HHS Public Access

Author manuscript

Clin Psychol Sci. Author manuscript; available in PMC 2019 July 09.

Published in final edited form as:

Clin Psychol Sci. 2016 March ; 4(2): 239–253. doi:10.1177/2167702615591768.

Racial/ethnic differences in youth depression indicators: An item response theory analysis of symptoms reported by White, Black, Asian, and Latino youths

Rachel A. Vaughn-Coaxum, **Patrick Mair**, and **John R. Weisz**

Department of Psychology, Harvard University, Cambridge, MA

Abstract

Accurate assessment of dysfunction is central to clinical psychological science, essential for valid conclusions about prevalence, risk, and appropriate intervention. Measures applied without adjustment across diverse racial/ethnic groups may risk errors if measurement equivalence has not been established. We tested this possibility in the domain of youth depression, applying item response theory (IRT) and differential item functioning (DIF) analyses to reports by White, Black, Latino, and Asian youths (N = 2,335) on the most widely-used measure of symptoms, the Children's Depression Inventory (CDI). Analyses revealed that 77% of CDI items were non-equivalent indicators of symptom severity across groups. CDI sum scores exhibited marked over-estimations of group differences and inappropriate classification as "clinically-elevated" for 29% of Latino, 23% of Black, and 10% of Asian youths. Applying DIF adjustment corrected these errors. The study demonstrates a useful strategy for ethnically sensitive assessment, applicable to other symptom domains and ethnic groups.

Keywords

child; adolescent; psychometrics; depression; measurement invariance

Depression is highly impairing and carries significant social, emotional and functional consequences. Indeed, the World Health Organization (2004) predicts that depression will be the world's most burdensome psychiatric disorder in the 21st century. This underscores the need for the most accurate assessment possible, to facilitate precise tracking of the extent and scope of the problem, and to inform efforts to prevent and treat depression. This is especially important in the school-age years, as rates of depression surge during the transition from childhood to adolescence (Costello, Mustillo, Erkanli, Keeler, & Angold,

Correspondence concerning this article should be addressed to Rachel Vaughn-Coaxum, Department of Psychology, Harvard University, Cambridge, MA 02138., rachelvaughn@fas.harvard.edu.

Author Contributions

R. A. Vaughn-Coaxum developed the study concept. All authors contributed to the study design. R. A. Vaughn-Coaxum was responsible for data analysis and interpretation under the supervision of P. Mair and J. Weisz, and drafting the manuscript. P. Mair was responsible for data analysis and critical revision of the manuscript. J. R. Weisz was responsible for original data collection and critical revision of the manuscript. All authors approved the final version of the manuscript.

Declaration of Conflicting Interests

The authors report no financial relationships or interests that may be affected by material in the manuscript, or which might potentially bias it.

2003; Weisz, McCarty, & Valeri, 2006); indeed, nearly 14% of American youths are diagnosed at some point during adolescence (Merikangas et al., 2010).

In the U.S., the need for accurate assessment of depressive symptoms in the childhood to adolescence transition readily focuses attention on race and ethnicity, given national demographic trends. Some 46% of Americans under age 18 are racial/ethnic minority youths (Mather, Pollard, & Jacbobsen, 2011), and this percentage is increasing steadily (U.S. Census Bureau, 2013). Research on rates of depression in different U.S. racial/ethnic groups has produced widely varying results, raising the question of whether standard measures using standard scoring actually produce equivalent assessment of symptoms across different racial/ethnic groups (Allen & Astuto, 2012; Crockett, Randall, Shen, Russell, & Driscoll, 2005; Merikangas & Knight, 2012). For example, while some have suggested that out-group stress and a historical “stigma of inferiority” increase susceptibility of Black youths to depressive symptoms (Wight et al., 2005), assessment studies disagree as to whether depressive symptomology is more pronounced in Black than White youths (Cole, Martin, Peeke, Henderson, & Harwell, 1998; Kistner, David, & White, 2003; Roberts, Roberts, & Chen, 1997), more pronounced in White youths (Saluja et al., 2004), or not significantly different between-groups (e.g., Paxton, Valois, Watkins, Huebner, & Drane, 2007; Twenge & Nolen-Hoeksema, 2002).

Somewhat more consistent findings indicate a particularly high level of depression symptomology among Latino youths (Allen & Astuto, 2012; Doi, Roberts, Takeuchi, & Suzuki, 2001; Paxton et al., 2007; Roberts et al., 1997; Saluja et al., 2004; Twenge & Nolen-Hoeksema, 2002; U.S. Department of Health and Human Services, 2001), and increased risk for suicidality (Pena, Matthieu, Zayas, Masyn, & Caine, 2012; Pena, Zayas, Cabrera-Nguyen, & Vega, 2012). Some suggest that immigration and acculturation pressures, as well as fatalism in the culture, leave Latino youths especially prone to depression (Allen & Astuto, 2012; Paxton et al., 2007; Roberts et al., 1997). However, assessment findings have differed on whether Latino youths’ symptom levels are higher (e.g., Allen & Astuto, 2012; Paxton et al., 2007) or not (Kubik, Lytle, Birnbaum, Murray, & Perry, 2003). For Asian youths, some evidence suggests greater risk compared to White or other minority youths (Kubik et al., 2003), while other research indicates no difference from other ethnic groups (Saluja et al. 2004; Wight et al., 2005). Some have suggested that these youths’ increased risk for depression symptomology is related to reduced help-seeking behavior and stressors associated with immigration (e.g., language barriers, isolation).

Such divergent findings across studies pose a challenge for those who seek to understand risk and prevalence in different racial/ethnic groups. The divergent findings may result from a number of factors, including cross-study differences in samples, inclusion criteria, and other study procedures, but the conflicting results have raised another concern: assessment of depression symptoms in different racial/ethnic groups, to date, has been carried out by scoring standard measures without any adjustment for group characteristics. This approach implicitly assumes that standard measures scored in the same standard way for different groups will generate symptom scores that have invariant meaning in all the groups. That assumption may warrant attention, and testing. As a case example, consider the Children’s Depression Inventory (CDI; Kovacs, 1992; 2004), the most widely used self-report measure

of depression symptoms in children and adolescents (Twenge & Nolen-Hoeksema, 2002). To date, the CDI has been used in more than 1000 studies (Google Scholar; PsycINFO; accessed September 7, 2014); to our knowledge, in none of those studies that included multiple racial/ethnic groups was there any adjustment in scoring based on group membership. This would not be a problem if standard item scoring generates scores with the same meaning for each racial/ethnic group; but if that were not the case, then standard scoring with no adjustment could produce an invalid picture of depression symptomology in one or more racial/ethnic groups. Thus, it is useful to know whether widely-used measures such as the CDI are invariantly measuring depression symptoms across different racial/ethnic groups.

Assessing measurement invariance is essential for understanding group differences. Measurement invariance is the state of affairs in which the function relating latent variables to observations is the same across all groups being compared (Borsboom, 2006). Borsboom (2006) indicates that tests for measurement invariance are necessary when evaluating group-differences in mean scores because the presence of bias may confound the very scores that are producing observed differences. Indeed, Kovacs (2004) encouraged investigation of cross-ethnic differences in CDI scores. Kovacs did not report evidence of measurement differences across racial/ethnic groups in the factor analytic structure of the CDI in the normative sample, but noted that further research was needed to inform conclusions about score interpretation cross-ethnically due to group differences that have been found in prior studies. A few very helpful studies have investigated cross-ethnic measurement invariance of the CDI, but in rather limited ways and without the use of the refined item response theory (IRT) methods that are now more readily available.

Cole and colleagues (1998) found evidence for invariant factor structure on the CDI for Black and White children using multi-group confirmatory factor analysis (MG-CFA). Steele and colleagues (2006) found invariant factor structure on the CDI for Black and White youths, but found that two of five factors on the CDI were distinct dimensions for Black but not White youths. Politano and colleagues (1986) found different factor structures and factor weights across Black and White adolescents. One study comparing White and Latino youths found evidence of measurement equivalence on the CDI by comparing regression parameters and intercorrelations between symptoms and other functional outcomes across groups (Knight, Viridin, Ocampo, & Rossa, 1994). A recent examination by Huang and Dong (2013) confirmed that few studies have appropriately compared factor structure across racial/ethnic groups, and that we still lack evidence that the CDI is cross-ethnically equivalent for youths. Moreover, while a few studies have compared White youths to one other racial/ethnic group, no studies, to our knowledge, have compared different minority groups to one another. So, the evidence to date is limited both methodologically and demographically.

On the methodological front, psychometric advances call for a more fine-grained approach to testing cross-ethnic measurement invariance than has been used with the CDI to date. IRT has been identified as a particularly appropriate psychometric method for testing measurement equivalence (de Ayala, 2009). IRT overcomes many limitations of classical test theory analyses—the traditional approach that includes the previously noted factor analytic

methods. Under classical test theory, observed scores on a measure are test-based and group-dependent; therefore true scores (i.e., latent variable scores) are heavily influenced by the characteristics of the sample. As a result, scores will change as the properties of the measure change (Hambleton & Jones, 1993). Under IRT, individuals' latent trait scores are estimated from the statistical properties of the scale items, which are independent of the groups they were estimated from. Resulting trait scores are thus less dependent on sample characteristics (Hambleton & Jones, 1993), and because item characteristics are also sample-independent, IRT estimates have superior generalizability. Further, IRT accounts for the ordinal level of data, which makes it a good fit to the CDI's ordinal scale approach.

In the present study, we used a sample of 2,335 early adolescents to evaluate whether the CDI exhibits measurement invariance across the four largest racial/ethnic groups in the U.S.: White, Black, Asian, and Latino. The study is the first, to our knowledge, to include more than two racial/ethnic groups, and the first to use current IRT methods to examine invariance on the CDI cross-ethnically. We tested the dimensional structure of the CDI to determine whether it is consistent cross-ethnically, and we tested whether the individual items functioned invariantly across groups. Overall, we sought to evaluate whether biases impact between-group mean differences in depression levels as well as categorical classification of youths into the clinically-elevated symptom range, and thus whether adjustments in scoring will be needed to generate an unbiased picture of depression symptomology in the different racial/ethnic groups.

Method

Participants and Procedure

Participants were 2,335 6th and 7th grade students, 53% female and with mean age of 11.74 years. The 6th and 7th grades (53% and 47% of the sample, respectively) provided a focus on early adolescence, the period when rates of depression begin to rise sharply (McLaughlin, Hilt, & Nolen-Hoeksema, 2007). Some 41% identified their race/ethnicity as White, 20% Black, 7% Asian, and 32% as Latino [Other racial/ethnic groups, including mixed race/ethnicity, were excluded due to small Ns.] Some 13% identified as first generation immigrants, 27% second generation, and 13% third generation. Previous work (e.g., Gil & Vega, 1996) suggests the need to document participant nationality. Among youths identifying as Latino, 46% identified their nationality as Mexican, 1.4% identified as Cuban, 18.9% identified as Puerto Rican, 25.8% identified as Central or South American, 15.3% identified as Dominican, and 6.9% identified as other Latin American background. Among youths identifying as Asian, 42.9% identified their nationality as Chinese, 13.6% identified as Japanese, 16.9% identified as Korean, 10.7% identified as Filipino, and 27.1% identified as other Asian background. Across all racial/ethnic groups, 53.2% of youths reported that they speak only English at home, 13.8% reported that they spoke mostly English and sometimes another language at home, while 14.5% of youths reported that they speak both English and another language equally at home. Some 11.9% of youths reported that they speak mostly another language at home and sometimes English, while 6.6% of youths reported speaking only a language other than English at home. Based on the 85th percentile cut-off for clinically-elevated scores on the CDI (see Kovacs, 2004), 393 participants (22%

White, 25% Black, 8% Asian, and 44% Latino) had clinically elevated depression symptoms.

Participants were recruited from 10 middle schools in California and Massachusetts, with the CDI administered orally at each school, in classroom-sized groups, in English. Informed parental consent and student assent were obtained, following IRB approval from the University of California at Los Angeles and the Judge Baker Children's Center, Harvard Medical School. Parental consent forms were provided to families in English and Spanish, depending on parental preference and language fluency/comprehension.

Children's Depression Inventory (CDI; Kovacs, 1992; 2004). The CDI, the most widely used self-report measure of depression, is supported by extensive reliability and validity data (e.g., Kovacs, 1992). Items are written as ordered, categorical sentences with three response categories (e.g., 0 = *Nobody really loves me*, 1 = *I am not sure if anybody loves me*, or 2 = *I am sure that somebody loves me*). Across studies, Cronbach's alpha has ranged from .80 to .94 (Saylor, Finch, Spirito, & Bennett, 1984) and test-retest reliability from .38 to .87 (Saylor et al., 1984). The single item (of 27 total) asking about suicidal ideation was removed due to the school officials' concerns about suggesting suicide to youths who might not otherwise have thought of it. Prior research has not shown the 26-item version to differ from the original 27-item version in terms of the network of social and psychological constructs the CDI is associated with (Twenge & Nolen-Hoeksema, 2002). Cronbach's alpha for the scale in the present study was .88.

Data Analyses

All statistical analyses were performed in the R environment for statistical computing (R Core Team, 2015).

Tests of unidimensionality.—To perform IRT analyses on the full CDI, there must be evidence that all of the items on the scale are measuring the same underlying construct. Therefore, we first determined the dimensional structure of the CDI using ordinal exploratory factor analysis (EFA) based on polychoric correlations as implemented in the psych package (Revelle, 2015) in R. Scale dimensions were initially determined for all four ethnic groups combined, then for each ethnic group individually, with the best-fitting dimensional structure across the full sample and the four groups used in subsequent analyses. These present methods are patterned after Hambrick et al. (2010) and van Beek, Hessen, Hutteman, Verhulp, and Leuven (2012).

Estimated IRT model and parameters.—Next, we used an IRT based approach to detect DIF across ethnic groups using the lordif package in R (Choi, Gibbons, & Crane, 2011). Internally this package uses the Graded Response Model (GRM; Samejima, 1969) in order to scale the items. Unidimensional CDI scale(s) were fit to GRM, which allows for variation in two types of IRT parameters estimated by the model. One of these is the item discrimination parameter. Discrimination refers to the strength of association between each item and the latent construct of depression symptom severity, establishing each item's capability to distinguish between respondents located at various points along the symptom severity continuum. The discrimination parameter (α) for each item can be understood in

classical test theory terms as the correlation between a particular item and the observed score (i.e. the total or sum score) on the measure (de Ayala, 2009). Item discrimination is analogous to a factor loading under classical test theory factor analytic methods.

The second set of parameters are the item category location parameters. Item location, referred to as "item difficulty" in proficiency or aptitude tests, models the distribution of items across the continuum of depression symptom severity, identifying where each item best captures symptom severity. Item categories with location parameters at the lower end of the continuum best capture depression severity in more normative—as opposed to clinically severe—symptom ranges. Items with category location parameters at the high end of the latent continuum best capture symptom severity in more severe or elevated symptom ranges. Both item discrimination and category location parameters are critical to scale evaluation. Traditional sum scores treat all items as equivalent, yet differences in these parameters influence estimates of individuals' true scores. Discrimination and category location parameters were estimated for each of the 26 CDI items in the present study. Further, for each participant, latent trait levels of symptom severity were estimated as a theta score (Θ) that was treated as the total score on the CDI.

Investigation of measurement invariance.—To investigate measurement invariance, the category locations and discrimination parameters for each item were compared across racial/ethnic groups. Statistically significant differences across groups indicate Differential Item Functioning (DIF). DIF (Agnoff, 1993) exists when an item displays different statistical properties across groups, after differences in the trait levels of the groups are accounted for. In the present study, DIF for a CDI item would indicate that the item differentially captures depression symptom severity across racial/ethnic groups—i.e., that latent symptom severity alone does not account for participants' individual responses to the CDI—and the absence of DIF would indicate that the item demonstrates measurement invariance (de Ayala, 2009).

We began the process of linking item and respondent characteristics across sub-groups using the lordif package in R. Items that were invariant across all four groups were identified and then used as anchors to re-calibrate all the CDI item responses to the same metric across groups so that the racial/ethnic group comparisons could be made appropriately. Next, a hybrid IRT/logistic regression approach (using the proportional-odds logistic regression method) to investigating DIF was performed (see Choi et al., 2011). This approach allowed for the detection of both uniform and non-uniform DIF. DIF is considered uniform when the effect is constant along the latent trait level of depression symptomology, and DIF is considered non-uniform when the effect varies conditionally along the latent trait level. A set of three ordinal regression models were fit for each item (as detailed below). In each model, the dependent variable was the probability of endorsing each of the three response categories for that item (P category 0, 1, 2), while the independent variables included trait levels of depression symptomology, race/ethnicity, and the interaction between these two terms. Each predictor was added incrementally in a new regression model, resulting in hierarchical nesting of the three models.

The models were compared by means of Likelihood Ratio (LR) χ^2 statistics. Also, for fit magnitude, McFadden's Pseudo R^2 (a proxy R^2 value used in logistic regression to estimate the gain in log-likelihood from the model's explanatory variables; Veall & Zimmermann, 1996) was calculated. Subsequently, for model comparison we examined differences between the pseudo R^2 values. In addition, we computed the absolute proportional change in point estimates for β_1 (Model 1 vs. Model 2). Model 1 included an intercept plus the latent trait level of depression symptom severity ($P_{\text{category 0, 1, 2}}$ = intercept + β_1* trait level), and this model is nested within Model 2. Model 2 included an intercept, the latent trait level of symptom severity, and racial/ethnic group membership ($P_{\text{category 0, 1, 2}}$ = intercept + β_1* trait level + β_2* group), and this model is nested within Model 3. Model 3 included an intercept, the latent trait level of depression symptom severity, ethnic group membership, and the interaction term between latent trait level and group membership ($P_{\text{category 0, 1, 2}}$ = intercept + β_1* trait level + β_2* group + β_2* trait level * group). The interaction term in Model 3 represents the specific test for DIF, evaluating whether trait levels of depression symptom severity vary by ethnic group membership across each item. The significance of this model is tested against Model 1 and Model 2. The comparisons of Model 1 to Model 2 and Model 1 to Model 3 test for uniform DIF, while the comparison of Model 2 to Model 3 tests for non-uniform DIF.

After all items were tested for DIF, empirically-derived cut-offs for the value of each test statistic (χ^2, β, and pseudo R^2) were used to evaluate whether DIF was meaningful. The cut-offs were derived from multiple Monte-Carlo simulated datasets, preserving observed group differences in trait level, under the null hypothesis that all 26 items were invariant. This method of simulation, described in detail by Choi and colleagues (2011), repeatedly computes various levels of magnitude across the simulated datasets from which the empirical distributions are derived—using the correlation structure of the data and reducing the probability of falsely rejecting the null hypothesis that all items are invariant. Test statistics that exceeded the Monte-Carlo cut-offs suggested clinically meaningful, as oppose to spurious, DIF. This approach is particularly appropriate for evaluating the magnitude of pseudo R^2 values. For most pseudo R^2 values an interpretation in terms of magnitude is problematic (see e.g., Mittlboeck & Schemper, 1996). Since there are no goodness-of-fit cut-offs, Monte-Carlo simulations provide an empirical way to evaluate these measures. Next, for items flagged for DIF, racial/ethnic group-specific item discrimination and category location parameters were estimated. Using these group-specific estimates, a new DIF-adjusted trait (Θ) estimate was produced for every participant, accounting for cross-ethnic measurement bias so that all latent trait estimates of symptom severity had comparable meaning and could be evaluated on the same scale.

Comparison of CDI scoring methods.—To test for impact of measurement bias, IRT latent trait estimates of depression symptom severity were compared to traditional total CDI scores. Total scores for each participant were calculated by summing item responses (0, 1, or 2) across all 26 items. An ANOVA was performed to investigate mean differences across racial/ethnic groups between total scores and DIF-adjusted Θ scores. Cohen's *d* effect sizes were calculated using mean differences and standard deviations for each pairwise group comparison to determine whether racial/ethnic group differences (or lack therefore) were

equivalent under classical test theory and IRT estimations. Next, we identified youths who met the CDI cutoff for clinically significant depression (i.e., >85th percentile of normative scores; see Kovacs, 2004). For all youths meeting this criterion, we calculated whether they also met the 85th percentile criterion for clinically-elevated, DIF-adjusted, trait estimates of symptom severity. Mismatched cases, i.e., those for which total scores were not classified into the same groups across scoring methods, reflected biased estimates of youth depression symptoms.

Results

The results of the ordinal EFA based on polychoric correlations supported unidimensionality of the CDI for the full sample, as well as for each racial/ethnic group individually. Eigenvalues for the full sample are given in Figure 1. The scree plot shows the eigenvalues of the full solution and, in addition, includes the eigenvalues of a random data matrix of the same size as the original one, computed from a parallel analysis. This analysis computed random data matrices of the same size and shape using re-sampled and normal data, allowing us to compare the results of our EFA to the average eigenvalues that would be produced by chance, strengthening the evidence that our results for dimensionality differ from chance. The one-factor solution produced the best fit, based on two goodness-of-fit test criteria that have been shown to best estimate the number of interpretable factors (Henson & Roberts, 2006; Zwick & Velicer, 1986): the Velicer Minimum Average Partial test (MAP; Velicer, 1976) and the Very Simple Structure criterion (VSS; Revelle & Rocklin, 1979). The MAP represents the squared, average partial correlation among items after removing the effect of the factors. The factor structure that minimizes the average partial correlations represents best fit. The VSS test degrades the factor solution to test how well the factor matrix fits the correlation matrix, with the maximum value achieved representing the ideal number of factors to extract.

The single factor solution for the full sample produced a MAP value of 0.01 and a maximum VSS value of 0.74. The results of the ordinal factor analysis for each racial/ethnic subgroup produced a similar picture with all eigenvalues, MAP and VSS values in the appropriate range for a unidimensional solution. Based on these measures and in conjunction with the “elbow criterion” (Thorndike, 1953; visual detection of the point where change is greatest and at which adding another factor results in minimal gain in variance accounted for) in the scree plot we concluded that the CDI was unidimensional. Accordingly, all subsequent IRT and DIF analyses were conducted on the full CDI as a single scale.

Graded Response Model, DIF detection, and test information across racial/ethnic groups

In order to verify the feasibility of the GRM as the base model in our subsequent DIF analysis, we tested the fit of the GRM for each of the four subgroups separately using the R package mirt (Chalmers, 2012). The fit indices—including RMSEA, RMSR, and comparative fit index (CFI) values—supported good fit (see Browne & Cudeck, 1993; Hu & Bentler, 1999) for White youths (RMSEA = .041 [95% CI: .038 – .045]; RMSR = .052; CFI = .97), Black youths (RMSEA = .043 [95% CI: .037 – .049]; RMSR = .059; CFI = .96), Asian youths (RMSEA = .025 [95% CI: .00 – .040]; RMSR = .067; CFI = .99), and Latino

youths (RMSEA = .043 [95% CI: .039 – .048]; RMSR = .053; CFI = .96). Based on these results we concluded that the GRM fit the data for each of the four groups, and we proceeded with a model for the full sample followed by DIF analyses. The IRT parameters from the GRM of the full sample indicated that discrimination parameters for all 26 items ranged from $\alpha = 0.91$ to $\alpha = 2.56$, falling within the range of "good" discrimination (0.80 to 2.50; de Ayala, 2009) and demonstrating that the construct under study is not too narrow in this sample. Although each item on the CDI initially included three response categories (0, 1, and 2), the number of response categories was collapsed from three to two on 13 of the 26 items as a result of low endorsement (i.e., < five observations) of the most extreme response (category 2; see Choi et al., 2011). On the full-scale level, Figure 2 illustrates the range of information on symptom severity that the CDI is able to capture for each racial/ethnic group across the full range of scores.

Prior to performing DIF analyses, the proportional odds assumption was tested to ensure that the DIF approach could be applied appropriately. We performed a graphical proportional odds inspection for the probability of endorsing each response category (Harrell, 2001) across each of the CDI items based on latent trait scores of depression symptom severity and racial/ethnic group membership. Results confirm that the proportional odds assumption was met. The DIF analyses revealed that only six of the 26 CDI items were invariant across racial/ethnic groups; these were items for sadness, crying spells, indecisiveness, sleep disturbance, loneliness, and lack of friends. The 20 remaining items displayed DIF across racial/ethnic groups. For each item displaying DIF, empirical cut-offs for test statistics based on the Monte-Carlo simulated datasets revealed that all of the χ^2, pseudo-R^2, and β values from the comparison of logistic regression models exceeded the thresholds for meaningful differences in item properties. Therefore, evidence suggested that the DIF for all 20 items was meaningful and warranted further investigation. The item parameters for all 20 items were re-calculated to produce ethnic group-specific item discrimination and category location estimates that account for the differential functioning of the items (see Table 1). Category location estimates are item location parameters referring to the DIF-adjusted trait level of symptom severity where respondents crossed the threshold from one response category to the next (i.e., from category 0 to 1, and from 1 to 2). Category locations ranged from 0.13 – 2.52 at the first threshold and from 1.87 – 5.17 at the second threshold. Results indicated a considerable range of trait levels captured, which provides added support for the suitability of the present IRT methods for this clinical construct of depression symptomology (Reise & Waller, 2009).

Magnitude and direction of DIF for non-equivalent items

For the items displaying DIF with test statistics exceeding Monte-Carlo thresholds for clinical significance, Differential Step Functioning analyses (Penfield, Gattamorta, & Childs, 2009) were conducted to determine which ethnic groups and item parameters differed from one another. Item category location parameters were compared across ethnic groups and group differences were evaluated using empirical cutoffs for small, medium, and large effect sizes. For each ethnic-group specific response category threshold on every item (crossing from 0 to 1 = b1, and from 1 to 2 = b2), difference scores were computed pairwise between ethnic groups. Between-group differences in each category location (Δb1 and Δb2)

constitute a small effect if less than 0.25, a medium effect if 0.26 to 0.50, and a large effect if greater than 0.50. This post-hoc testing is warranted, as the magnitude of DIF could be masked at the group level if there are differing patterns of group discrepancies for each response category. In effect, group differences in item parameters may vary with symptom severity and Differential Step Functioning provides an appropriate evaluation of whether this is the case.

Results of these analyses revealed that group differences varied widely across the symptom clusters defined by Kovacs (2004) as *interpersonal problems*, *ineffectiveness*, *negative self-esteem*, *negative mood*, and *anhedonia*, suggesting that group differences were not specific to symptom type. However, item-level group differences are consistent with prior research on racial/ethnic group differences in culturally-normative expressions of distress related depression, which is important for conceptualizing possible sources of DIF (McHorney & Fleishman, 2006). For example, an examination of the items with large effect sizes for group differences in item parameters reveals that Item 11 “Irritability” indicates significantly higher symptom severity for Asian, Black, and Latino youths compared to White youths. This pattern of results is consistent with existing literature indicating that irritability may be a more culturally normative expression of distress than sadness in ethnic minority youths (see reviews by Anderson & Mayes, 2010; Choi, 2002). Item 25 “Feeling Unloved” indicates the greatest level of severity for White youths compared to all ethnic minority groups, and this item indicates the lowest level of severity for Asian youths. This finding is consistent with literature suggesting that expressions of love are more implicit than explicit in many Asian cultures (Choi, 2002). Therefore, this symptom may not be a strong indicator of depression among these youths, for whom less frequent explicit statements of affection may be a relatively normative experience.

The Test Information Curve in Figure 2 provides information that is consistent with the general patterns of DIF for the item discrimination parameter across racial/ethnic groups. The discrimination parameter values for White and Asian youths tend to trend together, while discrimination values for Black and Latino youths tend to trend together. For the 20 items displaying DIF, discrimination was higher for Asian and White youths than Black and Latino youths on 18 items, and Asian youths represented the group with the highest item-discrimination on 10 of the items. Additionally, the category location thresholds reflect a similar pattern, with White and Asian youths tending to have similar locations and Black and Latino youths tending to lie at similar locations for each threshold (see Table 1). These results suggest that the strength of the association between symptoms and the underlying latent trait of depression is strongest for Asian and White youths. A possible explanation for the high level of test information among Asian youths is that their responses may be more consistent across the measure, resulting in higher inter-item correlations and the greatest amount of information available on the latent trait. Additionally, results suggest that the trait levels captured by the category location thresholds are most similar for White and Asian youths compared to Black and Latino youths. Both sets of findings reflect results from the Test Information Curve indicating that the CDI provides more information for White and Asian youths than Black and Latino youths.

Racial/ethnic group differences for total symptom severity

A one-way ANOVA revealed significant mean-differences in unadjusted CDI, raw, total scores across ethnic groups, $F(3, 2323) = 43.77$, $p = .00$. Unadjusted total scores were calculated by summing the raw responses (0, 1, or 2) from each of the 26 items. Total scores ranged from 0 to 44 in the current sample. White youths ($M = 5.13$, $SE = 0.81$) had significantly lower depression severity than any other group. Using Cohen's d criteria, the effect sizes for these group differences ranged from small ($d = 0.27$, White vs. Asian youths) to medium ($d = 0.49$, White vs. Black youths; $d = 0.50$, White vs. Latino youths). Asian youths ($M = 6.82$, $SE = 0.55$) had significantly lower depression severity than Black ($M = 8.25$, $SE = 0.32$) or Latino ($M = 8.31$, $SE = 0.25$) youths, although the effects were small ($d = 0.23$ and $d = 0.24$, respectively). Latino youths had the highest levels of depression severity followed by Black youths, with no significant difference between the two.

With DIF-adjusted trait estimates as the dependent variable, the overall main effect of ethnic group remained significant, but the patterns of group differences changed in some respects, $F(3, 2323) = 4.16$, $p = .01$. DIF-adjusted trait estimates were calculated by summing the recalibrated, IRT responses from each of the 26 items. DIF-adjusted scores ranged from −1.61 to 3.39 in the current sample. White youths still had significantly lower depression severity ($M = -0.07$, $SE = 0.03$) than Black ($M = 0.08$, $SE = 0.04$) and Latino ($M = 0.06$, $SE = 0.03$) youths, but the effect sizes were below threshold for even a small effect ($d = 0.17$ and $d = 0.14$, respectively). Further, White youths did not differ from Asian youths ($M = 0.04$, $SE = 0.07$) and Asian youths also did not differ from Black and Latino youths. The relative position of Black and Latino youths changed, with Blacks showing non-significantly higher depression symptom levels than Latinos.

Overall, racial/ethnic differences in total symptom severity were markedly smaller in magnitude after adjusting for DIF, with none of the statistically significant mean differences meeting the criterion for even a small effect size, and with the relative positions of Black and Latino youths reversed, relative to their position with unadjusted scores.

Racial/ethnic group differences for clinically-elevated depression symptoms

Following Kovacs (2004), we used the 85th percentile as a cutoff for clinically elevated depression symptoms. An analysis of the distribution for both types of scores resulted in the identification of threshold values to identify the 85th percentile of participants. In the present sample, this was a raw, total score of 13. The corresponding 85th percentile cut-off for the DIF-adjusted trait estimates was a Θ value of 0.99. Frequency distributions of only youths with elevated total scores revealed the following classification patterns: all of the White youths with clinically-elevated total scores also had DIF-adjusted trait scores in the clinically-elevated range ($\Theta > 0.99$). For Black youths with clinically-elevated total scores, only 77% had clinically-elevated DIF-adjusted trait scores. For Asian youths with clinically-elevated total scores, only 90% had clinically-elevated DIF-adjusted trait estimates. For Latino youths with clinically-elevated total scores, only 71% had clinically-elevated DIF-adjusted trait scores. In other words, relying on traditional sum scores and not accounting for DIF across items led to over-estimation of clinically-elevated symptom severity (more false positives) in all three ethnic minority groups. Thus, due to measurement bias 10–29% of

racial/ethnic minority youths were misclassified as having clinically-elevated depression symptoms (see Figure 3).

Associations of other demographic factors with DIF-adjusted trait estimates

Finally, we conducted post-hoc analyses probing whether other demographic factors were associated with DIF-adjusted trait estimates. Previous research suggests that differing experiences related to racial/ethnic background may be associated with immigration status, gender, and socioeconomic status (SES). Independent samples t-tests compared mean DIF-adjusted trait estimates by gender and immigration history (i.e., whether youths were first, second, or third-generation Americans). A one-way ANOVA was performed to compare DIF-adjusted trait estimates by language spoken in the home. Parental education and occupation were requested in study assessments, to provide SES data, but could not be included because of missing data; many youths did not know their parents' highest level of education or specific job.

Results indicated that students across all racial/ethnic groups who were first generation immigrants had significantly higher symptom severity than students born in the United States ($t = -3.82$, $p = .00$); the overall effect was small ($d = -0.24$). There were no differences in severity by immigration history within any of the four racial/ethnic groups individually. It is important to note, however, that there was significant heterogeneity in the nationalities that were identified by Asian and Latino youths and analysis of immigration history from differing countries was not included due to the small sample size of many sub-groups. DIF-adjusted symptom severity was significantly greater for youths who spoke only a language other than English with their parents compared to those speaking only English (mean difference = 0.26, $p = .01$). Youths speaking mostly another language with their parents also endorsed higher symptom severity than English speaking only youths (mean difference = 0.24, $p = .00$). Youths who spoke mostly English at home or equal amounts of English and another language did not differ from any other group. There were also no differences in symptom severity associated with language use when examining effects within the individual racial/ethnic groups. There were no gender differences with respect to mean DIF-adjusted trait estimates of depression severity across the full sample ($t = -0.69$, $p = .49$), or within individual racial/ethnic groups.

A linear regression was performed to test the association of DIF-adjusted trait estimates with student age. Student age was positively associated with depression symptom severity, such that older youths had higher scores on the CDI ($t = 4.56$, $p = .00$). However, age only accounted for 1% of the variance in DIF-adjusted trait estimates (adjusted $R^2 = .01$). In sum, demographic variables that are often associated with race/ethnicity had only very modest effects in relation to DIF-adjusted depression trait estimates in the present study.

Discussion

Psychological scientists have long questioned whether the experience and expression of youth depression could actually be alike across racially/ethnically diverse groups in America, given marked group differences in culture, social status, and experience in the United States. Studies have tackled this question, but limitations in measurement and data

analytic methods, and in sample composition, have left it unclear whether the (rather mixed) findings on depression severity in different racial/ethnic groups are meaningful and interpretable. We sought to shed new light on the question by examining whether raw scores on the CDI—the most widely used self-report measure of youth depression symptoms—were equivalent indicators of symptom severity for youths across the four largest racial/ethnic groups in the U.S.: White, Black, Asian and Latino. The findings indicate that (a) equivalent raw scores on the CDI do not indicate equivalent latent levels of depression symptom severity for youths in the four racial/ethnic groups, and (b) the strength of individual symptoms in characterizing expression of depression symptomology differs across the four groups.

Our ordinal EFA showed that there was a common latent dimension of depression across racial/ethnic groups (see also Cole et al., 1998; Crockett et al., 2005; Latzman et al., 2011; Steele et al., 2006; Trent et al., 2013); but our subsequent IRT findings showed that under comparable dimensional structure measurement variance may still be detected. The potential for detection of variance within similar factor structure supports the value of analytic applications beyond traditional factor analytic methods of testing invariance. Indeed, our findings showed that CDI items differed across racial/ethnic groups in (a) the level of depression severity indicated by the individual symptom items, and (b) the strength of association between each item and the latent trait of depression. When differential item functioning was properly adjusted for, what had previously appeared to be rather striking racial/ethnic differences in level of depression symptom severity shrank markedly, with no group difference meeting criteria for even a small effect. In addition, when these adjustments were made, it became clear that standard scoring of the CDI would lead to misclassification of many minority youths as “clinically elevated:” 10% of Asian youths, 23% of Black youths, and 29% of Latino youths. By contrast, no White youths were misclassified using standard CDI scoring. So, despite group similarity in dimensional structure of the CDI, differential item functioning across groups led to an exaggerated picture of group differences in overall symptom severity levels, and to inaccurate classification of many youths as falling into the clinical range of severity. Significantly, the distorting effect of racial/ethnic group differences in item functioning was seen only in the three minority groups.

These findings suggest that an uncritical application of standard scoring procedures for such self-report clinical measures as the CDI could lead to faulty conclusions about psychopathology, particularly in racial/ethnic minority groups, with possible consequences for both research and clinical practice. One research consequence, for example, could be inaccurate estimates of incidence and prevalence in epidemiologic research, resulting in inappropriate conclusions about relative risk in various population groups. On the clinical front, one consequence of a failure to correct for differential item functioning across groups could be an inflation in false-positive rates in clinical assessment, resulting in inappropriate identification of youths as candidates for targeted prevention, or treatment referral.

Such problems are not inevitable consequences of differential item functioning across racial/ethnic groups, on the CDI or other clinical measures. Although IRT is sometimes used in an effort to develop or revise measures to be invariant or “fair” for all respondents, DIF detection through IRT can also be used to adjust scoring procedures for measures whose

contents will not be altered. Algorithms can be programmed (e.g., within statistical software packages like R) that convert individuals' observed responses to DIF-adjusted trait estimates, thus generating scores that can be appropriately compared across different population groups. Our findings highlight the potential impact of pairing IRT analyses to identify DIF across groups with algorithm development to create more accurate and interpretable DIF-adjusted scores—certainly for the CDI, and also potentially for an array of other clinical measures.

In addition to these implications, our findings may suggest a useful strategy for identifying those symptom clusters that are more, and less, likely to differ as a function of race and ethnicity. We found, for example, that items displaying cross-group differences tended to fall into clusters the CDI manual identifies as *interpersonal problems* (e.g., misbehavior, reduced social interest), *ineffectiveness* (e.g., difficulty with school work, self-deprecation), and *negative self-esteem* (e.g., feeling unloved, increased pessimism). By contrast, items in the *negative mood* (e.g., sadness, irritability, indecisiveness) and *anhedonia* (e.g., sleep disturbances, loneliness) clusters were less likely to show DIF. This pattern might be seen as consistent with arguments in the literature that racial/ethnic group differences in depression symptomology reflect group differences in such constructs as "fatalism" resulting in greater pessimism (see Choi, 2002), perceptions of efficacy (Jenkin, Kleinman, & Good, 1991), and "stigma of inferiority" (Wight et al., 2005). So, the IRT approach used here could be part of a useful strategy for identifying the kinds of content that does, and does not, differ by racial/ethnic group, and thus identifying racial/ethnic group patterns in the experience and expression of depression.

The study's contributions should be viewed in the context of study limitations and strengths. There were important reasons for focusing specifically on early adolescence and eliminating the suicide item, but both features of the study placed certain limits on the generalizability of the findings, suggesting a need for additional research in the future. In addition, the focus of this study on racial/ethnic group differences may have overlooked factors that might account for more variance in CDI responses than race and ethnicity. Indeed, even future research that continues to focus on race and ethnicity might do well to dig more deeply into hypothesized explanatory variables, including, for example, experiences with perceived discrimination and "minority status", economic disadvantage, stressors associated with immigration and acculturation, and group differences in parenting behaviors and criticism (see Helms, Jernigan, & Mascher, 2005).

A related point is that the choice of any particular depression measure will place limits on the theoretical constructs that can be appropriately examined. Consider, for example, the theoretically important distinction in the depression literature between sociotropy (related to interpersonal concerns) and autonomy (related to achievement concerns) (see Robins & Luten, 1991). We began to explore this distinction, but found that results with our dataset did not map neatly onto this body of work. For example, the CDI items "loneliness" and "lack of friends" showed no DIF, and were equivalent across groups. Items involving "reduced social interest", "school performance decrement", and "self-deprecation" displayed generally small group difference effect sizes, and the direction of DIF differed across the items. It is possible that differences in item functioning across racial/ethnic groups are not related in any clear

way to the distinction between sociotropy and autonomy; but it is also possible that the CDI items, which were not designed to capture these two theoretical constructs, in fact do not capture them. Perhaps the best way to examine the sociotropy-autonomy distinction in relation to DIF would be to use a different depression measure, one that is designed specifically to capture those constructs.

The limitations of the study highlight the importance of not applying DIF-adjustments uncritically to youths from racial/ethnic minority backgrounds in research and clinical settings. Cultural and individual differences are quite complex. Sub-group analyses to explore differences by nationality and specific ethnicity (e.g., youths identifying as Mexican may differ significantly from those identifying as Puerto Rican) could not be included due to sample size limitations, but would be a valuable addition in future research with larger samples. Additionally, at the individual participant or patient level, relying on adjusted CDI scores alone may not be sufficient for accurate assessment, and more comprehensive assessments should be considered.

Several strengths of the study also warrant attention. The large, diverse sample of youths made it possible to overcome limitations of previous research on this topic, including samples that were too small for optimal use of IRT, and samples that included only White youths and one minority group. Our inclusion of youths from the four largest racial/ethnic groups in the U.S. allowed cross-ethnic comparisons between majority and minority groups, and across minority groups as well. The geographical diversity of participants may also have added to the generalizability of findings, reducing the likelihood that results would reflect the distinctive culture of a single region of the U.S. The study's focus on the theoretically important period of early adolescence, while limiting the developmental range of the sample, was a strength in other respects. It provided a precise look at the period when rates of depression begin to surge, and it reduced developmental variation in a way that supported the study's emphasis on racial/ethnic groups rather than age differences (see van Beek et al, 2012). Focusing on one discreet developmental period is consistent with the guidance of McLaughlin and colleagues (2007), who caution against making cross-racial/ethnic group comparisons irrespective of developmental period.

The findings suggest that caution should be used when interpreting and drawing conclusions about racial/ethnic group differences in CDI scores, and future research using similar methods may reach similar conclusions about other widely-used clinical measures. When symptom measures are indicative of different severity levels across groups, or are differentially discriminating, it may be unwise to take raw sum scores at their face value; doing so may produce estimates that mask or exaggerate the degree of symptom severity for youths from different racial/ethnic backgrounds. In such cases, the use of scoring algorithms that correct for differential item functioning can reduce bias and produce more interpretable data and more valid conclusions. IRT methods like those used in this study can provide psychological scientists with the tools needed to identify racial/ethnic group differences and correct for them, strengthening research on clinical dysfunction and improving the accuracy of clinical assessment.

Acknowledgments

Special thanks to Matthew Nock for his wise contributions to this study, and to the principals, staff, and children of the participating schools.

Funding

This research was supported was supported by a grant from the National Institute of Mental Health [R01 MH068806] to John Weisz.

References

Agnoff WH (1993) Perspectives on differential item functioning methodology. In Holland PW & Wainer H (Eds.), Differential Item Fucntioning (pp. 3–24). New Jersey: Lawrence Erlbaum Associates.

Allen L, & Astuto J (2012). Depression among racially, ethnically, and culturally diverse adolescents. In Nolen-Hoeksema S & Hilt LM (Eds.), Handbook of Depression in Adolescents (pp. 75–110). New York: Taylor & Francis Group, LLC.

Anderson ER, & Mayes LC (2010). Race/ethnicity and internalizing disorders in youth: A review. Clinical Psychology Review, 30(3), 338–348. doi:10.1016/j.cpr.2009.12.008 [PubMed: 20071063]

Borsboom D (2006). When does measurement invariance matter? Medical Care, 44(11 Suppl 3), S176–181. doi: 10.1097/01.mlr.0000245143.08679.cc [PubMed: 17060825]

Browne MW, & Cudeck R (1993). Alternative ways of assessing model fit. Sage Focus Editions, 154, 136.

Chalmers RP (2012). mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software, 48(6), 1–29.

Choi H (2002). Understanding adolescent depression in ethnocultural context. Advances in Nursing Science, 25(2), 71–85. [PubMed: 12484642]

Choi SW, Gibbons LE, & Crane PK (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. Journal of statistical software, 39(8), 1–30.

Chorpita BF, Yim L, Moffitt C, Umemoto LA, & Francis SE (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: a revised child anxiety and depression scale. Behavior Reseach and Therapy, 38(8), 835–855.

Cole DA, Martin JM, Peeke LG, Henderson A, & Harwell J (1998). Validation of depression and anxiety measures in White and Black youths: Multitrait-multimethod analyses. Psychological Assessment, 10(3), 261–276.

Costello EJ, Mustillo S, Erkanli A, Keeler G, & Angold A (2003). Prevalence and development of psychiatric disorders in childhood and adolescence. Archives of general psychiatry, 60(8), 837–844. [PubMed: 12912767]

Crockett LJ, Randall BA, Shen YL, Russell ST, & Driscoll AK (2005). Measurement equivalence of the center for epidemiological studies depression scale for Latino and Anglo adolescents: a national study. Journal of Consulting and Clinical Psychology, 73(1), 47–58. doi: 10.1037/0022-006X.73.1.47 [PubMed: 15709831]

de Ayala RJ (2009). The theory and practice of Item Response Theory New York: Guilford Press.

Doi Y, Roberts RE, Takeuchi K, & Suzuki S (2001). Multiethnic comparison of adolescent major depression based on the DSM-IV criteria in a U.S.-Japan study. Journal of the American Academy of Child and Adolescent Psychiatry, 40(11), 1308–1315. doi: 10.1097/00004583-200111000-00011 [PubMed: 11699805]

Donnelly M, & Wilson R (1994). The dimensions of depression in early adolesence. Personality and Individual Differences, 17(3), 425–430.

Foladare IS (1969). A clarification of “Ascribed Status” and “Achieved Status”. The Sociological Quarterly, 10(1), 53–61.

Gil AG, & Vega WA (1996). Two different worlds: Acculturation stress and adaptation among Cuban and Nicaraguan families. Journal of Social and Personal Relationships, 13(3), 435–456.

Hambleton RK, & Jones RW (1993). An NCME instructional module on comparison of classical test theory and Item Response Theory and their applications to test development. Educational measurement: issues and practice, 12(3), 38–47.

Hambrick JP, Rodebaugh TL, Balsis S, Woods CM, Mendez JL, & Heimberg RG (2010). Cross-ethnic measurement equivalence of measures of depression, social anxiety, and worry. Assessment, 17(2), 155–171. doi: 10.1177/1073191109350158 [PubMed: 19915199]

Harrell FE (2001). Regression modeling strategies New York: Springer Science & Business Media.

Helms JE, Jernigan M, & Mascher J (2005). The meaning of race in psychology and how to change it: A methodological perspective. American Psychologist, 60(27–36).

Helsel WJ, & Matson JL (1984). The assessment of depression in children: the internal structure of the Child Depression Inventory (CDI). Behavior Research and Therapy, 22(3), 289–298.

Henson RK, Roberts JK (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. Educational and Psychological Measurement, 66(3), 393–416.

Hu L, & Bentler PM (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55.

Iwata N, Turner RJ, & Lloyd DA (2002). Race/ethnicity and depressive symptoms in community-dwelling young adults: A differential item functioning analysis. Psychiatry Research, 110(3), 281–289. [PubMed: 12127478]

Kessler RC, & Wang PS (2009). The epidemiology of depression. In Gotlib IH & L Hammen C (Eds.), Handbook of depression, 2nd *Edition* (pp. 5–22). New York: Guilford Press.

Kistner JA, David CF, & White BA (2003). Ethnic and sex differences in children's depressive symptoms: mediating effects of perceived and actual competence. Journal of Clinical Child and Adolescent Psychology, 32(3), 341–350. doi: 10.1207/S15374424JCCP3203_03 [PubMed: 12881023]

Kovacs M (2004). Children's depression inventory (CDI) Toronto: Multi-Health Systems Inc.

Kovacs M (1992). The children's depression inventory New York: Multi-Health Systems.

Kubik MY, Lytle LA, Birnbaum AS, Murray DM, & Perry CL (2003). Prevalence and correlates of depressive symptoms in young adolescents. Americal Journal of Health Behavior, 27(5), 546–553.

Latzman RD, Naifeh JA, Watson D, Vaidya JG, Heiden LJ, Damon JD, … Young J (2011). Racial differences in symptoms of anxiety and depression among three cohorts of students in the southern United States. Psychiatry, 74(4), 332–348. doi: 10.1521/psyc.2011.74.4.332 [PubMed: 22168294]

Lee YS, Krishnan A, & Park YS (2012). Psychometric properties of the Children's Depression Inventory: An item response theory analysis across age in a nonclinical, longitudinal, adolescent sample. Measurement and Evaluation in Counseling and Development, 45(2), 84–100. doi: 10.1177/0748175611428329

Mather M, Pollard K, & Jacbobsen LA (2011). First results from the 2010 census *Population Reference Bureau: Reports on America*. Retrieved from: http://www.prb.org/Publications/Reports/2011/census-2010.aspx

McHorney CA, & Fleishman JA (2006). Assessing and understanding measurement equivalence in health outcome measures: issues for further quantitative and qualitative inquiry. Medical care, 44(11), S205–S210. [PubMed: 17060829]

McLaughlin KA, Hilt LM, & Nolen-Hoeksema S (2007). Racial/ethnic differences in internalizing and externalizing symptoms in adolescents. Journal of Abnormal Child Psychology, 35(5), 801–816. doi: 10.1007/s10802-007-9128-1 [PubMed: 17508278]

Merikangas K, He J, Burstein M, Swanson SA, Avenevoli S, Cui L, & … Swendsen J (2010). Lifetime prevalence of mental disorders in U.S. adolescents: Results from the National Comorbidity Survey Replication-Adolescent Supplement (NCS-A). Journal Of The American Academy Of Child & Adolescent Psychiatry, 49(10), 980–989. doi:10.1016/j.jaac.2010.05.017 [PubMed: 20855043]

Merikangas KR, & Knight E (2012). The epidemiology of depression in adolescents. In Nolen-Hoeksema S & Hilt LM (Eds.), Handbook of depression in adolescents (pp. 53–74). New York: Taylor & Francis Group, LLC.

Mittlboeck M, & Schemper M (1996). Explained variation for logistic regression. Statistics in Medicine, 15, 1987–1997. [PubMed: 8896134]

Nock MK, Deming CA, Chiu W, Hwang I, Angermeyer M, Borges G, & … Kessler RC (2012). Mental disorders, comorbidity, and suicidal behavior. In Nock MK, Borges G, Ono Y(Eds.), Suicide: Global perspectives from the WHO World Mental Health Surveys (pp. 148–163). New York, NY US: Cambridge University Press.

Nock MK, Hwang I, Sampson N, Kessler RC, Angermeyer M, Beautrais A, … & Williams DR (2009). Cross-national analysis of the associations among mental disorders and suicidal behavior: findings from the WHO World Mental Health Surveys. PLoS Medicine, 6(8), e1000123. [PubMed: 19668361]

Nguyen HT, Kitner-Triolo M, Evans MK, & Zonderman AB (2004). Factorial invariance of the CES-D in low socioeconomic status African Americans compared with a nationally representative sample. Psychiatry research, 126(2), 177–187. [PubMed: 15123397]

Paxton RJ, Valois RF, Watkins KW, Huebner ES, & Drane JW (2007). Sociodemographic differences in depressed mood: results from a nationally representative sample of high school adolescents. Journal of School Health, 77(4), 180–186. doi: 10.1111/j.1746-1561.2007.00189.x [PubMed: 17425520]

Peña J, Matthieu M, Zayas L, Masyn K, & Caine E (2012). Co-occurring risk behaviors among White, Black, and Hispanic U.S. high school adolescents with suicide attempts requiring medical attention, 1999–2007: Implications for future prevention initiatives. Social Psychiatry & Psychiatric Epidemiology, 47(1), 29–42. doi:10.1007/s00127-010-0322-z [PubMed: 21153018]

Peña JB, Zayas LH, Cabrera-Nguyen P, & Vega WA (2012). U.S. cultural involvement and its association with suicidal behavior among youths in the Dominican Republic. American Journal Of Public Health, 102(4), 664–671. doi:10.2105/AJPH.2011.300344 [PubMed: 22397348]

Penfield RD, Gattamorta K, & Childs RA (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. Educational Measurement: Issues and Practice, 28(1), 38–49.

Politano PM, Nelson WM, Evans HE, Sorenson SB, & Zeman DJ (1986). Factor analytic evaluation of differences between Black and European American emotionally disturbed children on the Children's Depression Inventory. Journal of Psychopathology and Behavioral Assessment, 8, 1–7.

R Core Team (2015). R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna, Austria URL http://www.R-project.org/.

Radloff LS (1977). The CES-D scale a self-report depression scale for research in the general population. Applied psychological measurement, 1(3), 385–401.

Revelle W (2015). psych: Procedures for psychological, psychometric, and personality research R package version 1.5.1

Revelle W, & Rocklin T (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behavioral Research, 14, 403–414. [PubMed: 26804437]

Reise SP, & Waller NG (2009). Item response theory and clinical measurement. Annual Review of Clinical Psychology, 5, 25–46. doi: 10.1146/annurev.clinpsy.032408.153553.

Roberts RE (1992). Manifestation of depressive symptoms among adolescents: A comparison of Mexican Americans with the majority and other minority populations. The Journal of nervous and mental disease, 180(10), 627–633. [PubMed: 1402840]

Roberts RE, Roberts CR, & Chen YR (1997). Ethnocultural differences in prevalence of adolescent depression. American Journal of Community Psychology, 25(1), 95–110. [PubMed: 9231998]

Roberts RE, & Sobhan M (1992). Symptoms of depression in adolescence: A comparison of Anglo, African, and Hispanic Americans. Journal of Youth and Adolescence, 21, 639–651. [PubMed: 24264167]

Robins CJ, & Luten AG (1991). Sociotropy and autonomy: Differential patterns of clinical presentation in unipolar depression. Journal of Abnormal Psychology, 100(1), 74–77. 10.1037/0021-843X.100.1.74 [PubMed: 2005274]

Saluja G, Iachan R, Scheidt PC, Overpeck MD, Sun W, & Giedd JN (2004). Prevalence of and risk factors for depressive symptoms among young adolescents. Archives of Pediatric and Adolescent Medicine, 158(8), 760–765. doi: 10.1001/archpedi.158.8.760

Samejima F (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph, 35(1), 139.

Saylor CF, Finch AJ, Spirito A, & Bennett B (1984). The Children's Depression Inventory: A systematic evaluation of psychometric properties. Journal of Consulting and Clinical Psychology, 52, 955–967. [PubMed: 6520288]

Sick J (2008). Rasch measurement in language education: Part 1. Japanese Assocation for Language Testing: Testing and Evaluation SIG Newsletter, 12(1), 1–6.

Steele RG, Little TD, Ilardi SS, Forehand R, Brody GH, & Hunter HL (2006). A confirmatory comparison of the factor structure of the Children's Depression Inventory between European American and African American youth. Journal of Child and Family Studies, 15, 779–794. doi: 10.1007/s10826-006-9054-9.

Thorndike RL (1953). Who Belong in the Family? Psychometrika, 18(4), 267–276.

Trent LR, Buchanan E, Ebesutani C, Ale CM, Heiden L, Hight TL, … Young J (2013). A measurement invariance examination of the Revised Child Anxiety and Depression Scale in a Southern sample: differential item functioning between African American and Caucasian youth. Assessment, 20(2), 175–187. doi: 10.1177/1073191112450907 [PubMed: 22855507]

Twenge JM, & Nolen-Hoeksema S (2002). Age, gender, race, socioeconomic status, and birth cohort differences on the children's depression inventory: a meta-analysis. Journal of Abnormal Psychology, 111(4), 578–588. [PubMed: 12428771]

U.S. Census Bureau. (2013). International Migration is Projected to Become Primary Driver of U.S. Population Growth for First Time in Nearly Two Centuries (CB13–89). Retrieved from: http://www.census.gov/newsroom/releases/archives/population/cb13-89.html.

U.S. Department of Health and Human Services. (2001). Mental Health: Culture, Race, and Ethnicity —A Supplement to Mental Health: A Report of the Surgeon General Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Center for Mental Health Services.

van Beek Y, Hessen DJ, Hutteman R, Verhulp EE, & van Leuven M (2012). Age and gender differences in depression across adolescence: real or 'bias'?. Journal of Child Psychology and Psychiatry, 53(9), 973–985. [PubMed: 22512614]

Velicer WF (1976) Determining the Number of Components from the Matrix of Partial Correlations. Psychometrika, 41(3), 321–327.

Weiss B, Weisz JR, Politano M, Carey M, Nelson WM, & Finch AJ (1991). Developmental differences in the factor structure of the Children's Depression Inventory. Psychological Assessment, 3, 38–45.

Weisz JR, McCarty CA, & Valeri SM (2006). Effects of psychotherapy for depression in children and adolescents: A meta-analysis. Psychological Bulletin, 132, 132–149. [PubMed: 16435960]

World Health Organization. (2004). The Global Burden of Disease: 2004 Update Geneva.

Zwick WR, & Velicer WF (1986). Comparison of Five Rules for Determining the Number of Components to Retain. Psychological Bulletin, 99(4), 432–442

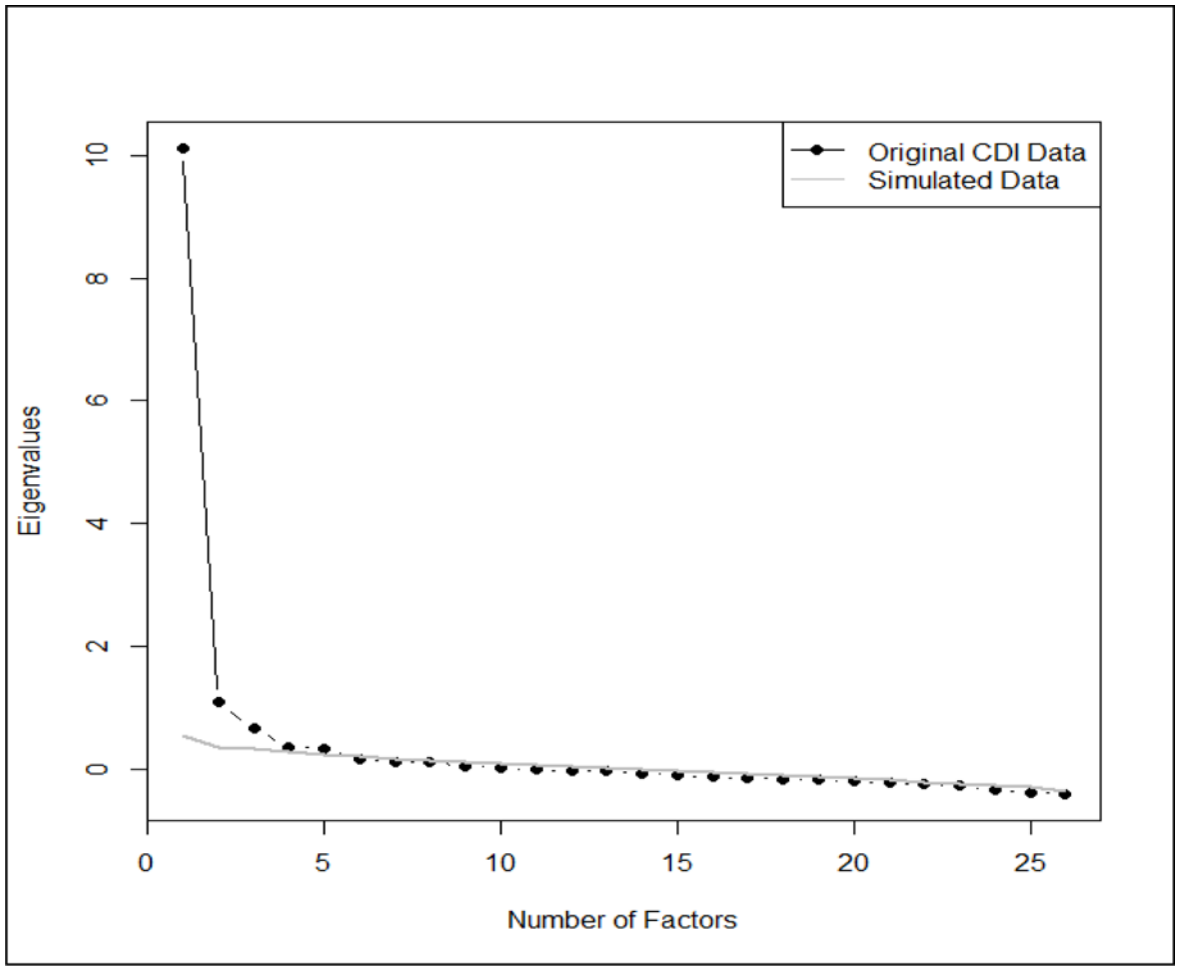

Figure 1.

Scree plot for ordinal Exploratory Factor Analysis (EFA). The ordinal factor analysis is based on polychoric correlations. Eigenvalues along the y-axis correspond to two sets of factors along the x-axis. The eigenvalues illustrated by the black line and labeled “Original CDI Data” represent the full-sample solution for the original empirical dataset. The eigenvalues illustrated by the grey line and labeled “Simulated data” represent a parallel analysis, for which the plotted values indicate the average eigenvalues computed from a series of random data matrices of the same size as the original one. The simulated solution represents the eigenvalues that would be derived from the factor-analysis by chance. The contrast between the two sets of eigenvalues indicates that our EFA solution differs from chance. Based on this plot in combination with the fit statistics for the EFA we conclude that the data for the full sample are unidimensional.

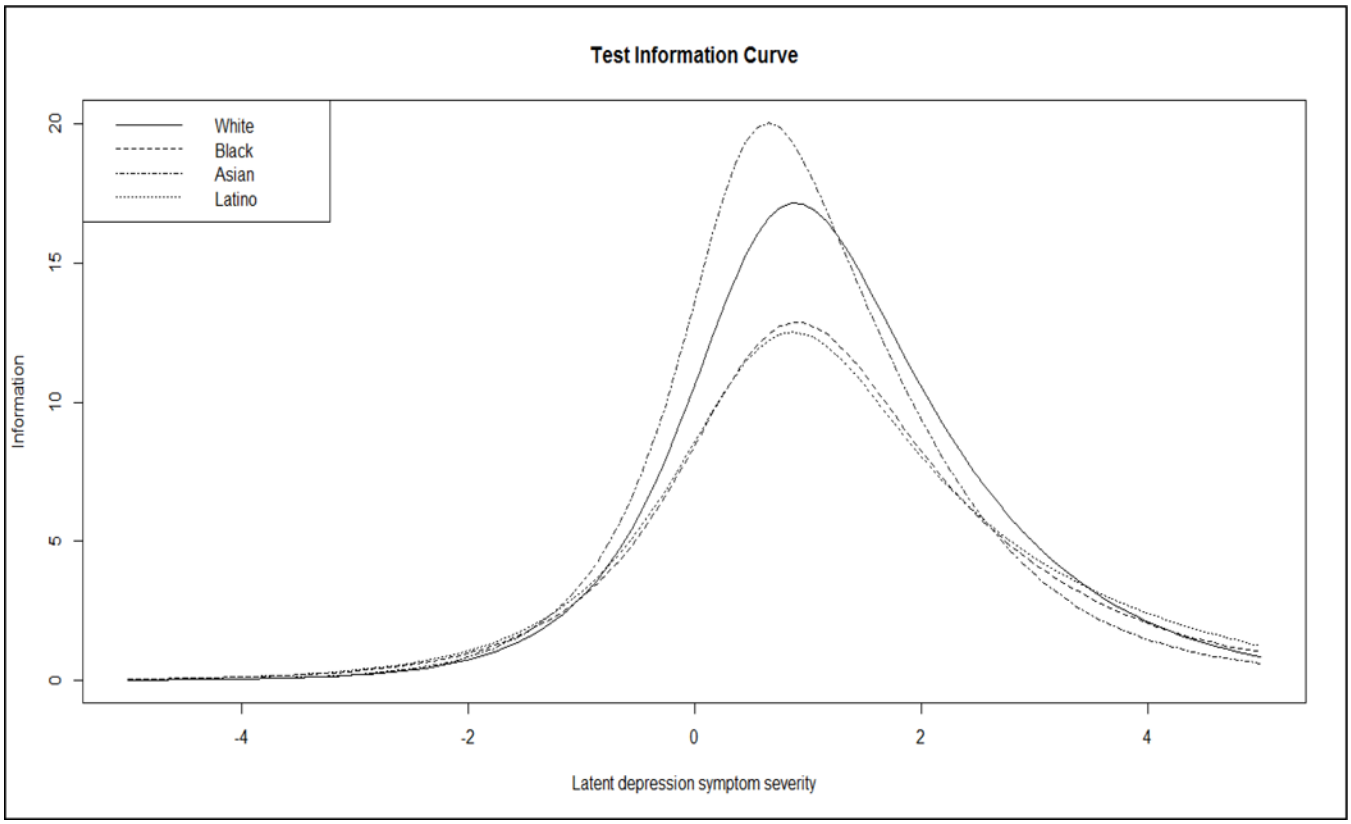

Figure 2.

Test information curve for the full CDI scale by racial/ethnic group. Test information curves depict the precision of the estimates of latent depression symptom severity across the full range of scores, for each group. Test information, or precision, on the y-axis is calculated by taking the reciprocal of the variance of the item parameters (discrimination and category-threshold location). The greatest amount of information is available for all four groups at symptom severity levels between approximately −0.50 and 2.50 (total sample $M = 0.01$) on the x-axis. Yet, the CDI demonstrates the greatest precision for measuring symptom severity in youths identifying as Asian, followed by youths identifying as White; the scale has lower precision for youths identifying as Black and Latino compared to youths identifying as Asian or White.

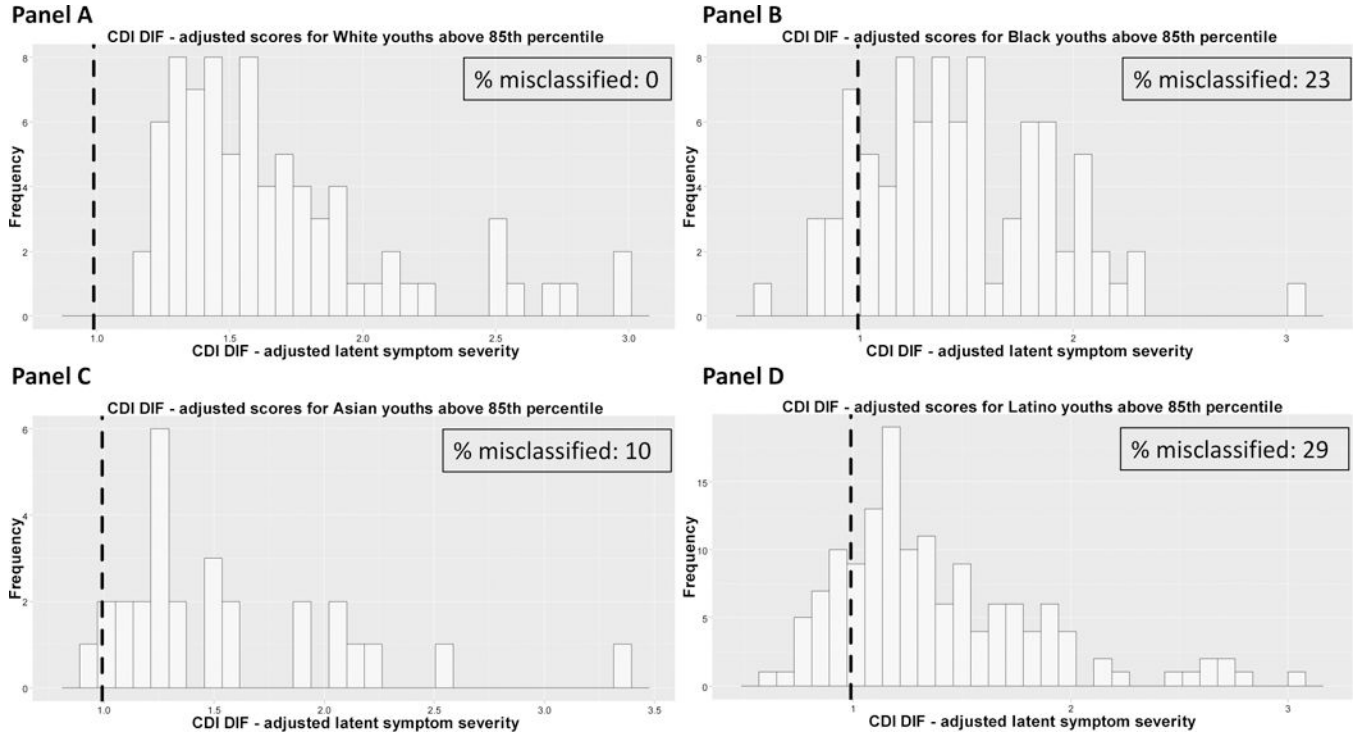

Figure 3.

Distribution of differential item functioning (DIF)-adjusted clinically-elevated symptom severity scores (for youths classified into the clinically-elevated range based on original CDI total scores) across racial/ethnic groups. Vertical dashed lines represent the DIF-adjusted cut-off for 85th percentile latent severity scores. All youths to the left of the cut-off were originally classified into the clinically-elevated range by CDI total scores, but do not meet the criterion for clinically-elevated symptoms once scores are adjusted for DIF. White youths were classified equivalently regardless of scoring method (Panel A). Black youths were not classified equivalently across scoring methods; 23% were misclassified into the clinically-elevated total score and did not meet DIF-adjusted criteria (Panel B). For Asian youths, 10% were misclassified (Panel C). For Latino youths, 29% were misclassified (Panel D).

Table 1.

Racial/ethnic group-specific item discrimination and location parameters for each item displaying Differential Item Functioning (DIF)

		White			Black			Asian			Latino		
CDI Items		α	b1	b2	α	b1	b2	α	b1	b2	α	b1	b2
1	I am sad all the time	--	--	--	--	--	--	--	--	--	--	--	--
2	Nothing will ever work out for me	1.95	1.08	NA	1.59	0.96	NA	1.60	0.98	NA	1.60	0.53	NA
3	I do everything wrong	2.44	1.79	NA	1.53	1.57	NA	3.33	1.58	NA	1.59	1.61	NA
4	Nothing is fun at all	1.45	1.42	NA	1.31	0.78	NA	1.33	0.96	NA	1.02	0.79	NA
5	I am bad all the time	1.47	2.28	NA	0.82	2.25	NA	1.45	1.86	NA	0.82	2.41	NA
6	I am sure that terrible	1.32	1.15	3.72	0.71	0.95	5.17	2.03	0.73	2.2	0.97	0.70	3.98
7	things will happen to me I hate myself	2.67	1.62	2.93	2.33	1.78	2.84	3.10	1.24	0 2.23	2.37	1.55	2.54
8	All bad things are my fault	1.67	1.71	NA	1.49	1.35	NA	1.77	1.27	NA	1.47	1.24	NA
10	I feel like crying every day	--	--	--	--	--	--	--	--	--	--	--	--
11	Things bother me all the time	1.69	0.61	2.46	1.33	0.95	2.64	1.90	0.75	1.87	1.51	1.01	2.50
12	I do not want to be with people at all	1.18	2.51	NA	1.11	1.56	NA	1.35	1.94	NA	1.36	1.90	NA
13	I cannot make up my mind about things	--	--	--	--	--	--	--	--	--	--	--	--
14	I look ugly	1.44	0.84	3.09	1.71	1.13	2.67	1.44	0.52	2.92	1.12	0.75	3.40
15	I have to push myself all the time to do my schoolwork	1.49	0.96	2.24	1.11	0.37	2.12	1.45	0.85	2.07	0.96	0.20	2.02
16	I have trouble sleeping every night	--	--	--	--	--	--	--	--	--	--	--	--
17	I am tired all the time	1.26	0.31	2.42	0.98	0.99	2.74	1.25	0.42	2.34	0.95	1.08	3.03
18	Most days I do not feel like eating	1.15	2.11	3.65	0.94	1.14	2.37	0.91	2.11	3.82	1.11	1.16	2.18
19	I do not worry about aches and pains	1.14	1.50	3.95	0.93	0.97	3.14	1.60	1.05	2.63	0.90	0.77	3.42
20	I do not feel alone	--	--	--	--	--	--	--	--	--	--	--	--
21	I never have fun at school	0.83	0.92	4.47	0.77	0.16	3.84	1.02	0.80	3.50	0.94	0.61	3.63
22	I do not have any friends	--	--	--	--	--	--	--	--	--	--	--	--
23	I do very badly in subjects I used to be good in	1.67	1.46	2.73	1.06	0.90	2.45	1.93	1.40	2.73	1.08	0.77	2.38
24	I can never be as good as other kids	1.69	1.01	2.54	1.18	0.51	2.77	1.33	0.82	2.32	1.14	0.22	2.50
25	Nobody really loves me	1.31	2.52	NA	1.93	1.93	NA	1.40	1.39	NA	1.62	1.77	NA
26	I never do what I am told	1.57	1.65	NA	0.79	1.12	NA	1.26	1.60	NA	0.96	0.84	NA
27	I get into fights all the time	1.70	2.05	NA	1.04	1.73	NA	1.83	1.88	NA	1.17	1.83	NA

Note. Items with missing parameters were not flagged for DIF and group-specific parameters were not needed. α = discrimination parameter, the strength of association between the item and the underlying construct of depression severity (analogous to the correlation between the item score and latent symptom severity). b1 = category threshold location 1, the level of symptom severity along the latent continuum where participants have a greater probability of endorsing a response of “1” compared to a response of “0”. b2 = category threshold location 2, the level of symptom severity along the latent continuum where participants have a greater probability of endorsing a response of “2” compared to a response of “1”. NA = second category threshold is not applicable because there were fewer than 5 observations in the response category and responses were collapsed to “0” and “1” only.