Published in final edited form as:
J Abnorm Child Psychol. 2012 April ; 40(3): 353–367. doi:10.1007/s10802-011-9569-4.

Measurement Equivalence Across Racial/Ethnic Groups of the Mood and Feelings Questionnaire for Childhood Depression

My K. Banh,
Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. Center for Child and Human Development, Georgetown University, 3300 Whitehaven Street, NW, Suite 3300, Washington, DC 20007, USA. Department of Psychiatry, Georgetown University Hospital, Washington, DC, USA

Paul K. Crane,
Department of Internal Medicine, University of Washington, Seattle, WA, USA

Isaac Rhew,
Social Development Research Group, University of Washington, Seattle, WA, USA

Gretchen Gudmundsen,
Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. Center for Child Health, Behavior and Development, Seattle Children's Hospital, Seattle, WA, USA

Ann Vander Stoep,
Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. Department of Epidemiology, University of Washington, Seattle, WA, USA

Aaron Lyon, and
Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA

Elizabeth McCauley
Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA, USA. Center for Child Health, Behavior and Development, Seattle Children's Hospital, Seattle, WA, USA

My K. Banh: My.Banh@gmail.com; Paul K. Crane: pcrane@u.washington.edu; Isaac Rhew: rhew@u.washington.edu; Gretchen Gudmundsen: gretchen.gudmundsen@seattlechildrens.org; Ann Vander Stoep: annv@u.washington.edu; Aaron Lyon: aaron.lyon@seattlechildrens.org; Elizabeth McCauley: eliz@u.washington.edu

Abstract

As research continues to document differences in the prevalence of mental health problems such as depression across racial/ethnic groups, the issue of measurement equivalence becomes increasingly important to address. The Mood and Feelings Questionnaire (MFQ) is a widely used screening tool for child and adolescent depression. This study applied a differential item functioning (DIF) framework to data from a sample of 6th and 8th grade students in the Seattle Public School District (*N*=3,593) to investigate the measurement equivalence of the MFQ. Several items in the MFQ were found to have DIF, but this DIF was associated with negligible individual- or group-level impact. These results suggest that differences in MFQ scores across groups are unlikely to be caused by measurement non-equivalence.

Correspondence to: My K. Banh, My.Banh@gmail.com.

Keywords

MFQ; IRT; DIF; Depression; Measurement equivalence

Introduction

Depression is a common mental health disorder among adolescents in the United States (US) (Anderson and Mayes 2010; Birmaher et al. 1996; Garrison et al. 1997). Prevalence estimates among different racial/ethnic groups vary widely. Saluja and colleagues (2004) found that American Indian youths reported the highest prevalence of depressive symptoms (29%), followed by Hispanic (22%), non-Hispanic White (18%), Asian American (17%), and African American (15%) youths. Roberts and colleagues (1997) found that the prevalence of depression among Mexican American adolescents was 12%, while for non-Hispanic white adolescents it was 6.6% without adjustment for impairment. Furthermore, the pattern of specific depressive symptoms endorsed by depressed youth has also been reported to vary across racial/ethnic groups. In comparison to white youth, higher proportions of Hispanic American youth endorse somatic symptoms (such as headaches and stomachaches), decreased pleasure, fatigue, low self-esteem, crying, and concentration difficulties (Choi and Gi Park 2006; Roberts and Sobhan 1992); higher proportions of African American youth endorse anhedonia (Choi and Gi Park 2006); and higher proportions of Asian American youth endorse depressed mood and low self-esteem (Choi and Gi Park 2006; Choi et al. 2006). Reported differences in symptom endorsement and varying prevalence of depressive disorders across groups may reflect actual ethnic/racial differences in symptom manifestation. However, inconsistencies in findings may also relate to how depression is measured across studies, with some research using symptom counts and others using different depression scales, often with varying clinical cut-off points. Moreover, the measurement properties of specific depression scales used in studies may also differ across racial/ethnic groups. In this paper, we explore the measurement properties of one such measure across four racial/ethnic groups.

Measurement Non-equivalence

Measurement non-equivalence can have important clinical and policy implications. According to Drasgow and Kanfer (1985), when measures are non-equivalent, observed scores from different groups or subgroups are on different scales. Therefore, interpretation of any observed differences in mean scores between groups would not be meaningful. To date, many depression screening measures have been developed and tested predominantly with white youth (Crockett et al. 2005). It is therefore unclear how well these measures assess adolescent depressive symptoms across different racial/ethnic groups. As noted previously, a few studies have suggested that certain symptoms of depression may be endorsed more frequently by depressed members of some groups (Choi and Gi Park 2006; Choi et al. 2006). If measures are non-equivalent across racial/ ethnic groups, group comparisons could be misleading. For example, if a measure of depression is valid for one group but less so for another, using a clinical cut-off score that is optimal for the group measured validly may result in under-or over-estimates of the prevalence of depression for the second group (Crockett et al. 2005). Thus, using differentially valid measures of depression can lead to inaccurate estimates of the prevalence of depression for some racial/ ethnic groups that could potentially misinform national initiatives to address child and adolescent depression.

Measurement Equivalence and the Mood and Feelings Questionnaire

The Mood and Feelings Questionnaire (MFQ) is a widely used screening measure of depressive symptomatology for children 8–18 years of age (Angold et al. 1995). Like many clinical instruments, its psychometric properties have been evaluated in samples consisting primarily of white children and adolescents (Daviss et al. 2006; Thapar and McGuffin 1998; Wood et al. 1995), although it has subsequently been used with more racially and ethnically diverse samples in the Great Smoky Mountains Study (Messer et al. 1995; Angold et al. 1995), including reports of racial/ethnic group differences in depressive symptomatology (e.g., Bisaga et al. 2005). However, to our knowledge, the equivalence of the MFQ across different racial/ethnic groups has not yet been examined. Until this step is completed, it is not possible to determine the meaning of between-group differences.

Previous studies have addressed measurement invariance in depression scales (e.g., Perreira et al. 2005), but the statistical tools they used have been limited. In many cases, the first step of these investigations involved selecting children who met some criterion for depression, thus selecting children at one end of the depression spectrum. This approach stratifies but does not further match on the level of underlying depression. For example, it is possible that the average depression level among depressed African American children might be much higher than that among depressed white children. If that were the case, then different items being endorsed by African American children than by white children might not reflect measurement bias, but instead might reflect the different average severity of disease in one group of children than another. Without matching on (rather than stratifying by) depression level, it is impossible to differentiate these two potential explanations, making it impossible to determine whether finding different patterns of endorsement reflects different depression severity (i.e. completely a measurement issue), reflects some sort of culturally specific manifestations of depression (i.e., completely a cultural issue), or some mixture of both of these explanations.

In this study, we propose to evaluate the measurement equivalence of the Mood and Feelings Questionnaire for depression across racial/ethnic groups from an item response theory (IRT)/differential item functioning (DIF) framework. To our knowledge, only one study has applied an IRT framework to the short form of the MFQ (SMFQ) (Sharp et al. 2006). Their IRT analyses confirmed the initial finding by Angold and colleagues (1995) that MFQ items appeared to be unidimensional. Sharp et al. (2006) also found that items in the SMFQ discriminated well among 7–11 year old children with more severe levels of depression. No study has yet applied an IRT/DIF approach to the long form of the MFQ.

Evaluating items in a scale using the IRT/DIF framework to identify potential test bias has the advantage of matching on the level of underlying depression. The necessity to control for the underlying trait level in evaluating measurement equivalence is a point made repeatedly in the literature (e.g. Camilli and Shepard 1994; Holland and Wainer 1993; Millsap and Everson 1993). Furthermore, DIF techniques consider the entire spectrum of depression, including levels of depression that are ignored when considering the prevalence of item endorsement only among depressed children. Finally, unlike classical test theory approaches (including the prevalence of item endorsement approach), DIF techniques permit us to comment on the possible impact of biased items, and whether scores are affected in a meaningful way by ignoring or accounting for DIF.

Item Response Theory and Differential Item Functioning

Over the past several decades, IRT has become the dominant paradigm in educational testing for construction and evaluation of academic test batteries. Only recently has IRT been used widely within psychological research to assess construct validity of psychological measures (Embretson and Reise 2000). An advantage of IRT is that it provides a detailed explication of the relationship between test items and levels of the latent or underlying trait measured by the test (Embretson and Reise 2000). IRT accomplishes this by modeling the probability of endorsing an item as a function of its location along a continuum of trait levels.

A central concept in IRT is the underlying trait measured by the instrument. This trait is represented by the Greek character theta (θ). Throughout this paper, we will refer to the specific underlying trait of interest as "depression level," IRT score, or θ. Depression level comprises the combined effects of experiential, environmental, and genetic factors that influence endorsement of depression test items, without modeling specific contributions from any one factor. An individual with high depression level (i.e., high θ) has a higher probability of endorsing each item than an individual with a low depression level (low θ). The item characteristic curve (ICC) is a plot of the association between depression level and the probability of endorsing the item.

IRT relies on two important assumptions that should be checked before using these models. The first assumption is that the scale can be considered to be unidimensional and, in turn, it is appropriate to model item responses using a single factor confirmatory factor analysis model. The second, and related, assumption is of local independence; that is, that given the underlying factor measured by test items, responses to the items in the scale are independent of each other (Embretson and Reise 2000).

Two people with the same depression level should have the same probability of endorsing any particular item, even if they represent different demographic groups (Embretson and Reise 2000). If the probability of endorsing an item given the same depression level varies systematically across demographic groups, this is known as differential item functioning (DIF) (Camilli 1994). Two types of DIF are described in the literature: uniform and non-uniform. Uniform DIF is analogous to the epidemiological concept of confounding by group membership (Crane et al. 2004) and occurs when the probability of endorsing an item is greater for one demographic group than another across all depression levels. If an item has uniform DIF, ICCs across demographic groups associated with DIF are parallel and will not cross (c.f. Fig. 1A). Non-uniform DIF operates much like an interaction effect or effect modification in epidemiological research, where group differences in the probability of endorsing an item vary across depression levels (Crane et al. 2004). If an item has non-uniform DIF, the ICC curves are not parallel with each other, and the ICCs may cross within the region of the scale examined. In sum, uniform DIF relates to differences in item difficulty across group, while non-uniform DIF relates to differences in item discrimination. Both can be present for the same item.

In the present study we analyzed data from the MFQ that has three response options for each item. The above discussion referred to the case of dichotomous items (endorsed/not endorsed). The extension to multiple response categories is straightforward; we used Samejima's graded response model (Samejima 1969, 1997), which uses item category characteristic curves (ICCCs) rather than ICCs.

DIF detection is an increasingly common analytical approach to evaluate item bias in psychological research, including depression scales (see Teresi et al. 2009, for a review). The existence of DIF for a large number of items may threaten the construct validity of the measure and the conclusions drawn from studies using the measure (Crane et al. 2004).

Assessment of item-level data on the presence of DIF and the impact of DIF are important steps in identifying biased scale items and determining their potential impact on conclusions drawn for individuals and groups. The finding of DIF presence allows scale developers to review particularly concerning items. Individual-level DIF impact could affect conclusions drawn about an individual person, while group-level DIF impact could affect conclusions drawn about different groups of people. See Crane et al. (2007) for a more detailed discussion.

Study Goal

The goal of this study was to assess measurement equivalence of the long form of the MFQ across racial/ ethnic groups by assessing MFQ items for DIF in Asian, African American, Hispanic, and non-Hispanic white students. We investigated the presence of uniform and non-uniform DIF and DIF impact related to gender, grade level, parents' birth place, and race/ethnicity.

Method

Study Population—This study used screening data from the Development Pathways Project (DPP) (*N*=2,187) and the High School Transition Study (HSTS) (*N*=2,665). Both studies were approved by the University of Washington Institutional Review Board and the Seattle Public School District Office of Research and Evaluation. Both studies used the MFQ to screen students for depression. The DPP study screened 6th graders while the HSTS study screened 8th graders. The combined dataset included 4,852 participants, of whom 1,259 participants were not eligible for these analyses, resulting in a final study sample size of 3,593. Reasons for exclusion include missing data on race/ethnicity (*N*=16) or parent's birth place (*N*=517). Other reasons for exclusion included small cell size (e.g., those who self-identified as Native American/Alaskan Native (*N*=32) or Native Hawaiian/Pacific Islander (*N*=55)) and self-identification of more than one race/ethnicity (*N*=642).

Setting—The study was carried out in Seattle, Washington, the largest urban center in the Pacific Northwest. A total of 46,730 students, 68% of school-aged children residing in Seattle, were enrolled in the Seattle Public School District (SPSD) during 2003–2004. According to data provided from the District, of the approximately 10,000 middle school students enrolled in 10 public schools, 23% were African Americans, 23% were Asian Americans, 10% were Hispanics, and 41% were non-Hispanic whites. In all, 35% of middle school students received free and reduced price lunches, an indicator of economic hardship.

Participants—Participants were recruited from SPSD middle schools. DPP recruited from four middle schools (47% female), and HSTS recruited from six middle schools (53% female). The middle schools were located in distinct geographic and demographic areas of Seattle, and had a racial/ethnic distribution very similar to that of the district as a whole. For both studies, students were eligible for participation if (1) they were enrolled in school at the time of screening, (2) they obtained written permission from their parent or guardian; and (3) they were determined to understand English at a 3rd grade level or higher. Students were not eligible for the DPP study if they had moderate or severe developmental delays. Students were also not eligible for the HSTS study if they were placed in a self-contained class for Serious Emotional Disturbance.

Procedures—Participation was voluntary in both studies and recruitment proceeded similarly. Recruitment materials, including a letter from the school principal, a study information sheet, and parent/caregiver consent forms, were sent to parents/guardians of eligible students by mail or given to students to bring home. For the DPP study, information about the research project was also distributed during assemblies, "back-to-school nights," and Parent Teacher Association meetings. For both studies, students whose parents consented were given the option to decline participation. Interested students completed an informed assent form prior to participating. Children who participated in either study completed a screening questionnaire administered in classrooms by study staff during one 50-min class period.

Treatment of Participants Who Were in Both Study Samples—Based on the number of schools participating in both DPP and HSTS, we estimated that a small number of students (9%) that were sixth graders at the time of the DPP study also participated as eighth graders in the HSTS study. The primary focus of the present analyses is on cross-sectional comparisons across racial/ethnic groups, and the fact that a student was included in both grade levels should not bias those comparisons using DIF analyses. Both datasets used codes to de-identify the study participants. However, the codes were different in the two studies when the participants were combined in this study, making it impossible for us to identify individuals who participated in both DPP and HSTS in this dataset. This de-identification also prevented us from determining which participants received free- or reduced-price lunch.

Mood and Feelings Questionnaire (MFQ)

The MFQ is a 33-item questionnaire developed to screen for depression in epidemiological studies of children and adolescents ages 8 through 18 (Costello and Angold 1988). It captures symptoms of depression included in the *DSM-IV* criteria for major depressive disorder (Angold et al. 1995). Additional items such as those assessing loneliness and feeling unloved were added due to perceived clinical significance. In both the DPP and HSTS, three suicide items (i.e., “I thought that life wasn’t worth living”, “I thought about death and dying”, and “I thought about killing myself”) were eliminated due to the research team’s inability to adequately follow up positive endorsements. The MFQ statements asked respondents to rate on a 3-point scale (0 = Not true, 1 = Sometimes, and 2 = True) how much they have felt or acted that way in the past two weeks. None of the items required reverse-coding. The items were designed to closely match the wording of the Diagnostic Interview Schedule for Children, a validated, structured psychiatric interview for children (Costello et al. 1985) and that of the DSM diagnostic criteria.

Data Analyses

Race/Ethnicity and Grade Level—Descriptive statistics were obtained for racial/ethnic group identification of study participants. In this study, the racial/ethnic categories were created based on responses to two close-ended questions that asked youth to mark “Yes/No” to “I am Latino(a) or Hispanic;” and, to mark “Asian, Black/African American, Native American/Alaskan Native, Native Hawaiian/Pacific Islander, or White/Caucasian” in response to the statement, “My race/ethnicity is (mark all that apply):” Using these data, we created four mutually-exclusive racial/ethnic categories: 1) non-Hispanic white, 2) non-Hispanic Asian, 3) non-Hispanic Black/African American, and 4) Hispanic, any race. For simplicity, we refer to these categories as white, Asian, African American and Hispanic throughout the remainder of this paper. In this study, grade level was also used as the closest proxy to age since participant’s actual ages were not available.

Parent’s Birth Place—Parent’s birth place was included as a potential source of DIF as there is a growing literature that shows that second-generation immigrants are more likely than first-generation immigrants to have higher levels of psychopathology, including depression, especially among Asian Americans (Alegria et al. 2008; Abe-Kim et al. 2007; Takeuchi et al. 2007a, b, c; Breslau and Chang 2006 & Harker 2001). Since we do not have information on generational status among the students, we used parent’s birth place as a proxy for each child’s generational status. Parent’s birth place was operationalized in the analyses as a dichotomous variable where US-born was defined as having at least one parent born in the United States; non-US-born was defined as having both parents born outside of the United States. US-born served as the reference group.

Prevalence of Depression—In this study, we used a clinical cut-off MFQ score of ≥27 to identify youth who were likely depressed. This cut-off has been shown to yield optimal sensitivity and specificity of the 33-item MFQ among a clinical sample of depressed youth (Wood et al. 1995). The prevalence of depression was calculated for each racial/ ethnic group. An omnibus chi-square test was performed to test the null hypothesis that there were no significant race/ ethnic group differences in the prevalence of depression.

Dimensionality Analyses—We analyzed MFQ item data using exploratory and confirmatory factor analytic (EFA and CFA) approaches to assess sufficient unidimensionality to use IRT. We used Mplus for all dimensionality analyses (Muthen and Muthen 1998–2004). We obtained eigenvalues from EFA and generated a scree plot (detailed analyses available on request from second author). For the CFA analyses, we used the weighted least squares with mean and variance (WLSMV) estimator applied to the polychoric correlation matrix to account for the categorical nature of the data (Beauducel and Herzberg 2006; Muthen et al. 1997). To assess model fit, we examined the confirmatory fit index (CFI), Tucker-Lewis index (TLI), and the root mean squared error of approximation (RMSEA). The CFI and TLI indicate good fit with values >0.90 (Bentler 1990), while the RMSEA indicates good fit with values <0.08 for categorical item response data (Brown and Cudeck 1993).

Mplus evaluates associations not modeled in the CFA to identify those that will have the greatest improvement in fit, and reports modification indices for each of these associations. We used these modification indices to suggest residual correlations to consider including in the CFA model. In two stages we examined the content of items with the largest modification indices and freed up residual correlations for those pairs that made clinical sense. For example, the largest modification index was for the residual correlation between items 4 (“I ate more than usual”) and 29 (“I slept more than usual”). These items address vegetative symptoms and include the same phrase “more than usual,” so a methods effect seemed plausible. The 7th largest modification index was for items 29 (“I didn’t sleep as well as I usually did”) and 23 (“I worried about aches and pains”). These items did not share any common phrase and did not seem thematically linked, so we did not include this residual correlation in comparator CFA models, as we did not include residual correlations between any pairs of items where methods effects seemed implausible or where thematic content did not seem clinically related. We compared fit for models with and without clinically sensible residual correlations, and we compared loadings on the primary factor and the magnitude of the standardized residual correlations.

Dimensionality analyses indicated sufficient unidimensionality of the MFQ. A one factor EFA model fit well. The first eigenvalue was 14.2, and the ratio of the first to the second eigenvalue was 8.24, which was easily over the standard rule of thumb of a ratio of 4 that is suggestive of unidimensionality for EFA (Reeve et al. 2007).

CFA results suggested that a single-factor CFA model fit well for the MFQ when residual correlations were included. Standardized factor loadings on the primary factor ranged from 0.25 to 0.88 for the model without residual correlations and from 0.24 to 0.86 for the model with residual correlations. Without the residual correlations included, CFI was 0.89, TLI was 0.98, and RMSEA was 0.05. With empirically guided residual correlations included, CFI was 0.93, TLI was 0.99, and RMSEA was 0.04 (full results available on request from second author). Most important, the factor loadings were minimally affected when including residual correlations. The fit statistics associated with the single factor model were good enough to be accepted by some standards used in analyses of categorical item response data (e.g. Bentler 1990; Brown and Cudeck 1993), though other work with factor analysis of continuous covariates suggests more stringent criteria (Hu and Bentler 1999). Loadings on the general factor for all of the items were very strong. Based on these analyses, we determined that the scale was sufficiently unidimensional to proceed with IRT analyses (Lai et al. 2006; Reeve et al. 2007). These findings are consistent with the design of the MFQ. The developers, Angold and colleagues (1995), found that long-form of the MFQ had a large single factor with moderate to high loadings on most items, while other factors were much smaller and unstable when subject to rotational techniques.

IRT and DIF Analyses—We used the graded response model (Samejima 1969, 1997) to obtain unadjusted IRT item parameters for the MFQ. The "a" parameters are directly analogous to factor loadings, and the "b" parameters are analogous to category threshold parameters. We used Parscale (Muraki and Bock 2003) for these analyses and employed expectation a posteriori scoring (Details can be obtained from the second author). The item category characteristic curves for the MFQ were generated from a previously developed Microsoft Excel spreadsheet, which is available from Dan Mungas (Mungas et al. 2003).

The approach to DIF assessment combined ordinal logistic regression and IRT (Crane et al. 2004, 2006) to determine whether MFQ items functioned differently for children of different ethnic/racial groups, gender, grade level, and parent's birth place. We utilized a published hybrid ordinal logistic regression/IRT algorithm for DIF detection (Crane et al. 2006, 2007). We initially generated IRT scores using Parscale (Muraki and Bock 2003). For each item, we then examined three ordinal logistic regression models for each demographic category (labeled here as "group") selected for analysis.

$$\text{Logit p}(\mathrm{Y}=1|\theta, \mathit{group})=\beta_1 * \theta+\beta_2 * \text{group}+\beta_3 * \theta * \text{group} \quad \text{(model1)}$$

$$\text{Logit p}(\mathrm{Y}=1|\theta, \mathit{group})\beta_1 * \theta+\beta_2 * \text{group} \quad \text{(model2)}$$

$$\text{Logit p}(\mathrm{Y}=1|\theta)=\beta_1 * \theta \quad \text{(model3)}$$

The models shown are logistic regression models; Stata (StataCorp 2009) uses the proportional odds model for ordinal logistic regression. In these equations, p(Y=1) is the probability of endorsing an item, θ is the IRT estimate of depression level, and group is the demographic category (full methods available on request from second author). Non-uniform DIF represents an interaction between group membership and depression level on item endorsement. This relationship is captured with the β_3 term in model 1. For non-uniform DIF, we applied a statistical significance criterion with α=0.01 when comparing the likelihoods for models 1 and 2. With our sample size, α=0.01 is very sensitive. Uniform DIF represents an interference by group membership of the relationship between depression level and probability of item endorsement. This relationship is captured with the difference in the β_1 term (i.e., the coefficient associated with depression level) in model 2 (that includes the group term) and model 3 (that does not include the group term). Large differences in the β_1 coefficient between models 2 and 3 imply that group membership has a strong interference with the relationship between depression level and probability of endorsing an item. For uniform DIF, we used a change in β coefficient criterion of 1% or 5% for models 2 and 3. The change in β coefficient is more consistent than a p value in terms of stability with respect to sample size. A 1% difference is truly trivial and very sensitive; 5% is also small but somewhat less sensitive than 1%. These sensitive criteria for uniform and non-uniform DIF were selected so as to err on the side of over-identification of items with DIF, which ensures thorough evaluation of items for measurement non-invariance. In this setting, a crucial issue is whether items with DIF detected using these sensitive criteria are associated with clinically relevant DIF impact (see below).

We determined the presence and extent of DIF for each covariate considered separately, beginning in each case with the unadjusted estimate of depression level, θ. These analyses included data from 3,593 participants. In analyzing DIF with respect to race/ethnicity, we compared Asian, African American, and Hispanic participants to white participants (i.e., we treated the race/ethnicity covariate as three separate dichotomous comparisons). For these analyses, we kept the criterion for non-uniform DIF consistent, with α=0.01 as the critical value when comparing the likelihoods for models 1 and 2.

We also determined DIF for all covariates simultaneously, resulting in a final depression level score that accounted for DIF with respect to all of the covariates we analyzed. We will refer to this as “multiple sources of DIF.” The covariates we considered for both sets of analyses were race/ethnicity, sex, grade level, and parent(s)’ birth place.

We have found similar DIF impact when using different threshold values for uniform DIF detection (Crane et al. 2007). For the single covariate at a time analyses, we used a very sensitive 1% change in β coefficient criterion to detect even miniscule amounts of DIF. When we evaluated all of the covariates, however, for items detected to have DIF, the algorithm we used calls for stratifying item responses for covariates with DIF when analyzing subsequent covariates. This leads to smaller effective sample sizes for items identified with DIF with respect to multiple covariates. Since DIF impact is similar at a 1% and a 5% change threshold (Crane et al. 2007), we chose the slightly less sensitive but still feasible 5% change in β coefficient threshold for uniform DIF when we evaluated all of the covariates for DIF. (Detailed methods for obtaining IRT scores that account for multiple sources of DIF can be obtained from the second author).

An important consideration often overlooked in the DIF literature is that individuals are members of many groups. Thus a single participant in our study could be in the Asian-American group, the male group, the 6th grade group, and the US-born parents group, all at once. When covariates are not matched across groups, it is important to account for DIF with respect to the unmatched covariates to ensure that DIF effects consider all sources of variability (Gibbons et al. 2011). In the present case, for example, there were grade differences across ethnic groups. We thus assessed DIF with respect to both grade and ethnic groups to ensure that our final estimates of DIF considered all sources of variation for which data were available.

We then subtracted the unadjusted IRT score from the final IRT score that accounted for multiple sources of DIF. The unadjusted IRT scores represented the scores not accounting for DIF with respect to any of the four covariates. The final IRT scores represented scores that accounted for DIF with respect to all four covariates considered. Thus any difference between these scores is entirely due to DIF. The difference between the scores for any individual provides an estimate of DIF impact for that person.

A box-and-whisker plot was used to show the distribution of individual-level DIF impact, indexed against the median standard error of measurement as an indicator of salient individual-level DIF impact, as we have done previously (Crane et al. 2006). There is little guidance in the literature on how to judge the magnitude of individual-level DIF impact. The rationale for choosing the median standard error of measurement is that every score is characterized by measurement error, which can be quantified with the standard error of measurement. The median standard error of measurement represents the amount of measurement error for the middle of the distribution of measurement errors observed in the sample. Differences in score smaller than the standard error of measurement are not reliably different. Another way of thinking about this is that investigators use these scores and tolerate the amount of measurement imprecision represented by the standard error of measurement. DIF impacts observed to be larger than the median measurement error would suggest that the signal related to DIF was greater than the noise related to measurement error.

We performed similar analyses of the difference between unadjusted IRT scores and IRT scores accounting for multiple sources of DIF for demographic subgroups to show the distribution of group-level DIF impact, indexed against the standard deviation of the larger group. For example, for race/ethnicity, we compared group-level impacts to the standard deviation for whites to provide an estimate of the effect size associated with DIF. As in individual-level DIF impact, there is little guidance from the literature on how to judge the magnitude of group-level DIF impact. We suggest indexing these impacts to the standard deviation of the larger group, based on considerations similar to those that suggest the use of z scores or effect sizes. Group-level DIF impacts larger than the standard deviation for the larger group suggest that DIF may distort comparisons of means. DIF impacts observed to be twice the standard deviation of the larger group suggest that statistical difference between groups may be caused by DIF. Finally, we determined mean depression level scores across subgroups for unadjusted IRT scores and for IRT scores that accounted for multiple sources of DIF.

Results

Table 1 summarizes demographic characteristics of the study population stratified by racial/ ethnic group. In general, the sex distribution was similar across all racial/ ethnic groups. Higher proportions of 8th graders than 6th graders were white or Hispanic. Most participants whose parents were not born in the US were Asian. Complete item responses for the MFQ were available for 3,593 participants. Participants who were excluded from the final analyses were similar in demographic characteristics to those whose data are included in this study.

Prevalence of Depression Across Racial/Ethnic Groups

Using Wood and colleagues' (1995) child-reported cut-off score of 27 or greater, 6.5% (N=261) of students would likely meet criteria for depression. The prevalence of depression was 6.5% among whites, 8.4% among Asians, 9.1% among African Americans, and 6.8% among His-panics. These proportions were not statistically different ($\chi_{df=3}{}^2$=5.73, p=0.13).

Differential Item Functioning Presence

For MFQ scores, we found DIF presence for several items with respect to gender, grade level, parent's birth place, and race/ ethnicity (i.e., Asian American vs. white, African American vs. white, and Hispanic vs. white). Item-level findings for DIF with respect to gender, grade, and parent's birth place analyzed separately are summarized in Table 2 and race/ethnicity in Table 3. We found that 14 items had DIF with respect to gender (Table 2) using very sensitive DIF thresholds. Uniform DIF items related to physical complaints (e.g., "I ate more than usual", "I felt so tired I sat around and did nothing", "I was very restless"), anhedonia (e.g., "I didn't enjoy anything at all", "I didn't have any fun at school"), and low self-esteem (e.g., "I felt I was a bad person," "I did everything wrong"). Males were more likely to endorse these items. Two items also had non-uniform DIF presence. They related to anhedo-nia and the belief that the family was better off without the person. Eleven items had DIF with respect to grade level. These items related mostly to physical/somatic complaints and low self-esteem. For any given level of depression, sixth graders were more likely to endorse these items.

Twenty items had DIF with respect to parent's birth place. The majority of items with uniform DIF related to physical/somatic complaints, low self-esteem, social withdrawal, irritability, and anhedonia. For any given level of depression, most of these items were more likely to be endorsed by students whose parents were US-born. Three items also had non-uniform DIF. They related to decreased appetite, feeling less talkative, and doing everything wrong.

We also found DIF presence with respect to racial/ethnic groups (Table 3). Comparing Asian American and white students, we found that 16 items had DIF. The items with the greatest magnitude of DIF related to irritability/parental conflicts (change in β coefficient=14%), tearfulness (change in β coefficient=5%), anhedonia (change in β coefficient=4%), and concentration difficulty (change in β coefficient=3%). For 10 of the 16 items with DIF, at a given depression level, a lower proportion of Asian American students than white students endorsed the item, while for the other 6 items, at a given depression level, a higher proportion of Asian American students than white items endorsed the item.

We provide item category characteristic curves (ICCC) across demographic groups in Fig. 1. For example, Fig. 1a showed the ICCC for the item "I felt grumpy and cross with my parents" for whites (black curves) and Asian-Americans (gray curves). The solid curves to the left (one in black and one in gray) showed the probability of endorsing "sometimes" rather than "not true" associated with each depression level, while the dashed curves to the right (one in black and one in gray) showed the probability of endorsing "true" rather than "sometimes" associated with each depression level. The black curves were consistently above the gray curves, which meant that for a given depression level, whites were more likely to endorse this item than Asian-Americans. This pattern illustrates uniform DIF. Four items also had non-uniform DIF. These items related to irritability (also illustrated in Fig. 1a), feeling less talkative, hating oneself, and doing everything wrong.

Sixteen items had uniform DIF with respect to African American vs. white race/ethnicity. For twelve of these, white students were more likely than African Americans to endorse the item. The greatest magnitude of uniform DIF was seen for irritability (change in β coefficient=10%), poor body image (change in β coefficient=8%), and loneliness (change in β coefficient=5%). Four items had uniform DIF in the opposite direction, such that African American students were more likely to endorse these items than white students at a given depression level. These items were "I ate more than usual", "I was talking more slowly than usual", "I thought bad things would happen to me", and "I slept a lot more than usual."

When comparing Hispanic and white students, items found to have uniform DIF included those addressing irritability, somatic complaints, anhedonia, concentration difficulty, and depressed mood. Of these, irritability had the greatest magnitude of uniform DIF with change in the β coefficient of 3%. For most of these items, white students were more likely to endorse the items for a given depression level than Hispanics.

Individual- and Group-Level DIF Impact for Each Covariate and for All Covariates

Overall, there was negligible individual- and group-level DIF impact for each covariate and for all covariates. The top part of Fig. 2 shows individual-level DIF impact for each of the covariates evaluated. Each box plot shows the distribution of the difference between unadjusted IRT scores and IRT scores that account for DIF related to a single covariate. A value of 0 means there was no difference between unadjusted scores and scores that account for DIF—that is, there was no DIF impact. Positive values indicate that accounting for DIF results in higher scores, or equivalently, that ignoring DIF results in underestimating the depression level. Likewise, negative values indicate that accounting for DIF results in lower scores, or, equivalently, that ignoring DIF results in overestimating the depression level. Vertical reference lines are placed at the median standard error of measurement, which served as our guideline for salient DIF impact as discussed in the Methods section.

In this study, the items with DIF related to gender, grade level, parents' birth place, and Asian American, African American, and Hispanic vs. white students were all associated with negligible DIF impact, as indicated by the entire distribution of differences smaller than the median standard error of measurement. We performed additional analyses to account for all sources of DIF simultaneously, as shown in the bottom of Fig. 2. None of the participants had scores that differed by as much as the median standard error of measurement.

We also performed analyses to determine group-level DIF impacts across demographic subgroups (Fig. 3). We found negligible group-level DIF impact when accounting for all of the sources of DIF considered here. Figure 3 illustrates group-level DIF impact between Asian American and white students. None of the participants had IRT scores that differed by as much as one standard deviation of the unadjusted IRT score of whites, the larger group. Differences between groups in mean depression levels were negligibly different when using unadjusted IRT scores and IRT scores accounting for multiple sources of DIF. Compared with the mean IRT scores for whites, differences in the means were negligibly impacted by DIF for Asian Americans (4% of the SD for whites), Hispanics (3%), and African Americans (2%). Similarly, differences in mean depression levels between students with non-US born parents and students with US-born parents only differed by 2% of the SD for those with US-born parents (where the differences between the unadjusted and adjusted IRT scores between the two groups was only 0.02 and the adjusted SD was 1.00); differences in mean depression levels between female and male students differed by 1% of the SD for males; and differences in mean depression scores between 8th graders and 6th graders differed by 0.1% of the SD for 6th graders. Thus, for the covariates considered here, accounting for DIF made a negligible difference in mean scores for demographic subgroups, making DIF unlikely to be the cause of differences in depression levels seen across these groups.

Discussion

Several items had DIF with respect to gender, grade level, parent's birth place, and racial/ethnic groups. Most of these items addressed physical/somatic complaints, low self-esteem, anhedonia, and irritability. Accounting for all of the sources of DIF considered here, differences in scores across groups were unlikely to be related to measurement non-equivalence. These findings suggest that the MFQ is a valid measure to use when comparing total or mean scores across diverse racial/ethnic populations, and that the MFQ provides estimates of depression severity that are appropriate for comparisons across groups.

Individual and Group-Level Impact Implications

In this study, findings support the measurement equivalence of the MFQ across groups. Overall, individual-level DIF impact was negligible. When we accounted for multiple sources of DIF, individual-level DIF impact was smaller than the median standard error of measurement (SEM). These findings suggest that a youth's individual score on the MFQ is not biased by measurement invariance, and this instrument may be suitable for screening diverse child and adolescent patient populations in clinical or school settings. DIF also did not lead to exaggerated or attenuated group differences across gender, grade level, parents' birth place, and racial/ethnic groups. The greatest difference was observed among Asian Americans vs. whites, where accounting for DIF affected the differences in mean scores for the two groups by only 4% of a SD for whites. These findings suggest that the MFQ demonstrates sufficient measurement equivalence for 6th or 8th graders, males and females, youth with US- and non-US born parents, and Asian American, African American, Hispanic American, and white youth. Researchers can have greater confidence that when using the MFQ, observed differences in overall depression scores across groups can be attributable to true differences rather than to DIF.

Cultural and Developmental Considerations

In this study, the prevalence of MFQ scores consistent with depression in a community sample of 6th and 8th graders was 7%. The prevalence varied between 7% and 9% across racial/ethnic groups with white students having the lowest rates. These estimates are comparable to findings of Roberts and colleagues (1997) who surveyed a sample of 5400, 6th–8th grade students using the DSM Scale for Depression. Roberts and colleagues (1997) found that the prevalence of depression was 6.3% for white youth and 9% for African American youth. Whereas Roberts et al. (1997) found that Hispanic students reported the highest prevalence of depression at 12% for Mexican American youth, we found that African American students reported the highest prevalence (9%). These findings lend support to a growing literature that documents higher depression among ethnic minority children (Brown et al. 2007; Siegel et al. 1998).

Overall, although items with DIF did not significantly affect the total MFQ scores for individuals or groups, our findings suggest patterns of differential symptom expression that may warrant further investigation. We found that for any level of depression, being male, Asian American, African American, and having non-US born parents were the characteristics associated with being more likely to endorse symptoms related to physical complaints (e.g., eating more than usual, talking more slowly, and sleeping more than usual). Several studies have suggested that Asian American youth may be more likely to somaticize psychological problems (Chen et al. 1998) due to cultural stigmas associated with mental illness. Physical complaints may be more likely to be reported as they are more socially accepted (Greenberger and Chen 1996). Choi and Gi Park (2006) also suggested that expression of physical complaints may be a more culturally acceptable method to express depression especially among Hispanic American boys who subscribe to the cultural construct of machismo, where these boys may perceive that the expression of internalizing symptoms would be viewed as a weakness whereas physical complaints would be acceptable. Therefore, from a clinical perspective, attention to reports of physical complaints may be especially important in treating these children as they may be indicative of depression. Subsequent qualitative work by scale developers could consider these findings in diverse groups of youth (Ercikan et al. 2010).

Our finding that there was negligible DIF impact may be, in part, a function of the low prevalence of depression found in early adolescence. While this study had a very large overall sample size, the prevalence of severe levels of depression was low, mitigating somewhat our power to address DIF among children with severe depressive levels. Studies have documented increasing prevalence of depression beginning in 9th grade (Birmaher et al. 1996; Lewinsohn et al. 1994). In adolescence, neurological development, physical growth, and sexual maturation contribute to increase risk of depression through the intensification of emotions and the incomplete development of cognitive and emotional coping skills needed to handle such strong emotions. Therefore, results might be different when DIF analyses are applied to older adolescents, who are likely to demonstrate a higher prevalence of depression (Lewinsohn et al. 1994).

Several limitations should be considered when interpreting our findings. Although we used fairly sensitive criteria for DIF detection, there is no universal agreement on DIF detection techniques. Different methods may yield different results, though most approaches (including the hybrid ordinal logistic regression/IRT approach used here) have produced similar findings as other approaches when applied to the same datasets (Millsap 2006). Furthermore, although this study had a substantial number of racial/ethnic minority participants, the sample was not large enough to assess intra-racial group differences. Each of the broad racial/ ethnic groupings within our sample was comprised of youth from many distinct nationalities and ethnic subgroups. For example within the Asian American group, ethnic groups included Chinese, Japanese, Korean, Vietnamese, Other Southeast Asian, and East Indian. Therefore, study inferences apply only across broad racial groupings. We also did not include individuals who endorsed mixed racial status, as we had small numbers of participants with any particular combination, which would not allow us to make any meaningful inferences about race/ethnicity based on this multi-ethnic group. Other richer data sets will be needed to capture those effects. Nevertheless, we suspect that the lack of important DIF impact despite our use of very sensitive DIF detection criteria makes large DIF effects from mixed racial groups somewhat unlikely. Furthermore, SES and acculturation data were not available for analyses. Therefore, we were not able to investigate the impact of socioeconomic status or other cultural factors, such as acculturation level, which could shed additional light on current findings.

In conclusion, we found that the MFQ appears to meet the goal of measuring depression equivalently for males and females, 6th and 8th graders, students of US- and non-US born parents, and Asian Americans, African Americans, Hispanic Americans, and whites. We found minimal individual- and group-level DIF impact, producing small and clinically insignificant differences in overall depression scores. Overall, our findings extend the current literature by suggesting that although there may be differences in symptom endorsements among depressed individuals across racial/ethnic groups, these differences do not impact overall scores (i.e. lead to a change in total depression score on the MFQ). Similar methods should be used with other samples to test whether the MFQ functions equivalently for racial/ethnic subgroups and with older adolescents.

References

Abe-Kim J, Takeuchi DT, Hong S, Zane N, Sue S, Spencer MS, et al. Use of mental health-related services among immigrant and US-born Asian Americans: results from the National Latino and Asian American Study. American Journal of Public Health. 2007; 97(1):91–98. [PubMed: 17138905]

Alegria M, Canino G, Shrout PE, Woo M, Duan N, Vila D, et al. Prevalence of mental illness in immigrant and non-immigrant U.S. Latino groups. The American Journal of Psychiatry. 2008; 165(3):359–369. [PubMed: 18245178]

Anderson ER, Mayes LC. Race/ethnicity and internalizing disorders in youth: a review. Clinical Psychology Review. 2010; 30(3):338–348. [PubMed: 20071063]

Angold A, Costello EJ, Messer SC, Pickles A. Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. International Journal of Methods in Psychiatric Research. 1995; 5(4):237–249.

Beauducel A, Herzberg PY. On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. Structural Equation Modeling. 2006; 13(2):186–203.

Bentler PM. Comparative fit indexes in structural models. Psychological Bulletin. 1990; 107:238–246. [PubMed: 2320703]

Birmaher B, Ryan ND, Williamson DE, Brent DA, Kaufman J, Dahl RE, et al. Childhood and adolescent depression: a review of the past 10 years. Part I. Journal of the American Academy of Child and Adolescent Psychiatry. 1996; 35(11):1427–1439. [PubMed: 8936909]

Bisaga K, Whitaker A, Davies M, Chuang S, Feldman J, Walsh B. Eating disorder and depressive symptoms in urban high school girls from different ethnic backgrounds. Journal of Developmental and Behavioral Pediatrics. 2005; 26(4):257–266. [PubMed: 16100498]

Breslau J, Chang DF. Psychiatric disorders among foreign-born and US-born Asian-Americans in a US national survey. Social Psychiatry and Psychiatric Epidemiology. 2006; 41(12):943–950. [PubMed: 16988789]

Brown, MW.; Cudeck, R. Alternative ways of assessing model fit. In: Bollen, KA.; Long, JS., editors. Testing structural equation models. Newbury Park: Sage; 1993. p. 445-455.

Brown JS, Meadows SO, Elder GH Jr. Race–ethnic inequality and psychological distress: depressive symptoms from adolescence to young adulthood. Developmental Psychology. 2007; 43:1295–1311. [PubMed: 18020812]

Camilli G. Origin of the scaling constant d = 1.7 in item response theory. Journal of Educational and Behavioral Statistics. 1994; 19(3):293–295.

Camilli, G.; Shepard, LA. Methods for identifying biased test items. Thousand Oaks: Sage; 1994.

Chen IR, Roberts RE, Aday LA. Ethnicity and adolescent depression: the case of Chinese Americans. The Journal of Nervous and Mental Disease. 1998; 186:623–630. [PubMed: 9788639]

Choi H, Gi Park C. Understanding adolescent depression in ethnocultural context: updated with empirical findings. ANS Advances in Nursing Science. 2006; 29(4):E1–12. [PubMed: 17135793]

Choi H, Meininger JC, Roberts RE. Ethnic differences in adolescents' mental distress, social stress, and resources. Adolescence. 2006; 41(162):263–283. [PubMed: 16981616]

Costello EJ, Angold A. Scales to assess child and adolescent depression: checklists, screens, and nets. Journal of the American Academy of Child and Adolescent Psychiatry. 1988; 27(6):726–737. [PubMed: 3058677]

Costello EJ, Edelbrock CS, Costello AJ. Validity of the NIMH diagnostic interview schedule for children: a comparison between psychiatric and pediatric referrals. Journal of Abnormal Child Psychology. 1985; 13(4):579–595. [PubMed: 4078188]

Crane PK, van Belle G, Larson EB. Test bias in a cognitive test: differential item functioning in the CASI. Statistics in Medicine. 2004; 23(2):241–256. [PubMed: 14716726]

Crane PK, Gibbons LE, Jolley L, van Belle G. Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. Medical Care. 2006; 44(11 Suppl 3):S115–123. [PubMed: 17060818]

Crane PK, Gibbons LE, Ocepek-Welikson K, Cook K, Cella D, Narasimhalu K, et al. A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. Quality of Life Research. 2007; 16(Suppl 1):69–84. [PubMed: 17554640]

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

Daviss WB, Birmaher B, Melhem NA, Axelson DA, Michaels SM, Brent DA. Criterion validity of the Mood and Feelings Questionnaire for depressive episodes in clinic and non-clinic subjects. J Child Psychol Psychiatry. 2006; 47(9):927–934. [PubMed: 16930387]

Drasgow F, Kanfer R. Equivalence of psychological measurement in heterogeneous populations. Journal of Applied Psychology. 1985; 70:662–680. [PubMed: 4086417]

Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah: L. Erlbaum Associates; 2000.

Ercikan K, Arim R, Law D, Domene J, Gagnon F, Lacroix S. Application of think aloud protocols for examining and confirming sources of differential item functioning identified by expert reviews. Educational Measurement: Issues and Practice. 2010; 29(2):24–35.

Garrison CZ, Waller JL, Cuffe SP, McKeown RE, Addy CL, Jackson KL. Incidence of major depressive disorder and dysthymia in young adolescents. J Am Acad Child Adolesc Psychiatry. 1997; 36(4):458–465. [PubMed: 9100419]

Gibbons LE, Crane PK, Mehta KM, Pedraza O, Tang Y, Manly JJ, et al. Multiple, correlated covariates associated with differential item functioning (DIF): accounting for language DIF when education levels differ across languages. Aging Research. 2011; 3(e4):19–25.

Greenberger E, Chen C. Perceived family relationships and depressed mood in early and late adolescence: a comparison of European and Asian Americans. Developmental Psychology. 1996; 32:707–716.

Harker K. Immigrant generation, assimilation, and adolescent psychological well-being. Social Forces. 2001; 79(3):969–1004.

Holland, PW.; Wainer, H. Differential item functioning. Hillsdale: Erlbaum; 1993.

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

Lai JS, Crane PK, Cella D. Factor analysis techniques for assessing sufficient unidimensionality of cancer related fatigue. Quality of Life Research. 2006; 15(7):1179–1190. [PubMed: 17001438]

Lewinsohn PM, Clarke GN, Seeley JR, Rohde P. Major depression in community adolescents: age at onset, episode duration, and time to recurrence. Journal of the American Academy of Child and Adolescent Psychiatry. 1994; 33(6):809–818. [PubMed: 7598758]

Messer SC, Angold A, Costello EJ, Loeber R. Development of a short questionnaire for use in epidemio-logical studies of depression in children and adolescents: Factor composition and structure across development. International Journal of Methods in Psychiatric Research. 1995; 5(4):251–262.

Millsap RE. Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. Medical Care. 2006; 44(11 Suppl 3):S171–175. [PubMed: 17060824]

Millsap RE, Everson HT. Methodology review: statistical approaches for assessing measurement bias. Applied Psychological Measurement. 1993; 17(4):297–334.

Mungas D, Reed BR, Kramer JH. Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. Neuropsychology. 2003; 17(3):380–392. [PubMed: 12959504]

Muraki, E.; Bock, D. PARSCALE for Windows (Version 4.1). Chicago: Scientific Software International; 2003.

Muthen, LK.; Muthen, BO. Mplus user's guide. 3. LA: Muthen & Muthen; 1998–2004.

Muthen, B.; du Toit, SHC.; Spisic, D. Robust inference using weighted least squared and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Psychometrika. 1997. Accepted for publication. Available from:www.StatModel.com

Perreira KM, Deeb-Sossa N, Harris KM, Bollen K. What are we measuring? An evaluation of the CES-D across race/ethnicity and immigrant generation. Social Forces. 2005; 83(4):1567–1602.

Reeve BB, Hays RD, Bjorner JB, Cook KF, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks: plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care. 2007; 45(5 Suppl 1):S22–31. [PubMed: 17443115]

Roberts RE, Sobhan M. Symptoms of depression in adolescence: a comparison of Anglo, African, and Hispanic Americans. Journal of Youth and Adolescence. 1992; 21(6):639–651.

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

Saluja G, Iachan R, Scheidt PC, Overpeck MD, Sun W, Giedd JN. Prevalence of and risk factors for depressive symptoms among young adolescents. Archives of Pediatrics & Adolescent Medicine. 2004; 158(8):760–765. [PubMed: 15289248]

Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement. 1969; 34(4 Part2 No 17)

Samejima, F. Graded response model. In: van der Linden, WJ.; Hambleton, RK., editors. Handbook of modern item response theory. NY: Springer; 1997. p. 85-100.

Sharp C, Goodyer IM, Croudace TJ. The Short Mood and Feelings Questionnaire (SMFQ): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. J Abnorm Child Psychol. 2006; 34(3):379–391. [PubMed: 16649000]

Siegel JM, Aneshensel CS, Taub B, Cantwell DP, Driscoll AK. Adolescent depressed mood in a multiethnic sample. Journal of Youth and Adolescence. 1998; 27:413–427.

StataCorp. Stata release 10. TX: College Station; 2009.

Takeuchi DT, Alegria M, Jackson JS, Williams DR. Immigration and mental health: diverse findings in Asian, black, and Latino populations. American Journal of Public Health. 2007a; 97(1):11–12. [PubMed: 17138903]

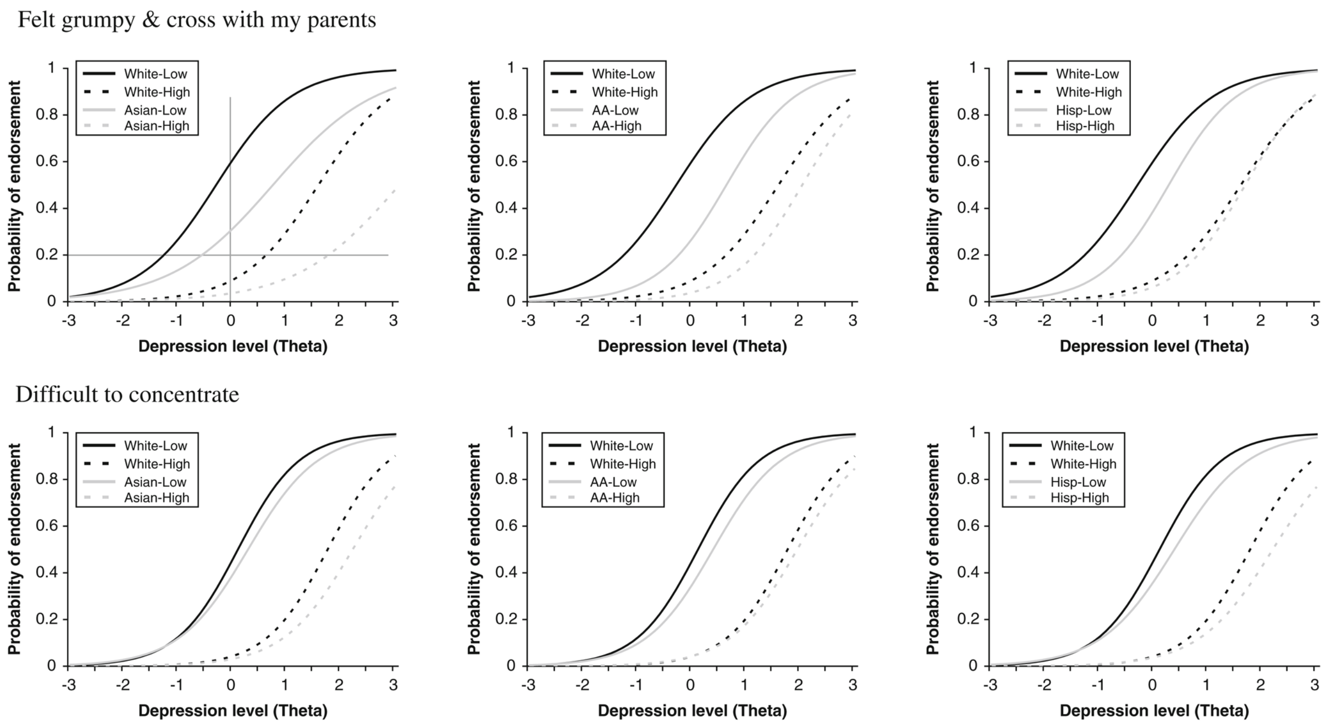
Takeuchi DT, Hong S, Gile K, Alegria M. Developmental contexts and mental disorders among Asian Americans. Research in Human Development. 2007b; 4(1 & 2):49. [PubMed: 20150976]

Takeuchi DT, Zane N, Hong S, Chae DH, Gong F, Gee GC, et al. Immigration-related factors and mental disorders among Asian Americans. American Journal of Public Health. 2007c; 97(1):84–90. [PubMed: 17138908]

Teresi JA, Ocepek-Welikson K, Kleinman M, Eimicke JP, Crane PK, Jones RN, et al. Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. Psychology Science Quarterly. 2009; 51(2):148–180. [PubMed: 20336180]

Thapar A, McGuffin P. Validity of the shortened Mood and Feelings Questionnaire in a community sample of children and adolescents: a preliminary research note. Psychiatry Research. 1998; 81(2):259–268. [PubMed: 9858042]

Wood A, Kroll L, Moore A, Harrington R. Properties of the mood and feelings questionnaire in adolescent psychiatric outpatients: a research note. J Child Psychol Psychiatry. 1995; 36(2):327–334. [PubMed: 7759594]



Felt grumpy & cross with my parents

Difficult to concentrate

Fig. 1.

Item category characteristic curves for two items that have DIF presence across racial/ethnic groups. The y-axis represents the probability of symptom endorsement and the x-axis represents the theta, θ, level. The solid black curves represent the item difficulty curves for whites at “not true” to “sometimes” theta level. The dashed black curves represent the item difficulty curve for whites at “sometimes” to “true” theta level. The solid gray curves represent the item category characteristic curves (ICCC) for Asian Americans, African Americans and Hispanic Americans at “not true” to “sometimes” theta level. The dashed gray curves represent the ICCCs for Asian Americans, African Americans and Hispanic Americans at “sometimes” to “true” theta level. In panel **A**, *top left*, ICCCs show a significant DIF presence with a change in β coefficient of 14% when comparing Asian Americans to whites for the item, “Felt grumpy and cross with my parents.” The vertical straight gray reference line illustrates that at 0 theta, there is 60% probability that whites endorsed the items versus 31% of Asian Americans. At 20% probability of symptom endorsement when endorsing “sometimes” or “true” on the item, theta was 0.6 for whites but 1.8 for Asian Americans (cf. horizontal gray line). The middle top panel shows significant DIF presence with change in β coefficient of 10% when comparing African Americans to whites. The right top panel illustrates negligible DIF presence change in β coefficient of 3% when comparing Hispanic Americans to whites. All three bottom panels illustrate relatively negligible DIF presence for the item, “Found it difficult to think properly or concentrate.” **A** White vs. Asian American youth. **B** White vs. African American youth. **C** White vs. Hispanic youth

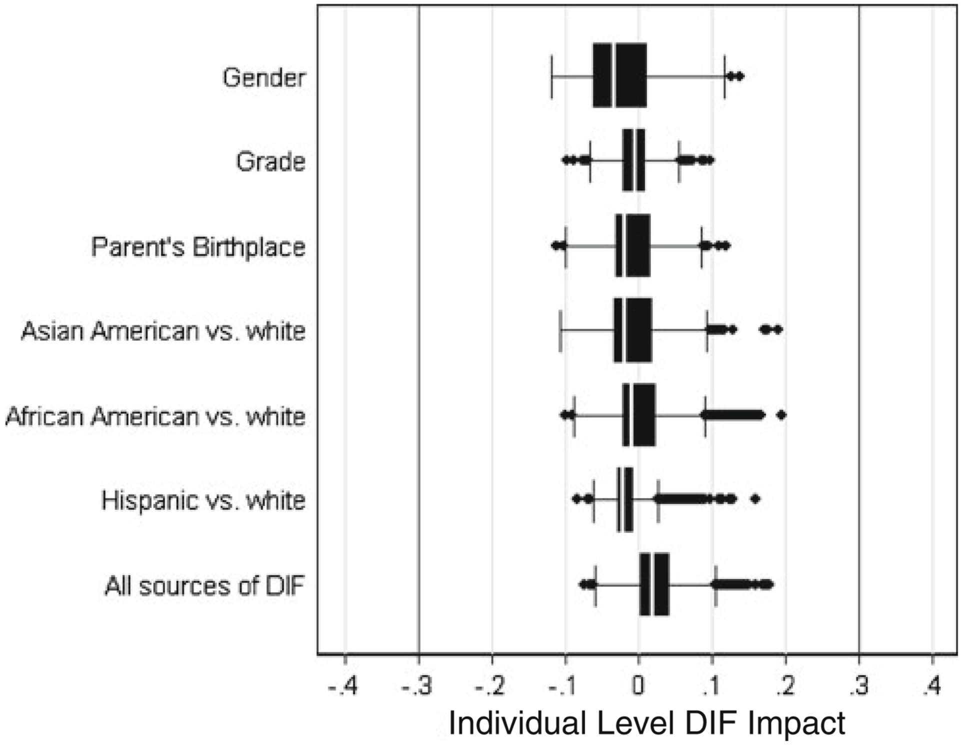


Fig. 2.

Individual-level DIF impact for each covariate and for all covariates. The first six box-and-whisker plots delineate individual-level DIF impact associated with each of the six covariates evaluated in turn, while the last plot delineates individual-level DIF impact associated with all the covariates considered here. The values plotted are the differences between the unadjusted IRT score and IRT scores that accounted for DIF associated with each covariate (first six plots) or with multiple covariates (last plot). A difference of 0 (the middle reference line) would mean that DIF made no difference for that person. Large positive values indicate that scores accounting for DIF were higher than scores that ignored DIF, which means that ignoring DIF resulted in underestimates of depression severity. Large negative values indicate that scores accounting for DIF were lower than scores that ignored DIF; thus ignoring DIF resulted in overestimates of depression severity. These box-and-whisker plots are indexed by 1x the median standard error of measurement (SEM) of the MFQ among these participants. Observations outside of ±.3 SEM indicate that a covariate has salient individual-level DIF impact (first six plots) or that the covariates evaluated for multiple sources of DIF considered together have salient individual-level DIF impact (last plot)

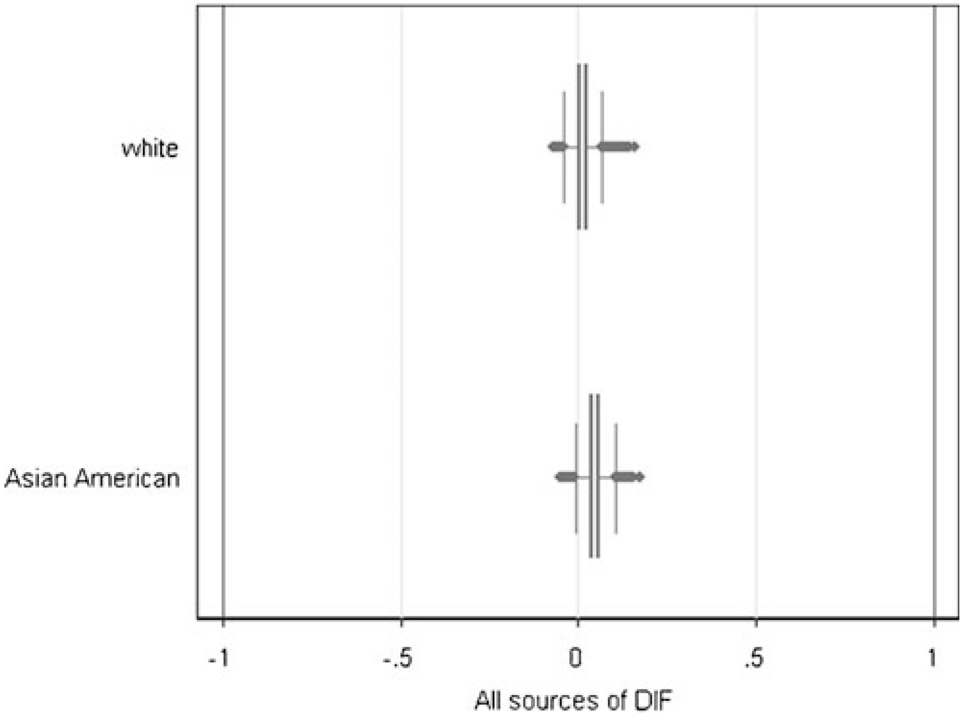


Fig. 3.

Group-level DIF impact for multiple covariates presented separately by Asian American versus white subgroup. Group-level DIF impact for multiple covariates presented separately by Asian American versus white subgroup. Difference scores were obtained as described in the note to Fig. 2; these are plotted separately for subgroups as indicated in the figure. Vertical reference lines are drawn at 1 standard deviation of the unadjusted IRT score for the largest subgroup. For example, for Asian American vs. white, the standard deviation for whites was 1.0, so vertical lines are drawn in Fig. 3 at 1.0 and −1.0

Table 1

Demographic of study participants stratified by race/ ethnicity

Characteristic	**White**		**Asian**		**African-American**		**Hispanic**	
***N*=3593**	***N*=2,090**	**%**	***N*=717**	**%**	***N*=505**	**%**	***N*=281**	**%**
Gender								
Male (*N*=1,774)	1,049	50	350	49	240	48	135	48
Female (*N*=1,819)	1,041	50	367	51	265	53	146	52
Grade								
6th grade (*N*=1,502)	806	39	330	46	261	52	105	37
8th grade (*N*=2,091)	1,284	61	387	54	244	48	176	63
Parent’s birth place								
US born (*N*=2,537)[a]	2,013	96	78	11	319	63	127	45
Non-US born (*N*=1,056)[b]	77	4	639	89	186	37	154	55

[a] At least one parent was born in the United States

[b] Both parents were born outside the United States

Of note: 517 participants were excluded due to having missing data on parent birth’s place

Table 2

Item-level DIF presence

MFQ Items	Gender			Grade			Parent’s Birth Place		
	NU	U	D^a	NU	U	D^a	NU	U	D^a
1. Felt miserable or unhappy	0.36	0.01		0.51	0.01		0.68	**0.03**	US
2. Didn’t enjoy anything at all	**0.01**	**0.03**	M	0.82	0.01		0.03	0.01	
3. Less hungry than usual	0.21	0.01		0.87	**0.03**	6th	**0.01**	0.01	
4. Ate more than usual	0.01	**0.06**	M	0.53	**0.02**	6th	0.52	**0.07**	NUS
5. So tired I sat around & did nothing	0.29	**0.01**	M	0.54	0.01		0.32	**0.01**	US
6. Moved & walked more slowly than usual	0.97	**0.02**	M	0.63	0.01		1	**0.01**	US
7. Was very restless	0.81	**0.02**	M	0.57	0.01		0.33	**0.01**	US
8. Felt I was no good anymore	0.38	0.01		0.34	0.01		0.57	0.01	
9. Blamed myself for things that weren’t my fault	0.11	0.01		0.08	0.01		0.14	0.01	
10. Hard for me to make up my mind	0.32	0.01		0.02	0.01		0.15	0.01	
11. Felt grumpy & gross w/my parents	0.51	**0.01**	F	0.5	0.01		0.28	**0.13**	US
12. Felt like talking less than usual	0.11	**0.03**	M	0.17	0.01		**0.01**	0.01	
13. Was talking more slowly than usual	0.89	**0.05**	M	0.14	**0.02**	6th	0.66	**0.01**	NUS
14. Cried a lot	0.23	0.01		0.12	**0.01**	6th	0.06	**0.03**	US
15. Thought there was nothing good for me in the future	0.04	**0.02**	M	0.57	0.01		0.53	0.01	
16. Thought my family was better of without me	**0.01**	0.01		0.16	**0.01**	6th	0.24	0.01	
17. Didn’t want to see my friends	0.38	0.01		0.39	0.01		0.98	**0.02**	US
18. Found it difficult to think properly or concentrate	0.58	0.01		0.68	0.01		0.08	**0.04**	US
19. Thought bad things would happen	0.16	0.02		0.08	**0.04**	6th	0.44	**0.01**	NUS
20. I hated myself	0.47	0.01		0.65	**0.01**	6th	0.03	0.01	
21. Felt that I was a bad person	0.48	**0.01**	M	0.42	0.01		0.33	0.01	
22. Thought I looked ugly	0.14	0.01		0.62	0.01		0.24	**0.02**	US
23. Worried about aches and pains	0.11	0.01		0.31	**0.02**	6th	0.83	**0.02**	US
24. Felt lonely	0.69	0.01		0.19	0.01		0.13	**0.01**	US
25. Thought nobody really loved me	0.45	0.01		0.17	0.01		0.64	0.01	
26. Didn’t have any fun at school	0.54	**0.06**	M	0.01	**0.01**	8th	0.63	**0.04**	US
27. Thought I would never be as good as other kids	0.06	0.01		0.84	0.01		0.33	0.01	

MFQ Items	Gender			Grade			Parent’s Birth Place		
	NU	**U**	**D[a]**	**NU**	**U**	**D[a]**	**NU**	**U**	**D[a]**
28. Did everything wrong	0.02	**0.01**	M	0.6	**0.01**	6th	**0.01**	0.01	
29. Didn’t sleep as well as I usually sleep	0.05	0.01		0.22	0.01		0.91	**0.02**	US
30. Slept a lot more than usual	0.83	**0.06**	M	0.63	**0.03**	6th	0.46	**0.01**	NUS

NU = non-uniform, U = uniform. Numbers in the NU columns represent p values for interaction terms; values <0.01 are in bold font. Numbers in the U columns represent the percent change in β coefficients from models that excluded and included the group terms; values >0.01 are in bold font. In the U column, there are some values of 0.01 that are in bold font and other values of 0.01 that are not in bold font. The items not in bold font had values between 0.005 and 0.009999 and were thus rounded up to 0.01; the items in bold font had values from 0.010 to 0.0149999, and thus are represented in bold font

[a] In the “D” column, we indicate the direction of uniform DIF. For sex, an M indicates that for a given level of depression, males are more likely to endorse the item, that is, the ICCCs for males are to the left of the ICCCs for females. Similarly, an F indicates that for a given level of depression, females are more likely to endorse the item, that is, the ICCCs for females are to the left of the ICCCs for males. 6th = sixth grade, 8th = eighth grade; US = US-born, NUS = Non-US born

Table 3

Item-level DIF presence by racial/ethnic groups

MFQ Items	Asian vs. white			African American vs. white			Hispanic vs. white		
	NU	**U**	**D^{a}**	**NU**	**U**	**D^{a}**	**NU**	**U**	**D^{a}**
1. Felt miserable or unhappy	0.57	**0.02**	W	0.69	**0.03**	W	0.81	**0.02**	W
2. Didn’t enjoy anything at all	0.19	0.01		0.45	0.01		0.72	0.01	
3. Less hungry than usual	0.08	0.01		0.32	**0.02**	W	0.88	0.01	
4. Ate more than usual	0.98	**0.10**	A	0.92	**0.06**	AA	0.09	0.01	
5. So tired I sat around & did nothing	0.58	**0.02**	W	0.8	**0.01**	W	0.61	0.01	
6. Moved & walked more slowly than usual	0.68	0.01		0.87	0.01		0.25	0.01	
7. Was very restless	0.03	0.01		0.06	0.01		0.03	0.01	
8. Felt I was no good anymore	0.93	0.01		0.19	**0.02**	W	0.59	0.01	
9. Blamed myself for things that weren’t my fault	0.25	0.01		0.51	0.01		0.17	0.01	
10. Hard for me to make up my mind	0.64	0.01		0.72	0.01		0.08	0.01	
11. Felt grumpy & gross w/my parents	**0.01**	***0.14***	W	0.05	***0.10***	W	0.06	***0.03***	W
12. Felt like talking less than usual	**0.01**	0		0.03	0.01		0.08	0.01	
13. Was talking more slowly than usual	0.34	**0.01**	A	0.97	**0.01**	AA	0.04	0.01	
14. Cried a lot	0.04	***0.05***	W	0.76	**0.03**	W	0.52	0.01	
15. Thought there was nothing good for me in the future	0.03	0.01		0.13	0.01		0.98	0.01	
16. Thought my family was better of without me	0.71	0.01		0.67	0.01		0.72	0.01	
17. Didn’t want to see my friends	0.93	**0.01**	W	0.02	0.01		0.02	0.01	
18. Found it difficult to think properly or concentrate	0.06	***0.03***	W	0.86	**0.03**	AA	0.13	**0.02**	W
19. Thought bad things would happen	0.87	0.01		0.85	**0.01**	W	0.46	0.01	
20. I hated myself	**0.01**	0.01		0.67	**0.02**	W	0.53	0.01	
21. Felt that I was a bad person	0.41	0.01		0.26	0.01		0.04	0.01	
22. Thought I looked ugly	0.07	**0.02**	W	0.56	**0.08**	W	0.04	0.01	
23. Worried about aches and pains	0.15	0.01		0.1	0.01		0.51	**0.02**	W
24. Felt lonely	0.02	**0.02**	W	0.17	**0.05**	W	0.86	0.01	
25. Thought nobody really loved me	0.34	0.01		0.48	0.01		0.68	0.01	
26. Didn’t have any fun at school	0.66	***0.04***	W	0.31	**0.02**	W	0.73	**0.02**	W
27. Thought I would never be as good as other kids	0.04	0.01		0.41	**0.02**	W	0.38	0.01	

MFQ Items	Asian vs. white			African American vs. white			Hispanic vs. white		
	NU	**U**	**D[a]**	**NU**	**U**	**D[a]**	**NU**	**U**	**D[a]**
28. Did everything wrong	**0.01**	0.01		0.23	0.01		0.25	0.01	
29. Didn’t sleep as well as I usually sleep	0.15	**0.02**		0.04	0.01		0.03	**0.01**	W
30. Slept a lot more than usual	0.42	**0.05**	A	0.39	**0.04**	AA	0.11	0.01	

NU = non-uniform, U = uniform. Numbers in the NU columns represent p values for interaction terms; values <0.01 are in bold font. Numbers in the U columns represent the percent change in β coefficients from models that excluded and included the group terms; values >0.01 are in bold font; greatest percent change in β coefficients are italicized. In the U column, there are some values of 0.01 that are in bold font and other values of 0.01 that are not in bold font. The items not in bold font had values between 0.005 and 0.009999 and were thus rounded up to 0.01; the items in bold font had values from 0.010 to 0.0149999, and thus are represented in bold font

[a] In the “D” column, we indicate the direction of uniform DIF. For Asian vs. white, an A indicates that for a given level of depression, Asians are more likely to endorse the item, that is, the ICCCs for Asians are to the left of the ICCCs for whites. Similarly, a W indicates that for a given level of depression, whites are more likely to endorse the item, that is, the ICCCs for whites are to the left of the ICCCs for Asian Americans. AA = African American; HA = Hispanic American