshold parameters. The threshold parameters are the points along the latent trait continuum where respondents have a .50 probability of responding above a threshold. Using the estimated parameters, category response curves can be plotted to describe the probability that a respondent with a certain trait level (θ) will endorse or agree with a statement (i.e., item) at each point using the Likert-type scale.

DIF analysis in IRT is used to assess differences between

different groups of respondents with regard to the difficulty of item endorsement. Each item has an estimated difficulty location measured on the same scale as the trait level (θ). Items with positive location estimates are harder to endorse, and those with negative location estimates are easier to endorse. Items are considered to be displaying DIF if the item location estimates for two or more groups of respondents are significantly different when all other parameters are held constant. In addition to IRT, we used CFA to verify the unidimensionality of depressive symptoms in our study samples. LISREL 8.80 (Scientific Software International, 2007) was used for CFA analyses.

In sum, by employing a combined data analysis approach of IRT and CFA in the same study, we used a rigorous and conservative method (see Hays et al., 2000; Stark et al., 2006) to explore the extent to which scalar equivalence exists for Black American and White American college students and for female and male college students. However, we also recognize that there are alternative methods that have been recommended as well (see Brown, 2006). Consistent with recommendations put forward by Gregorich (2006) and others, we used confirmatory factor analysis to determine if the construct validity of the BDI-II is invariant for our two population groups: race (Black American vs. White American college students) and gender (female and male college students). We also used confirmatory factor analysis to determine whether the group differences that emerge are true differences in the construct under investigation (depressive symptomatology) or are instead effects related to some other factor pertaining to the demographics of the population groups (e.g., group-specific attributes such as gender).

Results

Results from our CFA indicated that a single dominant factor underlies the BDI-II items. The one-factor CFA demonstrated an adequate fit of the data for all comparison groups (see **Tables 2** and **3**). Goodness-of-fit indices for the four groups are as follows. Results for Black Americans were comparative fit index (CFI) = .92, normed fit index (NFI) = .86, and nonnormed fit index (NNFI) = .91; results for White Americans were CFI = .95, NFI = .95, and NNFI = .95. Goodness-of-fit results for the female college students were CFI = .95, NFI = .94, and NNFI = .95; and for male college students results were CFI = .95, NFI = .94, and NNFI = .94.

BDI-II: Descriptive Item Statistics

As illustrated in **Table 4**, the BDI-II depressive items were compared between Black American and White American students using independent sample *t* tests. Significant mean differences were evidenced on seven items (BDI-II items 2, 3, 5,

Table 2.
Analyses for unidimensionality and reliability for beck depression Inventory-II in Black American and White American college student participants.

	Black American	White American
Cronbach's Alpha Values	.90	.92
CFI	.92	.95
NFI	.86	.95
NNFI	.91	.95

Note: CFI = comparative fit index; NFI = normed fit index; NNFI = nonnormed fit index.

Table 3.
Analyses for unidimensionality and reliability for beck depression Inventory-II in female and male college student participants.

	Female	Male
Cronbach's Alpha Values	.92	.92
CFI	.95	.95
NFI	.94	.94
NNFI	.95	.94

Note: CFI = comparative fit index; NFI = normed fit index; NNFI = nonnormed fit index.

Table 4.
Beck depression Inventory-II item scores in Black American and White American college student participants.

BDI-II Item	Mean ± Standard Deviation		T (*p* Value)
Score Range: 0 - 3	White American (*n* = 1,031)	Black American (*n* = 145)	
BDI01—Sadness	.34 ± .57	.33 ± .58	–.15 (.884)
BDI02—Pessimism	.46 ± .60	.34 ± .62	–2.24 (.025)
BDI03—Failure	.42 ± .63	.27 ± .56	–2.66 (.008)
BDI04—Loss of Pleasure	.33 ± .59	.32 ± .60	–.03 (.973)
BDI05—Guilt	.43 ± .62	.31 ± .58	–2.24 (.025)
BDI06—Punishment	.26 ± .61	.23 ± .61	–.38 (.702)
BDI07—Self-Dislike	.40 ± .70	.26 ± .60	–2.41 (.016)
BDI08—Self-Criticalness	.55 ± .73	.30 ± .57	–4.05 (<.0001)
BDI09—Suicidal Thoughts	.11 ± .36	.07 ± .28	–1.15 (.250)
BDI10—Crying	.37 ± .65	.40 ± .79	.50 (.619)
BDI11—Agitation	.42 ± .63	.33 ± .59	–1.56 (.120)
BDI12—Loss of Interest	.32 ± .59	.30 ± .50	–.23 (.820)
BDI13—Indecisiveness	.45 ± .78	.37 ± .63	–1.20 (.232)
BDI14—Worthlessness	.23 ± .57	.12 ± .40	–2.19 (.029)
BDI15—Loss of Energy	.52 ± .61	.56 ± .64	.75 (.456)
BDI16—Change in Sleep	.85 ± .76	.82 ± .81	–.37 (.712)
BDI17—Irritability	.35 ± .61	.35 ± .57	.01 (.991)
BDI18—Change in Appetite	.58 ± .77	.56 ± .77	–.34 (.733)
BDI19—Concentration Difficulty	.55 ± .76	.54 ± .78	–.24 (.814)
BDI20—Tiredness/Fatigue	.54 ± .61	.55 ± .69	.17 (.863)
BDI21—Loss of Interest in Sex	.20 ± .52	.37 ± .72	3.61 (.0003)

Note: Boldfaced values reflect a significant difference.

Table 5.
Beck depression Inventory-II scores in female and male college student participants.

BDI-II Item Score Range: 0 - 3	Mean ± Standard Deviation		T (*p* Value)
	Female (*n* = 684)	Male (*n* = 545)	
BDI01—Sadness	.34 ± .58	.33 ± .53	.30 (.761)
BDI02—Pessimism	.44 ± .60	.46 ± .62	–.61 (.542)
BDI03—Failure	.39 ± .62	.42 ± .65	–.76 (.445)
BDI04—Loss of Pleasure	.33 ± .57	.33 ± .62	.24 (.811)
BDI05—Guilt	.41 ± .61	.43 ± .64	–.58 (.563)
BDI06—Punishment	.25 ± .57	.26 ± .63	–.30 (.764)
BDI07—Self-Dislike	.38 ± .69	.40± .68	–.41 (.682)
BDI08—Self-Criticalness	.52 ± .71	.52 ± .73	.18 (.858)
BDI09—Suicidal Thoughts	.11 ± .36	.09 ± .33	.82 (.412)
BDI10—Crying	.41 ± .66	.32 ± .34	2.32 (.020)
BDI11—Agitation	.42 ± .63	.40 ± .63	.62 (.535)
BDI12—Loss of Interest	.30 ± .56	.35 ± .63	–1.47 (.141)
BDI13—Indecisiveness	.44 ± .76	.43 ± .71	.46 (.648)
BDI14—Worthlessness	.22 ± .58	.21 ± .54	.05 (.958)
BDI15—Loss of Energy	.52 ± .61	.53 ± .64	–.05 (.958)
BDI16—Change in Sleep	.83 ± .78	.84 ± .75	–.21 (.834)
BDI17—Irritability	.37 ± .60	.32 ± .61	1.48 (.140)
BDI18—Change in Appetite	.59 ± .76	.58 ± .79	.25 (.802)
BDI19—Concentration Difficulty	.54 ± .75	.56 ± .76	–.51 (.613)
BDI20—Tiredness/Fatigue	.55 ± .60	.55 ± .66	.01 (.992)
BDI21—Loss of Interest in Sex	.22 ± .54	.21 ± .57	.28 (.776)

Note: Boldfaced values reflect a significant difference.

7, 8, 14, and 21). White American students had higher mean scores on all items with the exception of item 21. BDI-II depressive items also were compared between female and male college students using *t* tests. Significant mean differences were evidenced on one item only, item 10. In this case, as shown in **Table 5**, females had higher mean scores for this item than male respondents.

BDI-II: Differential Item Functioning Analyses

To compare responses across the study samples, individual items on the BDI-II were assessed for DIF using the MULTILOG 7.03 program (Thissen, 1991). MULTILOG uses Marginal Maximum Likelihood (MML) estimation to evaluate the significance of item location differences between groups. All parameters except the location parameter were held constant for Black American and White American respondents. A chi-square statistic for the contrast between Black American and White American item locations was used to test for significance of the contrast. Items with chi-square values above the critical value at a .05 alpha level with one degree of freedom are considered to be displaying DIF. We followed the same analytic procedures for female and male respondents.

As illustrated in **Table 6**, significant differences were found in the item-level responses to the BDI-II based on race in the current study. More specifically, five items on the BDI-II (items 7, 8, 14, 15, and 21) displayed DIF in relation to Black American and White American respondents. Of the five items, only two items (BDI-II item 8, self-criticalness, and item 21, loss of interest in sex) exhibited DIF based on both methods: CFA and IRT analyses.

Table 7 shows that differences were also found in the item-level responses to the BDI-II based on gender in the current study. More specifically, two items on the BDI-II items (item 10, crying; and item 12, loss of interest) displayed DIF in relation to female and male student-respondents. Importantly, as can also be seen in **Table 7**, the two DIF items that emerged in our sample were observed from the CFA but not the IRT analyses.

Discussion

This study used a convenience cross-sectional sample of 1229 American student-respondents to examine race- and gender-related measurement equivalence (or bias) in the performance of the BDI-II (Beck et al., 1996). More specifically, using IRT and CFA, we tested the extent to which the BDI-II provides equivalent scalar measurement for depressive symptoms in Black American and White American college students and in female and male college students. The data from our convenience cross-sectional sample of American students point toward four main findings. We discuss these results in terms of our proposed research questions.

Our first main finding relates to our racial group comparisons. We used DIF analyses to examine research question 1: To what extent does the BDI-II (Beck et al., 1996) provide equivalent scalar measurement for depressive symptoms in Black American and White American college students? The data produced differences in symptom endorsement based on race. Twenty-three percent of the items on the BDI-II functioned differently based on at least one comparison method (i.e., CFA or IRT). More specifically, for these race-related comparisons, symptom endorsement varied on five BDI-II items: items 7, self-dislike; 8, self-criticalness; 14, worthlessness; 15, loss of energy; and 21, loss of interest in sex. Therefore, five items functioned differently, and 16 of the items functioned similarly in these racial group comparisons.

These results align with empirical findings as well as theoriz-

Table 6.
Differential Item Functioning (DIF) results from Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) methods for Black American and White American student comparisons.

Model	Chi-Square (Difference)	
	CFA ($\Delta df = 2$)	IRT ($\Delta df = 4$)
Baseline Model (Referent: Item BDI01—Sadness)	2319.3	15845.7
Comparison Models		
BDI02—Pessimism	4.6	11.0
BDI03—Failure	5.1	8.3
BDI04—Loss of Pleasure	4.6	7.7
BDI05—Guilt	4.9	9.9
BDI06—Punishment	2.8	5.6
BDI07—Self-Dislike	6.2	17.8^{DIF}
BDI08—Self-Criticalness	22.2^{DIF}	21.0^{DIF}
BDI09—Suicidal Thoughts	2.1	2.0
BDI10—Crying	10.7	8.1
BDI11—Agitation	1.5	5.4
BDI12—Loss of Interest	6.2	9.5
BDI13—Indecisiveness	.2	9.1
BDI14—Worthlessness	22.2^{DIF}	5.6
BDI15—Loss of Energy	13.3^{DIF}	10.0
BDI16—Change in Sleep	4.6	6.9
BDI17—Irritability	4.4	8.3
BDI18—Change in Appetite	.3	1.7
BDI19—Concentration Difficulty	2.2	1.9
BDI20—Tiredness/Fatigue	7.1	10.9
BDI21—Loss of Interest in Sex	22.6^{DIF}	23.6^{DIF}
Total number of DIF Items	4	3

Note: In CFA, DIF is flagged if chi-square (x^2) was > 11.98. In IRT, DIF is flagged if chi-square (x^2) was >16.51. Boldfaced values reflect DIF flagged for both CFA and IRT.

Table 7.
Differential Item Functioning (DIF) results from Confirmatory Factor Analysis (CFA) and Item Response Theory (IRT) methods for female and male student comparisons.

Model	Chi-Square (Difference)	
	CFA ($\Delta df = 2$)	IRT ($\Delta df = 4$)
Baseline Model (Referent: Item BDI01—Sadness)	2194.2	17167.1
Comparison Model		
BDI02—Pessimism	.7	3.9
BDI03—Failure	1.0	1.6
BDI04—Loss of Pleasure	1.3	10.0
BDI05—Guilt	1.3	6.7
BDI06—Punishment	.8	7.2
BDI07—Self-Dislike	.9	4.4
BDI08—Self-Criticalness	.2	1.3
BDI09—Suicidal Thoughts	1.9	3.0
BDI10—Crying	14.4^{DIF}	10.2
BDI11—Agitation	.6	2.1
BDI12—Loss of Interest	16.2^{DIF}	7.7
BDI13—Indecisiveness	.8	8.5
BDI14—Worthlessness	9.1	8.9
BDI15—Loss of Energy	1.3	1.3
BDI16—Change in Sleep	.1	7.1
BDI17—Irritability	3.9	7.9
BDI18—Change in Appetite	.8	2.3
BDI19—Concentration Difficulty	.5	2.4
BDI20—Tiredness/Fatigue	8.9	8.0
BDI21—Loss of Interest in Sex	2.3	4.1
Total number of DIF Items	2	0

Note: In CFA, DIF is flagged if chi-square (x^2) was > 11.98. In IRT, DIF is flagged if chi-square (x^2) was > 6.51. Boldfaced values reflect DIF flagged for both CFA and IRT.

ing in the literature related to the differential presentation and endorsement of depressive symptoms (i.e., scale scores and item scores) in adult population based on varied racial groups (see Teresi et al., 2008). Of significance, only a few studies have explored differences in depressive symptoms at the item level in particular using the BDI-II. More often, the comparisons have been done at the total score level. We can point to several studies comparing scale scores of the BDI-II based on race. Walker and Bishop (2005) found in their study, which was composed of college students, that White American students reported higher scores (BDI-II mean score = 9.1) on the BDI-II than Black American students did (BDI-II mean score = 8.3). Similarly, in the present investigation, our data indicated that White American students reported higher total scores (BDI-II mean score = 8.7) than did Black American college students (BDI-II mean score = 7.7). However, in another study composed of older adolescents (Miller & Taylor, 2011), depressive symptoms as measured by the CES-D (Radloff, 1977) revealed that older Black American adolescents had higher levels of depressive symptoms than did their White American counterparts. It remains unclear if differences evinced in the literature are true differences, differences based on the measure used, differences at the item level, or differences in sample or some other unmeasured factor (Boughton & Street, 2007; Santor et al., 1995).

Furthermore, those comparisons that have focused on DIF have been based primarily on age or gender comparisons (e.g., Carmody, 2005; Kim, Pilkonis, Frank, Thase, & Reynolds, 2002; Teresi et al., 2008), not racial comparisons. In one study that did include racial comparisons, Carmody (2005) investigated item bias in the endorsement of symptoms on the BDI-II among White, Asian, and Latino Americans. In his study, White American students scored higher on three items (BDI-II items 11, agitation; 14, worthlessness; and 17, irritability) than did Hispanic and Asian American students. However, Carmody's study made no comparisons between White American and Black American students.

In the analytical approach that we employed (i.e., IRT and CFA jointly) for research question 1, we found some discrepancies in the identification of DIF. More specifically, we found inconsistencies based on our CFA and IRT analyses. As seen in **Table 6**, the CFA identified four DIF items, whereas the IRT identified three DIF items. No clear guidelines exist when there are discrepancies in DIF results when multiple data analytic methods are employed (i.e., CFT and IRT), such as in our study (see Borsboom, 2006; Hambleton, 2006).

Our second main finding relates to our gender group comparisons. We used DIF analyses to examine research question 2: To what extent does the BDI-II (Beck et al., 1996) provide equivalent scalar measurement for depressive symptoms in female and male college students? The results from our data revealed slight differences in symptom endorsement between gender comparison groups. More specifically, for these comparisons, symptom expression varied on two BDI-II items: items 10, crying; and 12, loss of interest. Nineteen of the items on the BDI-II were found to function similarly for females and males. Consequently, for most items (90%) on the BDI-II, no bias was uncovered, and scalar equivalence was observed. Our findings are in partial agreement with Carmody's (2005) findings, although more differences emerged in his group comparisons. In Carmody's study, item-level scores differed based on gender; females had higher scores on BDI-II items 1, 10, 15, and 20.

Our results for between-group differences are consistent with other studies of college student populations (Gladstone& Koenig, 1994; Eisenberg et al., 2007; Nolen-Hoeksema, 1990). For example, as reviewed by Boughton and Street (2007), several studies have found gender differences to be invariant in populations specifically of college students. However, the dominant view and theorizing holds that gender differences in depression and depressive symptoms do exist (Boughton & Street; Kessler et al., 2003). Moreover, the amassed empirical literature (cross-sectional and epidemiological studies) on gender differences related to depression and depressive symptoms has suggested that symptom profiles—in most populations—are different more often than not (Carmody, 2005; Kessler et al., 2003; Leino & Kisch, 2005; World Health Organization, 2002). In the end, it is clear that much more research needs to be done to disentangle the effects that gender has on depressive symptoms in college student populations specifically.

As evidenced in our analysis related to race comparisons, we found various levels of DIF based on our CFA and IRT analyses. These emergent differences are important and require additional consideration in future investigations. Moreover, whether the items that showed DIF in our study need to be replaced remains unclear. Additional studies need to be conducted to test if the patterns evinced in the current study can be replicated.

In our third main finding, we demonstrated a method that can be employed in examining item and scalar equivalence in cross-cultural (e.g., race and gender) comparisons. Establishing item and scalar equivalence—even for commonly used instruments—is an important step that is often overlooked when comparisons are made (Harach et al., 2006). Many researchers and scholars have assumed that because instruments are used widely (e.g., CES-D [Radloff, 1977], BDI-II, Brief Symptom Inventory [BSI; Derogatis, 1993], and Health Outcomes Measure [SF-36; Ware & Sherbourne, 1992]) scalar equivalence is established. Additionally, some researchers conclude if total scale or subscale scores show no differences between diverse comparison groups then scalar equivalence is established (Teresi et al., 2008; van de Vijver & Tanzer, 2004). Drawing conclusions from these scale score and subscale score group comparisons could lead to faulty assumptions or erroneous conclusions (van de Vijver & Tanzer, 2004). Therefore, investigations at the item level are paramount; the importance cannot be overstated. Findings from our investigation add to the clinical and research literature base on the psychometric properties of the BDI-II and afford researchers and scholars alike confidence when screening for depressive symptoms in college student-respondents.

Finally, we found that the reliability of the BDI-II total scale score across all four groups was more than adequate. Cronbach's alpha values in the target groups in the current investigation demonstrated high internal reliability and ranged from .90 to .92. These values are also consistent with Beck et al.'s validation study (1996) and Dozois and colleagues' (1998) findings. These results—in conjunction with our other findings—also support item equivalence of the BDI-II in our samples, although alpha values by themselves should not be the sole method to establish item equivalence (Hui & Triandis, 1985; Vandenberg & Lance, 2000).

Study Limitations and Directions for Future Research

This study contributes to the literature by examining the extent to which one of the most commonly used measures to assess for depressive symptoms—the BDI-II (Beck et al., 1996)—produced scalar differences in Black American and White American college students and in female and male college students. In other words, did the participants' responses indicate that the BDI-II items function equivalently in the four samples?

Concurrent with our results, limitations of the study must be considered. First, the sample size of the two racial groups was limited. The unequal sample sizes of the two groups could have attenuated the results of the study. A second limitation, also related to the study sample, is that the participants were from one university and thus may not be representative of all college students. A third limitation is that our study was composed of a nonclinical population of college students. Although high levels of depressive symptoms and a clinical diagnosis of depression are often seen in college student populations, the majority of the sample reported low levels of depressive symptoms (see Beck et al., 1996; Dozois et al., 1998). Future studies should attempt to replicate these findings with clinical samples with clinical levels of depressive symptomatology.

We did not assess for social desirability—a fourth limitation. Some scholars have concluded that social desirability could explain DIF associated with racial and cultural groups (van de Vijver & Tanzer, 2004). Thus, future studies should consider the inclusion of a measure that assesses for social desirability.

A fifth limitation arises because the current study focused on depression and depressive symptoms; however, there are many other mental health disorders and problems with which college and university populations are faced.

A sixth limitation is that we compared two racial groups only. It remains unclear whether these findings are representative of findings that would be evinced in different racial and cultural groups, age groups, or clinical groups—or even in groups from different geographical regions. Future studies should include comparisons with additional racial and ethnic groups, including Latino individuals, one of the largest ethnic minority groups in the United States (see Humes, Jones, & Ramirez, 2011; Marin, Escobar, & Vega, 2006). We employed several statistical tests in the current study. As a result, some of our findings may be based on chance. Thus, the number of tests run in the current study serves as a seventh limitation.

Finally, it is plausible that our results were attenuated by the homogeneity of our college student sample. For example, the impact of the college experience, including the daily living experiences on a college campus, could have created more similarities than differences in our sample irrespective of race and gender. In other words, the strength of the common college experience could have been more powerful than the strength of the cultural experiences evidenced in our college student sample (Carmody, 2005; Kadison, 2004). Future research is needed to determine the applicability and generalizability of the results in the current investigation.

These findings have implications for future research. Studies that explore cross-cultural research, minority health and health disparities, and racial and cultural differences in medical conditions and mental health symptomatology must include measures that reliably capture the construct under investigation (Manly, 2006). To be clear, researchers and clinicians alike must be confident that the often-reported differences in depression and depressive symptoms are true differences and not artificial differences that are attributable to biased items on the measure used (i.e., DIF). Therefore, more studies that examine scalar and cultural equivalence of measures are needed (Harachi et al., 2006; McHorney & Fleischman, 2006; van de Vijver & Tanzer, 2004). The criticality of this need—even among the most commonly used instruments—cannot be overstated and therefore has far-reaching effects on science, practice, and policy (Eisenberg et al., 2007). Finally, and of significance, most of the most widely used instruments (e.g., BDI-II) were developed with predominantly or exclusively White American samples, in monocultural contexts, and most often with college student populations (see Beck et al., 1996; Boughton & Street, 2007; Carmody, 2005; Steer & Clark, 1997).

College students in general are an at-risk, high-priority population when it comes to the development of depressive symptoms and depression (Gore & Aseltine, 2003; National Research Council & Institute of Medicine, 2009). Cultural factors such as race and gender may further complicate or exacerbate mental health conditions, including depression. Future studies must ensure that racially diverse individuals are included in research studies that examine diagnosis and treatment methods and measures for medical conditions and mental health disorders such as depression (see Manly, 2006).

In addition, although our results found gender differences to be invariant future studies may want to consider a two-pronged culturally tailored approach to assess for gender differences in depression. For example, several researchers have intimated that the current DSM-IV criteria (American Psychiatric Association, 2000a) for depression (and thus indirectly the BDI-II) may fail to capture fully men's symptoms of depression. Several scholars have suggested that a singular assessment tool may be inadequate. Cochran and Rabinowitz (2003) suggested a culturally sensitive approach by asking male clients typical questions about depressive symptoms *and* questions that reflect masculine-specific distress and symptoms. For example, questions about increased anger and agitation, decreased motivation, increased somatic concerns, and a decrease in sexual *interest*, with little change in sexual *behavior*. To address the possible limitations of current assessments, Magovcevic and Addis (2008) developed the Masculine Depression Scale. In their development and refinement study, they found that males reported both typical depressive symptoms evinced in the current DSM-IV criteria (American Psychiatric Association, 2000a) as well as masculine-specific depressive symptoms (e.g., getting mad and feeling less confident). Other scholars have also suggested that typical depressive symptoms are often filtered through a masculine-focused lens and thus males may not be reporting symptoms that are evidenced on assessment tools (e.g., BDI-II and CES-D) and/or filtering their depressive symptoms through a masculine gender framework causing depression to go undiagnosed, undetected, and untreated. Future studies should consider a culturally tailoring approach—using multiple instruments—when examining depression in males (see Fields & Cochran, 2011).

Conclusion

The current investigation fills a gap in the depression literature. Our investigation is the first to assess scalar equivalence of the BDI-II (Beck et al., 1996) in college students using two rigorous methods jointly (see Hays et al., 2000; Stark et al., 2006). More specifically, this was the first study to examine DIF with the BDI-II using IRT and CFA concurrently. In our

study sample we found DIF based on race and gender comparisons. Despite these differences, overall our results, suggest the BDI-II scores appear to be a reliable and valid measure of depressive symptoms for Black and White American college students and female and male college students. This research advances clinical knowledge about the utility, reliability, and validity of the BDI-II scores in a racially and culturally diverse college student population. Using DIF we found that 16 of the 21 items on the BDI-II functioned similarly in our racial group comparisons, and 19 of the 21 items on the BDI-II functioned similarly in our gender group comparisons. Before this investigation, most cross-cultural comparison studies focused on total BDI-II scale score analyses and therefore may have missed important differences at the item level. We can conclude with some confidence that many of the items on the BDI-II functioned equivalently in our sample. Based on our preliminary results, it appears that the BDI-II items do not need to be significantly revised based on gender groups in college populations. However, researchers may want to consider to what extent the BDI-II items need to be slightly tailored based on gender and racial groups or used in conjunction with other measures (e.g., Masculine Depression Scale; Magovcevic & Addis, 2008) in college populations, although more studies replicating these findings are warranted before any changes are implemented.

REFERENCES

American College Health Association (2009). *National college health assessment II*. Reference Group Executive Summary, Fall 2008. Baltimore, MD: American College Health Association.

American Psychiatric Association (2000a). *Diagnostic and statistical manual of mental health disorders* (4th ed., text revision). Washington, DC: American College Health Association.

American Psychiatric Association (2000b). *Practice guideline for the treatment of patients with major depressive disorder*. Washington, DC: American College Health Association.

Anderson, E. R., & Mayes, L. C. (2010). Race/ethnicity and internalizeing disorders in youth: A review. *Clinical Psychology Review, 30,* 338-348. doi:10.1016/j.cpr.2009.12.008

Arria, A. M., O'Grady, K. E., Caldeira, K. M., Vincent, K. B., Wilcox, H. C., & Wish, E. D. (2009). Suicide ideation among college students: A multivariate analysis. *Archives of Suicide Research, 13,* 230-246. doi:10.1080/13811110903044351

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *BDI-II manual*. San Antonio, TX: The Psychological Corporation.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 12,* 57-62.

Blanco, C., Okuda, M., Wright, C., Hasin, D. S., Grant, B. F., Liu, S. M., & Olfson, M. (2008). Mental health of college students and their non-college-attending peers: Results from the National Epidemiologic Study on Alcohol and Related Conditions. *Archives of General Psychiatry, 65,* 1429-1437. doi:10.1001/archpsyc.65.12.1429

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44,* s176-s181. doi:10.1097/01.mlr.0000245143.08679.cc

Boughton, S., & Street, H. (2007). Integrated review of the social and psychological gender differences in depression. *Australian Psychologist, 42,* 187-197. doi:10.1080/00050060601139770

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Carmody, D. P. (2005). Psychometric properties of the Beck Depression Inventory-II with college students of diverse ethnicity. *International Journal of Psychiatry of Clinical Practice, 9,* 22-28. doi:10.1080/13651500510014800

Centers for Disease Control and Prevention (2010). National Center for Inquiry Prevention and Control, Web-Based Injury Statistics Query and Reporting System. http://www.cdc.gov/injury/wisqars/index.html

Chao, R. K., & Otsuki-Clutter, M. (2011). Racial and ethnic differences: Sociocultural and contextual explanations. *Journal of Research on Adolescence, 21,* 47-60. doi:10.1111/j.1532-7795.2010.00714.x

Cochran, S. V., & Rabinowitz, F. E. (2003). Gender-sensitive recommendations for assessment and treatment of depression in men. *Professional Psychology: Research and Practice, 34,* 132-140. doi:10.1037/0735-7028.34.2.132

Coyne, J. C., & Marcus, S. C. (2006). Health disparities in care for depression possibly obscured by the clinical significance criterion. *American Journal of Psychiatry, 163,* 1577-1579. doi:10.1176/appi.ajp.163.9.1577

Culbertson, F. M. (1997). Depression and gender. An international review. *American Psychologist, 52,* 25-31. doi:10.1037/0003-066X.52.1.25

Day, J. (1996). *Population projections of the United States by age, sex, race, and Hispanic origin: 1995-2050*. Washington, DC: US Government Printing Office.

Derogatis, L. R. (1993). *Brief Symptom Inventory: Administration, scoring, and procedures manual*. Minneapolis, MN: National Computer Systems, Inc.

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

Dunlop, D. D., Song, J., Lyons, J. S., Manheim, L. M., & Chang, R. W. (2003). Racial/ethnic differences in rates of depression among preretirement adults. *American Journal of Public Health, 93,* 1945-1952. doi:10.2105/AJPH.93.11.1945

Eaton, N. R., Keyes, K. M., Krueger, R. F., Balsis, S., Skodol, A. E., Markon, K. E., & Hasin, D. S. (2011). An invariant dimensional liability model of gender differences in mental disorder prevalence: Evidence from a national sample. *Journal of Abnormal Psychology*. doi:10.1037/a0024780

Eisenberg, D., Gollust, S. E., Golberstein, E., & Hefner, J. L. (2007). Prevalence and correlates of depression, anxiety, and suicidality among university students. *American Journal of Orthopsychiatry, 77,* 534-542. doi:10.1037/0002-9432.77.4.534

Fields, A. J., & Cochran, S. V. (2011). Men and depression: Current perspectives for health care professionals. *American Journal of Lifestyle Medicine, 5,* 92-100.

Furr, S. R., Westefeld, J. S., McConnell, G. N., & Jenkins, J. M. (2001). Suicide and depression among college students: A decade later. *Professional Psychology, Research and Practice, 32,* 97-100. doi:10.1037/0735-7028.32.1.97

George, L. K., & Lynch, S. M. (2003). Race differences in depressive symptoms: A dynamic perspective on stress exposure and vulnerability. *Journal of Health and Social Behavior, 44,* 353-369. doi:10.2307/1519784

Gladstone, T. R., & Koenig, L. (1994). Sex differences in depression across the high school to college transition. *Journal of Youth and Adolescence, 23,* 643-669. doi:10.1007/BF01537634

Gonzalez, H. M., Vega, W. A., Williams, D. R., Tarraf, W., West, B. T., & Neighbors, W. (2010). Depression care in the United States. *Archives of General Psychiatry, 67,* 37-46. doi:10.1001/archgenpsychiatry.2009.168

Gore, S., & Aseltine, R. H. (2003). Race and ethnic differences in depressed mood following the transition from high school. *Journal of Health and Social Behavior, 44,* 370-389. doi:10.2307/1519785

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement variance using the confirmatory factor analysis framework. *Medical Care, 44,* s78-s94. doi:10.1097/01.mlr.0000245454.12228.8f

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44,* s182-s188. doi:10.1097/01.mlr.0000245443.86671.c4

Hankin, B. L. (2002). Gender differences in depression from childhood through adulthood: A review of course, causes, and treatment. *Primary Psychiatry, 9,* 32-36.

Hankin, B. L., & Abramson, L. Y. (2001). Development of gender differences in depression: An elaborated cognitive vulnerability-transactional stress theory. *Psychological Bulletin, 127,* 773-796. doi:10.1037/0033-2909.127.6.773

Harachi, T. W., Choi, Y., Abbott, R. D., Catalano, R. F., & Bliesner, S. L. (2006). Examining equivalence of concepts and measures in diverse samples. *Prevention Science, 7,* 359-368. doi:10.1007/s11121-006-0039-0

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38,* s1128-s1142. doi:10.1097/00005650-200009002-00007

Hirsch, J. K., Webb, J. R., & Jeglic, E. L. (in press). Forgiveness, depression, and suicidal behavior among a diverse sample of college students. *Journal of Clinical Psychology*.

Hooper, L. M. (2010). The unmet needs of depressed adolescent patients: How race, gender, and age relate to evidence-based depression care in rural areas. *Primary Health Care Research & Development, 11,* 339-348. doi:10.1017/S1463423610000277

Hooper, L. M., & Doehler, K. (2011). The mediating effects of differentiation of self on body mass index and depressive symptomatology among an American college sample. *Counselling Psychology Quarterly, 24,* 71-82. doi:10.1080/09515070.2011.559957

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology—A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16,* 131-152. doi:10.1177/0022002185016002001

Humes, K. R., Jones, N. A., & Ramirez, R. R. (2011). Overview of race and Hispanic origin: 2010. http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Institute of Medicine (2002). *Unequal treatment: Confronting racial and ethnic disparities in health care.* Washington DC: National Academies Press.

Iwata, N., & Buka, S. (2002). Race/ethnicity and depressive symptoms: A cross-cultural/ethnic comparison among university students in East Asia, North and South America. *Social Science and Medicine, 55,* 2243-2252. doi:10.1016/S0277-9536(02)00003-5

Kadison, R. (2004). The mental-health crisis: What colleges must do. *The Chronicle of Higher Education*, B20.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H. U., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Results from the National Comorbidity Survey. *Archives of General Psychiatry, 51,* 8-19. doi:10.1001/archpsyc.1994.03950010008002

Kessler, R. R., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., & Wang, P. S. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *Journal of the American Medical Association, 289,* 3095-3105. doi:10.1001/jama.289.23.3095

Kilmartin, C. (2005). Depression in men: Communication, diagnosis and therapy. *The Journal of Men's Health & Gender, 2,* 95-99. doi:10.1016/j.jmhg.2004.10.010

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential item functioning of the Beck Depression Inventory in late-life patients: Use of item response theory. *Psychology and Aging, 17,* 379-391. doi:10.1037/0882-7974.17.3.379

Kisch, J., Leino, V. E., & Silverman, M. M. (2005). Aspects of suicidal behavior, depression, and treatment in college students: Results from the spring 2000 national college health assessment survey. *Suicide and Life-Threatening Behavior, 35,* 3-13. doi:10.1521/suli.35.1.3.59263

Leino, E. V., & Kisch, J. (2005). Correlates and predictors of depression in college students: Results from the spring 2000 national college health assessment. *American Journal of Health Education, 36,* 66-74.

Magovcevic, M., & Addis, M. E. (2008). The masculine depression scale: Development and psychometric evaluation. *Psychology of Men and Masculinity, 9,* 114-132. doi:10.1037/1524-9220.9.3.117

Manly, J. J. (2006). Deconstructing race and ethnicity: Implications for measurement of health outcomes. *Medical Care, 44,* s10-s16. doi:10.1097/01.mlr.0000245427.22788.be

Marin, H., Escobar, J. I., & Vega, W. A. (2006). Mental illness in Hispanics: A review of the literature. *Focus, 4,* 23-37.

McHorney, C. A., & Fleischman, J. A. (2006). Assessing and understanding measurement equivalence in health outcomes measures. Issues for further quantitative and qualitative inquiry. *Medical Care, 44,* s205-s210. doi:10.1097/01.mlr.0000245451.67862.57

National Institute of Mental Health (2010). Office for Research on Disparities and Global Mental Health: Overview. http://www.nimh.nih.gov/about/organization/od/office-for-research-on-disparities-and-global-mental-health-ordgmh.shtml

National Institutes of Health. (2002). Outreach notebook for the inclusion, recruitment and retention of women and minority subjects in clinical research. URL. http://orwh.od.nih.gov/pubs/outreach.pdf

National Research Council & Institute of Medicine (2009). *Preventing mental, emotional, and behavioral disorders among young people: Progress and possibilities*. Washington, DC: National Academies Press.

Nolen-Hoeksema, S. (1990). *Sex differences in depression*. Stanford, CA: Stanford University Press.

Nolen-Hoeksema, S., & Girgus, J. S. (1994). The emergence of gender differences in depression during adolescence. *Psychological Bulletin, 115,* 424-443. doi:10.1037/0033-2909.115.3.424

Osman, A., Downs, W. R., Barrios, F. X., Kopper, B. A., Guiterrez, 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

Paniagua, F. A. (1994). *Assessing and treating culturally diverse clients: A practical guide.* Thousand Oaks, CA: Sage.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385-401. doi:10.1177/014662167700100306

Rao, U., & Chen, L. (2009). Characteristics, correlates, and outcomes of childhood and adolescent depressive disorders. *Dialogues in Clinical Neuroscience, 11,* 45-62.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika,* Monograph Supplement No. 17.

Santor, D. A., Zuroff, D. C., Cervantes, Palacios, J., & Ramsay, J. O. (1995). Examining scale discrimability in the BDI and CES-D as a function of depression severity. *Psychological Assessment, 7,* 131-139. doi:10.1037/1040-3590.7.2.131

Schmidt, F. L., & Hunter, J. E. (2003). History, development, evolution, and impact of validity generalization and meta-analysis methods. In K. R. Murphy (Ed.), *Validity and generalization: A critical review* (pp. 31-65). Mahwah, NJ: Erlbaum.

Silverstein, B. (1999). Gender differences in the prevalence of clinical depression: The role played by depression associated with somatic symptoms. *American Journal of Psychiatry, 156,* 480-482.

Sperry, L. (2010). Culture, personality, health, and family dynamics: Cultural competence in the selection of culturally sensitive treatments. *Family Journal, 18,* 316-320. doi:10.1177/1066480710372129

Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory STAI.* Palo Alto, CA: Mind Garden.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1972). *Manual for the State-Trait Anxiety Inventory Self-Evaluation Questionnaire.* Palo Alto, CA: Consulting Psychologists Press.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91,* 1292-1306. doi:10.1037/0021-9010.91.6.1292

Steer, R. A., & Clark, D. A. (1997). Psychometric properties of the Beck Depression Inventory-II with college students. *Measurement and Evaluation in Counseling and Development, 30,* 128-137.

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

Teresi, J. A., Ramirez, M., Lai, J. S., & Silver, S. (2008). Occurrences and sources of differential item functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life, and general health. *Psychological*

Science Quality, 50, 538-600.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Chicago: Scientific Software.

Tjia, J., Givens, J. L., & Shea, J. A. (2005). Factors associated with undertreatment of medical student depression. *Journal of American College Health, 53,* 219-224. doi:10.3200/JACH.53.5.219-224

US Department of Health and Human Services (n.d.) Healthy People 2020: Proposed Objectives. http://www.healthypeople.gov/hp2020/default.asp

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-69. doi:10.1177/109442810031002

van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Review for Applied Psychology, 54,* 119-135. doi:10.1016/j.erap.2003.12.004

Walker, R. L., & Bishop, S. (2005). Examining a model of the relation between religiosity and suicidal ideation in a sample of African American and white college students. *Suicide and Life-Threatening Behavior, 35,* 630-639. doi:10.1521/suli.2005.35.6.630

Ware, J. E., & Sherbourne, C. D. (1992). The MOS 36 Item Short Form Health Survey (SF 36): Conceptual framework and item selection. *Medical Care, 30,* 473-483. doi:10.1097/00005650-199206000-00002

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

Wilcox, H. C., Arria, A. M., Caldeira, K. M., Vincent, K. B., Pinchevsky, G. M., & O'Grady, K. E. (2010). Prevalence and predictors of persistent suicide ideation, plans, and attempts during college. *Journal of Affective Disorders, 127,* 287-294. doi:10.1016/j.jad.2010.04.017

World Health Organization (2002). *The World Health Report 2002: Reducing risks, promoting healthy life*. Geneva: World Health Organization.