Psychological Test and Assessment Modeling, Volume 58, 2016 (1), 141-181

Psychometric properties and performance of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Depression short forms in ethnically diverse groups

Jeanne A. Teresi[1,2,3,4], *Katja Ocepek-Welikson*[3], *Marjorie Kleinman*[2], *Mildred Ramirez*[3,4] & *Giyeon Kim*[5]

Abstract

Short form measures from the Patient Reported Outcomes Measurement Information System® (PROMIS®) are used widely. The present study was among the first to examine differential item functioning (DIF) in the PROMIS Depression short form scales in a sample of over 5000 racially/ethnically diverse patients with cancer. DIF analyses were conducted across different racial/ethnic, educational, age, gender and language groups.

Methods: DIF hypotheses, generated by content experts, informed the evaluation of the DIF analyses. The graded item response theory (IRT) model was used to evaluate the five-level ordinal items. The primary tests of DIF were Wald tests; sensitivity analyses were conducted using the IRT ordinal logistic regression procedure. Magnitude was evaluated using expected item score functions, and the non-compensatory differential item functioning (NCDIF) and T1 indexes, both based on group differences in the item curves. Aggregate impact was evaluated with expected scale score (test) response functions; individual impact was assessed through examination of differences in DIF adjusted and unadjusted depression estimates.

Results: Many items evidenced DIF; however, only a few had slightly elevated magnitude. No items evidenced salient DIF with respect to NCDIF and the scale-level impact was minimal for all group comparisons. The following short form items might be targeted for further study because they were also hypothesized to evidence DIF. One item showed slightly higher magnitude of DIF

[1] *Correspondence concerning this article should be addressed to:* Jeanne A. Teresi, Ed.D, Ph.D., Columbia University Stroud Center at New York State Psychiatric Institute, 1051 Riverside Drive, Box 42, Room 2714, New York, New York, 10032-3702, USA; email: Teresimeas@aol.com; jat61@columbia.edu

[2] New York State Psychiatric Institute

[3] Research Division, Hebrew Home at Riverdale; RiverSpring Health

[4] Department of Geriatrics and Palliative Medicine, Weill Cornell Medical Center

[5] Center for Mental Health and Aging, Department of Psychology, University of Alabama, Tuscaloosa

for age: nothing to look forward to; conditional on depression, this item was more likely to be endorsed in the depressed direction by individuals in older groups as contrasted with the cohort aged 21 to 49. This item was also hypothesized to show age DIF. Only one item (failure) showed DIF of slightly higher magnitude (just above threshold) for Whites vs. Asians/Pacific Islanders in the direction of higher likelihood of endorsement for Asians/Pacific Islanders. This item was also hypothesized to show DIF for minority groups. The impact of DIF was negligible. Conditional on depression, the items, worthless and hopeless were more likely to be endorsed in the depressed direction by respondents with less than high school education vs. those with a graduate degree; the magnitude of DIF was slightly above the T1 threshold, but not that of NCDIF. These items were also hypothesized to show DIF in the direction of more feelings of worthlessness by groups with lower education. While the magnitude and aggregate impact of DIF was small, in a few instances, individual impact was observed.

Information provided was relatively high, particularly in the middle upper (depressed) tail of the distribution. Reliability estimates were high (> 0.90) across all studied groups, regardless of estimation method.

Conclusions: This was the first study to evaluate measurement equivalence of the PROMIS Depression short forms across large samples of ethnically diverse groups. There were few items with DIF, and none of high magnitude, thus supporting the use of PROMIS Depression short form measures across such groups. These results could be informative for those using the short forms in minority populations or clinicians evaluating individuals with the depression short forms.

Key Words: depression, PROMIS®, differential item functioning, item response theory, ethnic diversity

Introduction

The Patient Reported Outcomes Measurement Information System® (PROMIS®) measures (Cella et al, 2007; Reeve et al., 2007) are being promoted internationally for use both clinically and in research. However, little information is available regarding their performance among ethnically diverse groups. Examination of item-level measurement equivalence is a central first step in evaluating the performance of measures because scale means should not be compared unless item-level measurement invariance is established (Meredith, 1993; Millsap & Meredith, 1992; Meredith & Teresi, 2006; van de Vijver & Leung, 1997). Methods for establishing invariance include analyses of differential item functioning (DIF; Holland & Wainer, 1993). Because PROMIS item banks were developed using item response theory (IRT) and are on a 5-point ordinal scale, a graded response model (Samejima, 1969) was used to examine DIF. The specific method was a comparison of parameters from nested DIF models using a variant of the Wald test based on Lord's chi-square (Cai, Thissen, & du Toit, 2011; Langer, 2008; Lord, 1980; Teresi, Kleinman, & Ocepek-Welikson, 2000; Woods, Cai, & Wang, 2013).

DIF in depression item banks and short form measures

The use of item banks and short forms derived from such banks for depression assessment is growing (Choi, Reise, Pilkonis, Hays, & Cella, 2010; Forkmann et al., 2013), and it is important to examine these banks for DIF. For example, using an analysis of variance approach to DIF examination based on Rasch (1960) analyses, Forkmann et al. reported no DIF for age or gender for their depression item bank.

DIF was examined in the 32 item PROMIS depression bank by four teams of measurement statisticians using five methods (Teresi et al., 2009). Age, gender and education DIF were examined using the IRT log-likelihood ratio tests (Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998; Thissen, Steinberg, & Gerard, 1986; Thissen, Steinberg, & Wainer, 1993). Other methods used in sensitivity analyses were item response theory ordinal logistic regression (IRTOLR; Crane, van Belle, & Larson, 2004), Differential Functioning of Items and Tests (DFIT; Raju, 1999; Raju, van der Linden, & Fleer, 1995), Simultaneous Item Bias Test, (SIBTEST; Shealy & Stout, 1993a, 1993b) and Multiple Indicator Multiple Cause (MIMIC; Jones, 2006; Muthén, 1984). Most items (22 / 32) showed significant DIF for at least one method or comparison. Significance tests were accompanied by magnitude measures such as the non-compensatory DIF (NCDIF) index (Raju et al.; Flowers, Oshima, & Raju, 1999; Oshima, Kushubar, Scott, & Raju, 2009). Only items with high magnitude and hypothesized, consistent DIF were flagged. A consistent finding across all methods was of gender DIF associated with the item, "I felt like crying." The item was a more severe indicator of depression for men than for women, a finding both hypothesized by PROMIS content experts, and found in the literature on DIF in depression measures. The item, "I had trouble enjoying the things I used to enjoy" was hypothesized to have higher conditional endorsement in men. This was confirmed by two analyses. The item was found to have salient DIF in several of the analyses for one or more gender, age and/or education comparisons. "I felt that I had no energy," hypothesized by content experts to possibly show gender and age DIF was confirmed by several methods to show age, gender or education DIF. Conditional on depression, those 65 and over were more likely to report no energy. Those with lower education were more likely to endorse the item. Scale level impact was assessed using expected scale scores, expressed as group differences in the total scale response functions. Group impact of DIF in the PROMIS Depression item bank was found to be minimal when mean scale or latent trait scores were examined with and without adjustment for DIF. This result was confirmed examining the expected scale score functions. Individual impact was observed for about 100 people. Based on the results, review of hypotheses generated by content experts and findings from the literature, items with high magnitude DIF were removed from the PROMIS item bank and from depression short forms (Teresi et al., 2009).

Short forms are frequently developed based on item bank parameters. The PROMIS Depression and Anxiety short forms have been examined for clinical validity (Schalet, et al., in press), and minimally important differences established for PROMIS cancer scales (Yost, Eton, Garcia, & Cella, 2011). However, differential item functioning analyses have been limited, and non-existent for ethnically diverse groups. In one of the few studies of DIF in the PROMIS short forms extant, Bjorner, Rose, Gandek, Stone, Junghaenel,

& Ware (2014) examined the eight item PROMIS physical function, fatigue and depression short forms among a sample of adults with chronic obstructive pulmonary disease. They performed DIF analyses of administration mode: interactive voice response, paper questionnaires, personal digital assistant and personal computer on the Internet. Multigroup confirmatory factor analyses (MG-CFA), examining thresholds and factor loadings across mode (of administration) groups was conducted. DIF effects (impact) were examined by fixing and freeing IRT threshold and slope parameters. Equivalence of response was observed across response modes.

DIF in traditional depression measures

Briefly reviewed are findings regarding DIF in depression measures in general. Although a complete review is beyond the scope of this paper, it is noted that many depression scales have been found to have items with DIF (Chan, Orlando, Ghosh-Dastidar, & Sherbourne, 2004; Cole, Kawachi, Maller, & Berkman, 2000; Grayson, Mackinnon, Jorm, Creasey, & Broe, 2000; Pickard, Dalal, & Bushnell, 2006; Yang & Jones, 2007). Items related to sadness showed DIF based on physical disorder and interview mode (Grayson et al.; Chan et al.). DIF was also observed in the "crying" items with respect to gender (Cole et al.; Gelin & Zumbo, 2003; Reeve, 2000; Yang & Jones), race/ethnicity (Spanish-speakers; Azocar, Areán, Miranda, & Muñoz, 2001; Teresi & Golden, 1994), physical disorder (Grayson et al.), and stroke (Pickard et al.). The impact of DIF has been found to be substantial in some studies (Azocar et al.; Chan, et al.; Cole et al.; Kim, Pilkonis, Frank, Thase, & Reynolds, 2002). A more detailed review of DIF in depression, anxiety and quality-of-life measures can be found in Teresi, Ramirez, Lai, and Silver (2008).

Aims of the analyses

The purpose of these analyses was to examine the item-level performance of the short form PROMIS Depression scale among different educational, ethnic/racial, gender, age and language groups, focusing on differential item functioning.

Methods

Sample generation and description

The sample sizes for the depression DIF analyses were as follows. The studied (also called the focal) group was males in the analyses of gender; the sample sizes for the groups were 3,241 females and 2,183 males. In the analyses of education, the reference group was graduate degree (n = 640). The studied groups were less than high school (n = 965), high school (n = 1,047), some college (n = 1,750) and college degree (n = 984). The reference group for age was 21 to 49 (n = 1,198); the studied groups were 50 to 64 (n = 2,003) and 65 to 84 (n = 2,223). For the analyses of ethnicity the reference group

was non-Hispanic White (n = 2,263); the studied groups were non-Hispanic Blacks (n = 1,116), Hispanics (n = 1,039) and Asians/Pacific Islanders (n = 906). Within the Hispanic subsample, there were 334 interviews conducted in Spanish and 700 in English.

Measures

Depressive symptoms assessment was a subdomain of emotional distress. The PROMIS short form Depression scales were developed by selecting items that maximized measurement precision and were most informative regardless of their location on the trait (Choi et al., 2010; Pilkonis, Choi, Reise, Stover, Riley, & Cella, 2011). Short form items were selected from the item bank based on the rank-order of IRT information provided and frequency of administration in the computerized adaptive test (CAT). The eight item short form was almost as precise as the CAT in the middle and upper part of the distribution, and less so at the extremes of the distribution. The current study included items from several short form scales identified in Table 1. In addition to eight short form items, two items were selected for this study based on their rank-ordering in terms of information. All ten items were examined in the analyses. The timeframe for all items was the past seven days. Items were administered using a five point response scale: *never*, *rarely*, *sometimes*, *often* and *always*.

Procedures and statistical approach

Qualitative analyses and hypotheses generation

Extensive qualitative analyses, including focus groups and cognitive interviews were performed with respect to PROMIS items, which target a sixth grade reading level (see DeWalt, Rothrock, Yount, & Stone, 2007). DIF hypotheses were generated for these analyses by asking a set of clinicians and other content experts to indicate whether or not they expected DIF to be present, and the direction of the DIF with respect to several comparison groups: gender, age, race/ethnicity, language and education. Hypotheses with respect to diagnostic groups were also elicited; however, all hypotheses related to the diagnosis of cancer. Because all patients carried such a diagnosis, the sample sizes did not permit DIF evaluation within diagnostic categories. The hypotheses for diagnosis are included for completeness, and in the event future studies permit examination of cancer diagnoses.

A definition of DIF was provided, and the following instructions related to hypotheses generation were given:

> Differential item functioning means that individuals from different sociodemographic groups with the same underlying trait (state) level will have different probabilities of endorsing an item. Put another way, reporting a symptom (e.g., crying frequency) should depend only on the level of the trait (state), e.g., depression, and not on membership in a group, e.g., male or female. Very specifically, randomly selected persons from each of two groups (e.g., males and females) who are at the

Table 1:
DIF hypotheses generated by 9 content experts for depression (Italicized entries are those with 2 or more ratings in the same direction.) (The items from the PROMIS Depression short forms are shown in the item stem column)

#	Stem	Gender	Age	Race/Ethnicity	Language	Education	Diagnosis
1	I felt worthless (4a, 6a, 8a, 8b)	4[a] *Women more worthless* (4)[b]		4 *Blacks & Latinos, Japanese, minorities* more worthless (2)	3 *Spanish more worthless* (3)	2 *Lower education more worthless* (2)	2 *Cancer more worthless* (2)
2	I felt that I had nothing to look forward to (8a, 8b)	2 *Women less to look forward to* (2)	4 *Older less to look forward to* (3)	2 *Whites, Japanese less to look forward to*	2 *Spanish more to look forward to*		2
3	I felt helpless (4a, 6a, 8a, 8b)	3 *Women more helpless* (2)	2	4 *Black & Latinos, Japanese, Minorities more helpless*	3 *Non-English more helpless*	3 Inconsistent direction	2 *Cancer more helpless*
4	I felt sad (8b)	3 *Women more sad* (2)					
5	I felt like a failure (6a, 8a, 8b)			5 *Asians, Blacks, Japanese more like failure*	3	3 Inconsistent direction	2 *Cancer more like failure*
6	I felt depressed (4a, 6a, 8a, 8b)	3 *Women more depressed* (2)		3 *Japanese, Whites more depressed*	2		
7	I felt unhappy (6a, 8a, 8b)	3 *Women more unhappy* (2)	0		2		
8	I felt hopeless (4a, 6a, 8a, 8b)			2 *Latinos less; Whites more hopeless*	2	2 *Lower education more hopeless*	
9	I felt discouraged about the future		2 Inconsistent direction				2 *Cancer/ Terminally ill more discouraged*
10	I felt disappointed in myself						

a Number indicates total number of hypotheses; b Number indicates number of directional hypotheses

same (e.g., mild) level of depression should have the same likelihood of reporting crying often. If it is theorized that this might not be the case, it would be hypothesized that the item has gender DIF.

A grid containing a row for each of the items and separate columns for each of the referenced groups was developed and distributed to content experts for completion in order to facilitate the rating. Forms were completed by nine content experts for depression (three clinical or counseling psychologists and six public health practitioners). The goal was to identify items that might have a different meaning or not be understood well and/or equivalently by individuals of any of the groups referenced. A summary of the DIF hypotheses is given in Table 1. A summary table (available from the authors) was also developed arraying the hypotheses and findings from the literature.

Quantitative analyses

Model Assumption of Unidimensionality: Item response theory assumptions include unidimensionality and local independence. The latter implies that the items are independent, conditional on the trait level. Model assumptions and fit were tested. Unidimensionality was examined using split samples, constructed by selection of two random halves in order to use one sample for cross-validation of results. The random first half of the sample was used for the exploratory factor analyses with principal components estimation and tests of scree, with cross-loadings permitted; and the second half was used to obtain the confirmatory and bi-factor solution. Essential unidimensionality was examined through a merged exploratory factor analysis (EFA) and confirmatory factor analysis (Asparouhov & Muthén, 2009) performed by fitting a unidimensional model with polychoric correlations using MPlus (Muthén & Muthén, 2011).

The confirmatory analyses of the unidimensional model and evaluation of the Comparitive Fit Index (CFI) was performed in the context of invariance testing and model fit (Bentler, 1990; Cook, Kallen, & Amtmann, 2009; Meade, Johnson, & Bradley, 2008). A bifactor model was compared with a unidimensional model. The bifactor model assumes that a single general trait explains most of the common variance but that group traits explain additional common variance for item subsets (Reise, Moore, & Haviland, 2010). A Schmid-Leiman (S-L; 1957) transformation using the "psych" R package (Rizopoulus, 2009) was performed in order to find an alternative set of group factors for the bi-factor model (Reise, et al.). All items were specified to load on the general factor, and the loadings on the group factors were specified following the Schmid-Leiman solution. M-PLUS (Muthén & Muthén, 2010) was used to both estimate the polychoric correlations based on the underlying continuous normal variables and to perform the final bi-factor modeling.

The explained common variance (ECV) provides information about whether the observed variance covariance matrix is close to unidimensionality (Sijtsma, 2009), and is estimated as the percent of observed variance explained (Reise, 2012; Reise, et al., 2010).

Local dependence (LD): The generalized, standardized local dependency chi-square statistics (Chen & Thissen, 1997) provided in IRTPRO, version 2.1 (Cai et al., 2011) was used to test the local independence assumption. Because local dependencies can result in

false DIF detection (Houts & Edwards, 2013), sensitivity analyses removing one item each from two pairs of items with higher LD values was performed.

IRT-model Fit: Model fit for the IRT model was examined using the root mean square error of approximation (RMSEA) from IRTPRO (Cai et al., 2011).

Descriptive Analyses: Prior to any formal tests of DIF, following a best practice recommended by Hambleton (2006), item frequencies were examined within each subgroup and for the total sample to detect possible problems with sparse data and skew.

Anchor Items and Linking: The first step in the analyses was to link the comparison groups in terms of depression and to estimate the mean and variance for the target groups studied (while setting the reference group mean to 0 and variance to 1). There are several methods for accomplishing this (Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Wang, Shih, & Sun, 2012; Woods, 2009), most of which rely upon anchor items, assumed to be DIF-free. An iterative process was used in selection of the anchor items for theta estimation. The method that was used in these analyses is a modified "all-other" anchor method in which initial DIF estimates were obtained by treating each item as a "studied" item, while using the remainder as "anchor" items. The purification process was also iterative, such that the analyses were repeated using the final subset of items identified as free of DIF as the "purified" anchor set. Items with DIF from the original anchor set were removed. This process continued until no changes in DIF status were observed. Items identified as DIF-free in the final model were not added back into the anchor set. The number and identity of the anchor items may be different for each socio-demographic comparison, e.g., race/ethnicity, education, and were determined by the number of iteratively identified DIF-free items. It has been suggested that at least four anchor items be used to measure an underlying latent variable (Cohen, Cohen, Teresi, Marchi, & Velez, 1990) and to serve as an anchor for linking metrics in DIF detection (Wang & Yeh, 2003) because greater numbers of anchor items increases power for DIF detection (Shih & Wang, 2009; Thissen et al., 1988). Anchor item selection procedures are presented in the methods overview article in this issue (Teresi & Jones, 2016).

Sensitivity Analyses for Anchor Item Selection: Sensitivity analyses for anchor item selection included the LR/f test (Woods, 2009) and rank order test statistic. Another method, a variant of the iterative backward all-other test approach was used, and p-values examined to select or remove the anchors (see Kopf, Zeileis, & Stobl, 2015). The difference in chi-square statistics resulting from two models was calculated, the first model with all parameters fixed to be equal for comparison groups, and the second, freeing all parameters for the studied item. The resulting log-likelihood ratio chi-square statistic was evaluated for significance. Because all items had the same number of response categories, in this case the results did not differ from the LR/f test or the rank order test statistics method.

Model for DIF Detection: The graded response model (Samejima, 1969) was used for the analyses of DIF. The item characteristic curve (ICC) that relates the probability of an item response to the underlying state, e.g., depression, measured by the item set is characterized by: a discrimination parameter, proportional to the slope of the curve (denoted a) and location (severity) parameter(s) (denoted b). An item shows DIF if people from different subgroups but at the same level of the attribute (denoted θ) have unequal prob-

abilities of endorsement. Put another way, the presence of DIF is demonstrated by ICCs that are different for the subgroups examined. The formula is given in the methods overview in this issue.

DIF Detection Tests: The primary method used for DIF detection was the Wald test for examination of group differences in IRT item parameters. For each studied item, a model was constructed with all parameters (except the studied item) constrained to be equal across comparison groups for the anchor items, and item parameters for the studied item freed to be estimated distinctly. An overall simultaneous joint test of differences in the *a* or *b* parameters was performed followed by step down tests for group differences in the *a* parameters, followed by conditional tests of the *b* parameters. Uniform DIF was detected when the *b* parameters differed and non-uniform DIF when the *a* parameters differed. Because tests of *b* parameters are performed constraining the *a* parameters to be equal, severity (*b)* parameters were interpreted as uniform DIF only if the tests of the *a* parameters were not significant.

Because there were three or more groups (three age, four race/ethnicity and five education), and the interest was in comparing the studied groups to the reference group, non-orthogonal rather than orthogonal contrasts were used. The final p values were adjusted using Bonferroni (1936) methods. In this case, the p value was adjusted for examination of 10 depression items ($p = 0.005$). Other methods such as Benjamini-Hochberg (B-H) have been used in sensitivity analyses (Benjamini & Hochberg, 1995; Thissen, Steinberg, & Kuang, 2002) .

Sensitivity Analyses for DIF Detection: A second DIF-detection method used in sensitivity analyses was based on ordinal logistic regression (OLR; Swaminathan & Rogers, 1990; Zumbo, 1999), which typically conditions on an observed variable. Uniform DIF is defined in the OLR framework as a significant group effect, conditional on the depression state; non-uniform DIF is a significant interaction of group and state. Three hierarchical models are tested; the first examines depression state (1), followed by group (2) and the interaction of group by state (3). Non-uniform DIF is tested by examining model 3 vs. 2; uniform DIF is tested by examining the incremental effect of model 2 vs. 1, with a chi-square (1 degree of freedom) test (Camilli & Shepard, 1994). A modification applied in these analyses, IRTOLR (Crane, Gibbons, Jolley & van Belle, 2006; Crane et al., 2004; Mukherjee, Gibbons, Kristiansson, & Crane, 2013) uses the depression estimates from a latent variable IRT model, rather than the traditional observed score conditioning variable, and incorporates effect sizes into the uniform DIF detection procedure. The software, lordif (Choi, Gibbons, & Crane, 2011) was used to perform IRTOLR.

Evaluation of DIF Magnitude and Effect Sizes: The magnitude of DIF refers to the degree of difference in item performance between or among groups, conditional on the trait or state being examined. Expected item scores can be examined as measures of magnitude. (See Figure 1 for examples.) An expected item score is the sum of the weighted (by the response category value) probabilities of scoring in each of the possible categories for the item. The method used for quantification of the difference in the average expected item scores was the non-compensatory DIF index (Raju et al., 1995) used in DFIT (Oshima et al., 2009; Raju, 1999; Raju et al., 2009). Additional effect size measures proposed by

Wainer (1993) and extended for polytomous data by Kim, Cohen, Alagoz, and Kim (2007) were also examined. For example, also reported here is the T1 effect size measure (Wainer), for which a recommended cutoff value is 0.10. However, primary reliance was on the NCDIF magnitude measure because little research has been conducted on the performance of T1. For a detailed description of these measures see Kleinman and Teresi (2016).

Cutoff values established based on simulations (Fleer, 1993; Flowers et al., 1999) can be used in the estimation of the magnitude of item-level DIF. For example, for the data presented here, the cutoff values were 0.0960 for polytomous items with five response options (Raju, 1999). Because NCDIF is expressed as the average squared difference in expected scores for individuals as members of the focal group and as members of the reference group, the square root of NCDIF provides an effect size in terms of the original metric. Thus, for a polytomous item with five response categories, the recommended cutoff of 0.0960 would correspond to an average absolute difference of 0.310 (almost one third of a point) on a five point scale (see Raju, 1999; Meade, Lautenschlager, & Johnson, 2007). Lower cutoff values have been identified based on simulations, e.g., Flowers et al., and item parameter replication methods have been recommended to derive sample-specific estimates (Seybert & Stark, 2012) in the context of power for DIF detection. However, in the context of magnitude measures, the practical meaning of group differences in expected scores reflected in NCDIF is also of consideration; thus the higher threshold values were used here as recommended by Raju.

Prior to application of the DFIT software, the estimates of the latent trait (theta) were calculated separately for each group, and equated together with the item parameters. Baker's (Baker, 1995) EQUATE program was used in an iterative fashion in order to equate the theta and item parameter estimates for the two groups and place them on a common metric. If DIF was detected, the item showing DIF was excluded from the equating algorithm, and new DIF-free equating constants were computed, and purified iteratively. Iterative purification of equating constants has been shown to reduce Type I error (Seybert & Stark, 2012).

Evaluation of DIF Impact: Aggregate-level impact was evaluated, examining expected scale score functions. Expected item scores were summed to produce an expected scale score (also referred to as the test or scale response function), which provides evidence regarding the effect of DIF on the total score. Group differences in these test response functions provide overall aggregated measures of DIF impact.

Impact at the individual level was examined by comparing DIF-adjusted and unadjusted estimates of the latent depression state scores. Estimates were adjusted for *all* items with DIF, not just for those with DIF after adjustment for multiple comparisons or those with high DIF magnitude. Individual impact was evaluated by fixing and freeing parameters to account for DIF, and comparing the results with and without DIF adjustment. Two different theta estimates were compared: 1) thetas based on the equated item parameters for the subgroups for all the items and 2) thetas based on free estimation of the parameters for items showing DIF. The latter estimate produced different subgroup parameters for items with DIF. The impact can be presented in two different ways: 1) the number of individual theta estimates that differ by more than 0.5 or 1.0 standard deviations; 2)

based on a threshold value. An example of the latter is the use of an arbitrary cut-off value such as theta = 1.0 to classify individuals as those with and without depression symptomatology. A measure of individual impact is the number of individuals who change designations when theta is estimated based on DIF item parameters estimated freely vs. thetas when all item parameters are equated (set equal for comparison groups).

Evaluation of Reliability and Information: Reliability was evaluated by decomposing the scale score into the sum of the item scores, and the contribution of the common term or communality. McDonald's (McDonald, 1999) Omega Total (ω_t), a reliability estimate that is based on the proportion of total common variance explained, was also calculated. Both Cronbach's alpha (Cronbach, 1951) and ordinal alpha based on polychoric correlations (Zumbo, Gadermann, & Zeisser, 2007) were calculated. Additionally, IRT-based reliability measures were examined at selected points along the underlying latent continuum. Finally, the item and test information functions were calculated and graphed.

Results

Qualitative results

Table 1 shows the hypotheses generated for the depression items. Conditional on depression, it was hypothesized that women would express greater feelings of worthlessness, helplessness, sadness, depression, and unhappiness. It was hypothesized that women and older people would be more likely to report feeling that there was nothing to look forward to.

For race/ethnicity, conditional on depression, it was hypothesized that minority groups as contrasted with the White majority would express more feelings of worthlessness, helplessness and feeling like a failure. Spanish speakers were hypothesized to express greater worthlessness, helplessness and nothing to look forward to. Conditional on depression, it was posited that those with lower education would express more feelings of being worthless and hopeless.

Finally those with a diagnosis of cancer were posited to express greater feelings of worthlessness, helplessness, discouragement about the future and feeling like a failure, conditional on depression.

Quantitative results

Tests of model assumptions

Unidimensionality: As shown in Table 2, there was strong support for essential unidimensionality across all comparison socio-demographic groups. (The test of scree for the total sample is given in Appendix[6], Figure 1.) The first random half of the split sample

[6] To access online appendices, please use the following url: http://www.research-hhar.org/Tables/DEP-PTAM-appendix.htm

was used to perform exploratory principal components analyses and to fit a unidimensional confirmatory factor analyses (CFA). The principal components analyses showed that the ratio of component 1 to 2 was large (18.6 to 29.2) for all groups. The first component across comparison groups accounted for between 83 % and 88 % of the variance for all groups, supporting the essential unidimensionality of the item set across comparison subgroups.

As an additional test of dimensionality a bifactor model was examined using the second random half of the sample. Examination of the confirmatory factor analyses results in Table 3 show that the loadings on the single common factor were very similar to those observed on the general factor from the bifactor analyses, which provides additional evidence for unidimensionality. Additionally, the communality values were large, ranging from 0.81 to 0.91. The model fit indices (CFIs) for the unidimensional CFA from MPlus ranged from 0.988 to 0.994 across groups (see Appendix, Table 1); the ECVs ranged from 71.68 to 79.71 (see Table 4).

Table 2:
PROMIS depression short form item set: Tests of dimensionality from principal components analysis (eigenvalues by subgroup)

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio Component 1/ Component 2
Total Sample (n = 5459)					
Eigenvalues	8.612	0.315	0.283	0.187	27.3
Explained Variance	86.1 %	3.1 %	2.8 %	1.9 %	
Random First Half Sample (n = 2729)					
Eigenvalues	8.607	0.319	0.286	0.186	27.0
Explained Variance	86. 1%	3.2 %	2.9 %	1.9 %	
Females (n = 3241)					
Eigenvalues	8.529	0.347	0.286	0.199	24.6
Explained Variance	85.3 %	3.5 %	2.9 %	2.0 %	
Males (n = 2183)					
Eigenvalues	8.731	0.299	0.245	0.171	29.2
Explained Variance	87.3 %	3.0 %	2.5 %	1.7 %	
Age 21-49 (n = 1198)					
Eigenvalues	8.615	0.307	0.258	0.198	28.1
Explained Variance	86.1 %	3.1 %	2.6 %	2.0 %	
Age 50-64 (n = 2003)					
Eigenvalues	8.614	0.324	0.295	0.169	26.6
Explained Variance	86.1 %	3.2 %	2.9 %	1.7 %	
Age 65-84 (n = 2223)					
Eigenvalues	8.551	0.325	0.298	0.213	26.3
Explained Variance	85.5%	3.2%	3.0%	2.1%	

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio Component 1/ Component 2
Race/Ethnicity: Non-Hispanic White (n = 2263)					
Eigenvalues	8.579	0.334	0.264	0.217	25.7
Explained Variance	85.8 %	3.3 %	2.6 %	2.2 %	
Race/Ethnicity: Non-Hispanic Black (n = 1116)					
Eigenvalues	8.693	0.359	0.262	0.149	24.2
Explained Variance	86.9 %	3.6 %	2.6 %	1.5 %	
Race/Ethnicity: Hispanic (n = 1039)					
Eigenvalues	8.546	0.321	0.304	0.186	26.6
Explained Variance	85.5 %	3.2 %	3.0 %	1.9 %	
Race/Ethnicity: Non-Hispanic Asians/Pacific Islander (n = 906)					
Eigenvalues	8.759	0.319	0.255	0.155	27.5
Explained Variance	87.6 %	3.2 %	2.6 %	1.6 %	
Education: Less Than High School (n = 965)					
Eigenvalues	8.555	0.365	0.271	0.169	23.4
Explained Variance	85.6 %	3.7 %	2.7 %	1.7 %	
Education: High School (n = 1047)					
Eigenvalues	8.614	0.316	0.273	0.195	27.3
Explained Variance	86.1 %	3.2 %	2.7 %	1.9 %	
Education: Some College (n = 1750)					
Eigenvalues	8.689	0.324	0.257	0.19	26.8
Explained Variance	86.9 %	3.2 %	2.6 %	1.9 %	
Education: College Degree (n = 984)					
Eigenvalues	8.283	0.445	0.315	0.237	18.6
Explained Variance	82.8 %	4.4 %	3.1 %	2.4 %	
Education: Graduate Degree (n = 640)					
Eigenvalues	8.303	0.416	0.33	0.263	20.0
Explained Variance	83.0 %	4.2 %	3.3 %	2.6 %	
Hispanics Interviewed in English (n = 700)					
Eigenvalues	8.548	0.334	0.321	0.189	25.6
Explained Variance	85.5 %	3.3 %	3.2 %	1.9 %	
Hispanics Interviewed in Spanish (n = 334)					
Eigenvalues	8.525	0.344	0.285	0.205	24.8
Explained Variance	85.3 %	3.4 %	2.9 %	2.1 %	

Table 3:

PROMIS depression short form item set: Item loadings (λ) from the unidimensional confirmatory factor analysis (MPlus) for the first half of the random sample (n=2729), Schmid-Leiman (S-L) bi-factor model with two and three group factors (performed with R for the second random half sample) and MPlus bi-factor two group factor solution for the second random half sample (n= 2730)

Item Description	One Fact.* λ (s.e.)	Schmid-Leiman Bi-Factor Three and Two Group Factor Solutions									MPlus Bi-Factor Two Group Factor Solution (Based on S-L** Result)		
		G λ	F1 λ	F2 λ	F3 λ	h^2	G λ	F1 λ	F2 λ	h^2	G λ (s.e.)	F1 λ (s.e.)	F2 λ (s.e.)
I felt worthless	0.91 (0.005)	0.89		0.29		0.87	0.88	0.27		0.85	0.89 (0.006)	0.34 (0.019)	
I felt that I had nothing to look forward to	0.94 (0.004)	0.92		0.23		0.90	0.92	0.23		0.89	0.92 (0.005)	0.23 (0.014)	
I felt helpless	0.91 (0.005)	0.89		0.24		0.86	0.89	0.22		0.84	0.90 (0.006)	0.22 (0.015)	
I felt sad	0.92 (0.004)	0.89	0.28			0.87	0.90		0.22	0.86	0.90 (0.005)		0.26 (0.016)
I felt like a failure	0.93 (0.004)	0.91				0.88	0.91	0.20		0.87	0.94 (0.004)	0.04 (0.014)	
I felt depressed	0.94 (0.003)	0.90	0.28			0.90	0.92		0.23	0.90	0.91 (0.005)		0.29 (0.014)
I felt unhappy	0.93 (0.003)	0.90	0.27			0.89	0.91		0.23	0.89	0.91 (0.005)		0.246 (0.015)
I felt hopeless	0.95 (0.003)	0.93				0.91	0.94			0.91	0.96 (0.003)		
I felt discouraged about the future	0.91 (0.004)	0.89			0.20	0.84	0.89			0.83	0.92 (0.004)		
I felt disappointed in myself	0.90 (0.005)	0.88			0.28	0.85	0.88			0.81	0.91 (0.005)		

* Geomin (oblique) rotation ** Schmid-Leiman bi-factor model; Fact. = Factor

Note: The comparative fit index (CFI) for the MPlus one-factor solution is 0.992 and for the bi-factor solution, 0.997

h^2 is the communality. G λ are the loadings on the general factor; F1 λ through F3 λ are the loadings on the group factors

Table 4:
PROMIS depression short form item set. Reliability statistics: Cronbach's alpha, ordinal alpha, McDonald's Omega Total, and explained common variance (ECV) for the total sample and demographic subgroups ("Psych" R package)

	Cronbach's Alpha	Ordinal Alpha	McDonald's Omega	ECV
Total Sample	0.968	0.982	0.982	77.804
Random Second Half Sample	0.968	0.982	0.982	77.670
Age 21 to 49 years	0.969	0.982	0.982	78.552
Age 50 to 64 years	0.969	0.982	0.982	78.155
Age 65 to 84 years	0.965	0.981	0.981	76.187
Male	0.970	0.984	0.984	78.702
Female	0.967	0.981	0.981	77.153
Non-Hispanic White	0.965	0.982	0.982	76.584
Non-Hispanic Black	0.968	0.983	0.984	78.135
Hispanic	0.969	0.981	0.981	78.198
Non-Hispanic Asian/Pacific Islander	0.972	0.984	0.985	79.710
Less Than High School	0.970	0.981	0.981	78.550
High School Degree	0.969	0.982	0.982	78.312
Some College	0.969	0.983	0.983	78.345
College Graduate	0.957	0.977	0.977	72.404
Graduate Degree	0.955	0.977	0.977	71.677
Hispanics Interviewed in English	0.968	0.981	0.981	77.883
Hispanics Interviewed in Spanish	0.969	0.981	0.981	78.346

Local Independence: In general, the local dependency statistics (not shown) were in the acceptable range. However, in sensitivity analyses, one item from a pair that evidenced higher LD values was removed. Item 7 – "I felt unhappy" evidenced the highest LD values with item 6 – "I felt depression" for the Black (28.2) and low education (24.4) subgroups. The results of the DIF analyses after item removal varied only slightly in terms of the parameter estimates, and the DIF p values were very similar, resulting in no change in DIF designations.

Tests of model fit

The fit statistics (RMSEA's) from IRTPRO for the IRT models (see Appendix, Table 1) ranged from 0.04 to 0.07 across DIF subgroup comparison models, indicating good to acceptable fit.

Table 5:
PROMIS depression short form item set: Item response theory (IRT) reliability estimates at varying levels of the attribute (theta) estimate based on results of the IRT analysis (IRTPRO) for total sample and demographic subgroups

Depression (Theta)	IRT Reliability																
	Total	F	M	Age 21-49	Age 50-64	Age 65-84	NHW	NHB	Hisp.	NH API	<HS	HS	Some Coll.	Coll.	Grad.	Lang. Engl.	Lang. Span.
-1.2	0.65	0.72	0.57	0.77	0.69	0.59	0.63	0.61	0.77	0.62	0.78	0.68	0.65	0.61	0.60	0.72	0.87
-0.8	0.84	0.89	0.75	0.92	0.87	0.77	0.82	0.80	0.91	0.82	0.92	0.87	0.84	0.78	0.76	0.89	0.95
-0.4	0.95	0.95	0.92	0.97	0.95	0.92	0.94	0.94	0.97	0.95	0.97	0.96	0.95	0.91	0.90	0.96	0.98
0.0	0.98	0.98	0.97	0.98	0.98	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.97	0.96	0.95	0.98	0.98
0.4	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.98	0.98
0.8	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
1.2	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
1.6	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98
2.0	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.97	0.98	0.98	0.98	0.98	0.98	0.96
2.4	0.97	0.97	0.98	0.95	0.97	0.98	0.98	0.97	0.93	0.97	0.90	0.97	0.98	0.98	0.98	0.95	0.90
2.8	0.90	0.90	0.91	0.84	0.89	0.95	0.95	0.91	0.80	0.90	0.72	0.89	0.92	0.97	0.98	0.83	0.76
Overall (Average)	**0.93**	**0.94**	**0.91**	**0.94**	**0.93**	**0.92**	**0.93**	**0.92**	**0.93**	**0.92**	**0.92**	**0.93**	**0.93**	**0.92**	**0.91**	**0.93**	**0.94**

Note: Reliability estimates were calculated for theta levels for which there were respondents
NHW = Non-Hispanic Whites; NHB = Non-Hispanic Blacks; Hisp. = Hispanic; NHAPI = Non-Hispanic Asian/Pacific Islander
Coll. = college; Lang. = language; Engl. = English; Span, = Spanish

Reliability estimates

The reliability estimates were high. The Omega total values (Table 4) ranged from 0.977 to 0.985, the Cronbach's alphas ranged from 0.955 to 0.972, and the ordinal alphas based on the polychoric correlations were 0.977 to 0.984. Finally, the reliability estimates (precision) at points along the latent trait (theta) reflective of where respondents were observed were high, ranging from 0.72 to 0.99, except for at the lowest point (theta = - 1.2) where the estimates were lower, ranging from 0.57 to 0.87. The overall reliability estimate was 0.93 for the total sample ranging from 0.91 to 0.94 for the individual subgroups (see Table 5).

IRT parameter estimates, tests of DIF and assessment of magnitude and impact

Shown in Table 6 are the graded response item parameters and their standard errors for the total sample. Appendix Table 2 shows the discrimination (*a)* parameters across subgroup comparisons. As shown, the *a* parameters vary somewhat across items and groups, ranging from 3.60 to 6.46 across items for the total sample and from 3.13 ("I felt worthless" for non-Hispanic Blacks) to 7.45 ("I felt hopeless" for those with some college).

Table 6:
PROMIS depression short form item set: Item response theory (IRT) item parameters and standard error estimates (using IRTPRO) for the total sample (n = 5,459)

Item Description	***a***	**s.e. of *a***	***b1***	**s.e.**	***b2***	**s.e.**	***b3***	**s.e.**	***b4***	**s.e.**
I felt worthless	3.77	0.10	0.35	0.02	0.80	0.02	1.54	0.03	2.30	0.04
I felt that I had nothing to look forward to	4.63	0.12	0.36	0.02	0.81	0.02	1.51	0.02	2.18	0.04
I felt helpless	4.23	0.11	0.29	0.02	0.75	0.02	1.46	0.02	2.15	0.04
I felt sad	4.14	0.10	-0.25	0.02	0.35	0.02	1.22	0.02	1.99	0.03
I felt like a failure	4.69	0.13	0.42	0.02	0.89	0.02	1.56	0.03	2.15	0.04
I felt depressed	4.67	0.12	0.01	0.02	0.53	0.02	1.28	0.02	1.97	0.03
I felt unhappy	4.65	0.11	-0.17	0.02	0.45	0.02	1.32	0.02	2.07	0.04
I felt hopeless	6.46	0.20	0.38	0.02	0.83	0.02	1.44	0.02	2.06	0.03
I felt discouraged about the future	4.09	0.10	0.06	0.02	0.60	0.02	1.34	0.02	2.00	0.03
I felt disappointed in myself	3.60	0.09	0.20	0.02	0.73	0.02	1.45	0.03	2.15	0.04

a = item discrimination; *b* = item severity, s.e.= standard error

DIF results

Appendix Tables 3-6 show the detailed DIF results for race/ethnicity, education, age and gender, respectively. Tables 7-9 are summaries of the DIF results. Table 7 shows the results for race/ethnicity. As shown, eight items showed DIF for both IRTPRO (Wald tests after Bonferroni correction) and for lordif (ordinal logistic regression). These items were: worthless, nothing to look forward to, helpless, failure, unhappy, hopeless, discouraged about the future and disappointed in myself. Conditional on depression, Asians/Pacific Islanders (as contrasted with non-Hispanic Whites) had a higher probability of responding in the depressed direction to the items: worthless, nothing to look forward to, helpless, failure, unhappy and disappointed. These items evidenced lower *b* parameters and were less severe indicators of depression for Asians/Pacific Islanders than for the reference group. For the item, discouraged about the future, all minority groups in contrast to the White non-Hispanic reference group evidenced a lower probability of a depressed response. Non-Hispanic Blacks in contrast to the reference group evidenced a lower probability of endorsing the item, worthless, conditional on depression. Hispanics as contrasted with non-Hispanic Whites evidenced a lower probability of item endorsement for the items: hopeless and disappointed. Conditional on depression, the likelihood of endorsing the item, unhappy was lower for Hispanics vs. non-Hispanic Whites. Only one item showed DIF of higher magnitude (just above threshold on the T1 statistic) for non-Hispanic Asians/Pacific Islanders vs. Whites: "I felt like a failure" (see Table 7). However, the magnitude of DIF was small and the NCDIF statistic was not above threshold or large. The impact of DIF was negligible, as shown by the overlapping curves (see Figure 1).

DIF analysis was performed for Hispanics only contrasting those interviewed in English with those interviewed in Spanish. Spanish speakers were hypothesized to express greater feelings of worthlessness, helplessness and nothing to look forward to; however, no significant DIF was observed. Five items were selected as anchor items: "I felt that I had nothing to look forward to"; "I felt helpless"; "I felt depressed"; "I felt unhappy"; "I felt discouraged".

For education (Table 8), five items were identified with DIF after Bonferroni correction using the Wald test (worthless, unhappy, hopeless, discouraged and disappointed). Only three were consistently identified with both the Wald test and the OLR procedure: worthless, hopeless and disappointed. The item, hopeless had a lower discrimination for the patients with less than high school education than for the post-graduate reference group. Conditional on depression, the item, worthless was more likely to be endorsed in the depressed direction by the patients with less than high school education vs. the patients with a graduate degree. However, the items, unhappy, discouraged and disappointed in myself were less likely to be endorsed in the depressed direction by the patients with less than high school education compared with those with a graduate degree. The item, "I felt disappointed in myself" showed DIF of higher T1 magnitude for the graduate school vs. no high school groups. However, the NCDIF statistic was not above threshold.

Table 7:
PROMIS depression short form item set: Differential item function (DIF) results. Race/Ethnicity subgroup comparisons

Item description	IRTPRO			*lordif*			Magnitude (NCDIF)			Effect Size T1		
	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI
I felt worthless	NU*; U*	U	U*	U*; NU*	U*	NU*	0.0109	0.0013	0.0094	-0.0577	0.0090	-0.0656
I felt that I had nothing to look forward to			U*	U*	U*; NU	NU*	0.0042	0.0094	0.0079	0.0444	0.0530	-0.0343
I felt helpless			U*	NU	U*; NU	NU*	0.0034	0.0052	0.0117	-0.0291	0.0013	-0.0554
I felt sad					U*	U; NU	0.0006	0.0102	0.0011	-0.0106	-0.0906	0.0052
I felt like a failure	U		U*	U	U*	U*; NU*	0.0065	0.0174	0.0572	-0.0414	-0.0984	**-0.1719**†
I felt depressed					U*	U*; NU*	0.0014	0.0019	0.0047	-0.0272	-0.0357	0.0364
I felt unhappy		U*	U*		U*; NU*	U*	0.0018	0.0107	0.0029	-0.0068	0.0704	0.0229
I felt hopeless		NU; U*	U		U*	NU*	0.0008	0.0134	0.0032	-0.0081	-0.0890	-0.0179
I felt discouraged about the future	U*	U*	U*	U*	U*	U*; NU	0.0181	0.0077	0.0228	0.1074*	0.0736	0.1192†
I felt disappointed in myself	U	U*	NU; U*	U; NU*	U*; NU*	U*	0.0070	0.0111	0.0044	0.0467	0.0774	0.0477

All NCDIF values were smaller than the threshold (0.0960)
*Asterisks indicate significance after adjustment for multiple comparisons. † Indicates value above threshold of 0.10; bolded values are above 0.15.
NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters.
Hisp.= Hispanic; NHAPI = Non-Hispanic Asian/ Pacific Islander
For the *lordif* analyses, the Uniform and non-uniform DIF was determined using the likelihood ratio chi-square test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R2 measures or the change in Beta criterion.

Table 8:
PROMIS depression short form item set: Differential item function (DIF) results. Education subgroup comparisons

Item description	IRTPRO				*Lordif*				Magnitude (NCDIF)				Effect Size T1			
	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS
I felt worthless			NU	U*				U*	0.0001	0.0027	0.0068	0.0214	0.0020	-0.0330	-0.0649	-0.1222†
I felt I had nothing to look forward to								U*	0.0019	0.0020	0.0057	0.0083	-0.0256	-0.0210	-0.0545	-0.0575
I felt helpless				NU; U		NU		U*	0.0046	0.0083	0.0086	0.0155	-0.0240	-0.0214	-0.0492	-0.1011†
I felt sad								U*	0.0014	0.0029	0.0039	0.0035	0.0104	-0.0331	0.0276	0.0435
I felt like a failure		U		U		U*		U*	0.0008	0.0092	0.0015	0.0017	0.0061	0.0631	0.0249	0.0088
I felt depressed								U*	0.0000	0.0025	0.0008	0.0008	0.0007	-0.0366	0.0028	-0.0060
I felt unhappy			U	U*					0.0007	0.0019	0.0055	0.0127	-0.0028	0.0306	0.0468	0.0864
I felt hopeless				NU*; U				U*	0.0018	0.0022	0.0047	0.0218	-0.0222	-0.0294	-0.0489	-0.1161†
I felt discouraged about the future				U*					0.0037	0.0010	0.0041	0.0182	0.0397	0.0139	0.0495	0.1129†
I felt disappointed in myself		U		U*		U*			0.0042	0.0173	0.0112	0.0333	0.0410	0.0961	0.0832	**0.1568**†

All NCDIF values were smaller than the threshold (0.0960) *Asterisks indicate significance after adjustment for multiple comparisons. † Indicates value above threshold of 0.10; bolded values are above 0.15.
NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters. GD = graduate degree; HS = high school; Coll. = college
For the *lordif* analyses, the uniform and non-uniform DIF was determined using the likelihood ratio chi-square test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R2 measures or the change in Beta criterion.

Table 9:
PROMIS depression short form item set: Differential item function (DIF) results. Gender and age subgroup comparisons

Item description	IRTPRO			*lordif*			Magnitude (NCDIF)			Effect Size T1		
	Gender	Age		Gender	Age		Gender	Age		Gender	Age	
		21-49 vs. 50-64	21-49 vs. 65-84		21-49 vs. 50-64	21-49 vs. 65-84		21-49 vs. 50-64	21-49 vs. 65-84		21-49 vs. 50-64	21-49 vs. 65-84
I felt worthless			U*			U*	0.0020	0.0024	0.0128	-0.0317	-0.0355	-0.0786
I felt that I had nothing to look forward to		U*	U*		U*	U*; NU	0.0017	0.0112	0.0357	-0.0262	-0.0781	-0.1292†
I felt helpless				U*; NU			0.0023	0.0006	0.0017	0.0302	0.0167	-0.0244
I felt sad	U*	U	U*	U*; NU*	U*	U*; NU	0.0266	0.0077	0.0276	0.1321†	0.0754	0.1378†
I felt like a failure				NU			0.0055	0.0002	0.0005	-0.0471	-0.0107	-0.0042
I felt depressed				U*; NU		U*	0.0040	0.0004	0.0041	0.0497	0.0119	0.0499
I felt unhappy				U*		U*	0.0009	0.0002	0.0038	0.0187	0.0051	0.0453
I felt hopeless						U	0.0009	0.0005	0.0004	-0.0175	0.0142	0.0005
I felt discouraged about the future							0.0049	0.0004	0.0007	-0.0472	0.0060	-0.0182
I felt disappointed in myself						U*	0.0025	0.0012	0.0045	-0.0328	-0.0037	0.0276

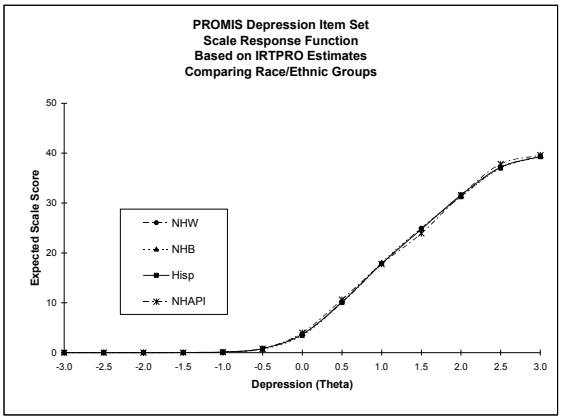
All NCDIF values were smaller than the threshold (0.0960) *Asterisks indicate significance after adjustment for multiple comparisons. † Indicates value above threshold of 0.10.
NU= Non-uniform DIF involving the discrimination parameters; U=Uniform DIF involving the location parameters.
For the *lordif* analyses, the uniform and non-uniform DIF was determined using the likelihood ratio chi-square test. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF was not detected using the pseudo R2 measures or the change in Beta criterion.

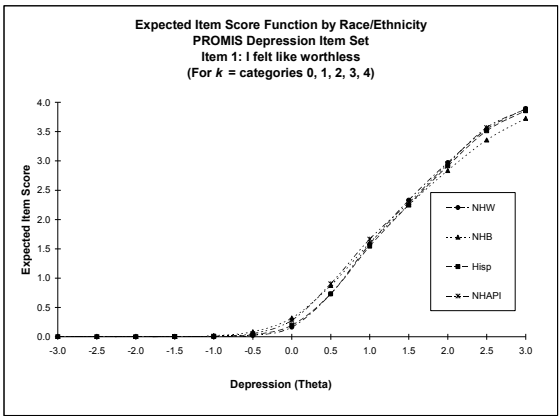
In summary, conditional on depression, the items, worthless and hopeless were more likely to be endorsed in the depressed direction by the patients with less than high school education vs. the patients with a graduate degree; the magnitude of DIF was slightly above the T1 threshold. These items were also hypothesized to show DIF in the direction of more feelings of worthlessness by groups with lower education. However, the NCDIF statistic was not above threshold. The impact of DIF on the scale was trivial (see Table 8 and Figure 1).

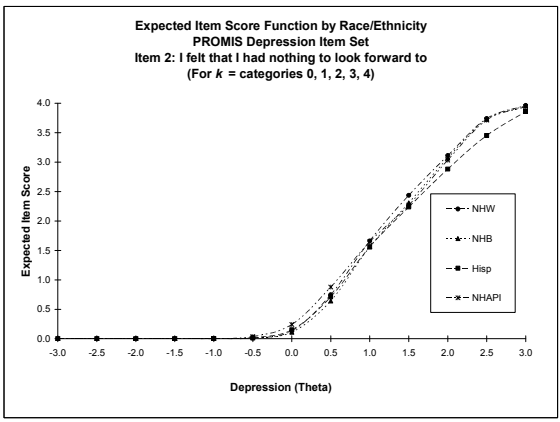
As shown in Table 9, one item (sad) showed gender DIF after Bonferroni correction and three showed age DIF (worthless, nothing to look forward to and sad). Conditional on depression, women were more likely than men to respond to the item, sad in the depressed direction. This item evidenced consistent DIF of slightly higher magnitude as evidenced by a value slightly above the T1 threshold for gender, although the NCDIF index was not above threshold. This item, sad, was also hypothesized to show DIF in the expected direction.

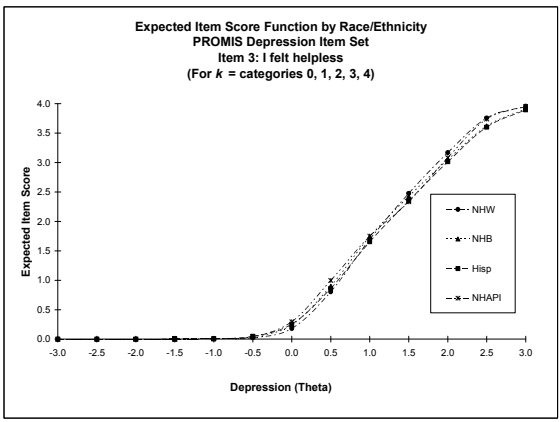
Conditional on depression, older respondents were more likely to respond in the depressed direction to the items worthless and nothing to look forward to, while the item, sad was more likely endorsed in the depressed direction by the younger age groups. One item showed slightly higher magnitude for age: nothing to look forward to; however, the NCDIF magnitude measure was not above threshold. The scale level impact was trivial (see Table 9 and Figure 1). Conditional on depression, this item was more likely to be endorsed in the depressed direction by both older groups in contrast with the cohort aged 21-49.

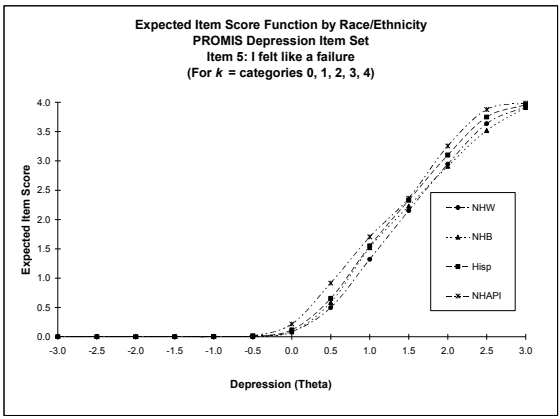
Figure 1:
PROMIS depression short form item set: Expected scale and item scores for race/ethnicity subgroups

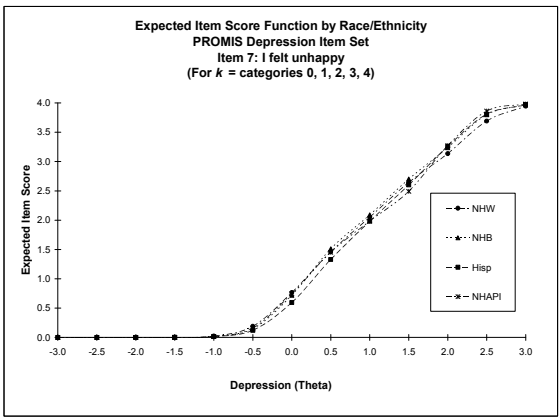


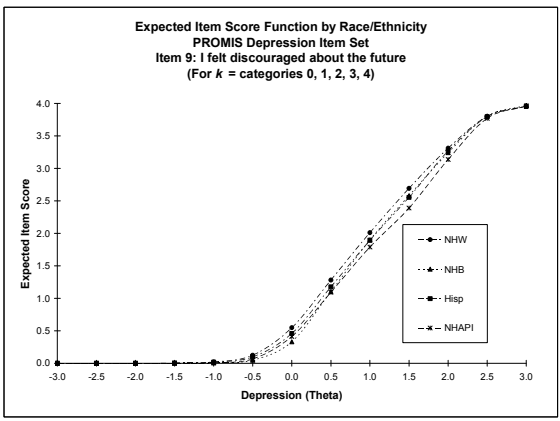

Expected Item Score Function by Race/Ethnicity
PROMIS Depression Item Set
Item 1: I felt like worthless
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
Depression (Theta)
NHW
NHB
Hisp
NHAPI


Expected Item Score Function by Race/Ethnicity
PROMIS Depression Item Set
Item 2: I felt that I had nothing to look forward to
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
Depression (Theta)
NHW
NHB
Hisp
NHAPI


Expected Item Score Function by Race/Ethnicity
PROMIS Depression Item Set
Item 3: I felt helpless
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
Depression (Theta)
NHW
NHB
Hisp
NHAPI


Expected Item Score Function by Race/Ethnicity
PROMIS Depression Item Set
Item 5: I felt like a failure
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
Depression (Theta)
NHW
NHB
Hisp
NHAPI


Expected Item Score Function by Race/Ethnicity
PROMIS Depression Item Set
Item 7: I felt unhappy
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
Depression (Theta)
NHW
NHB
Hisp
NHAPI


Expected Item Score Function by Race/Ethnicity
PROMIS Depression Item Set
Item 9: I felt discouraged about the future
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
Depression (Theta)
NHW
NHB
Hisp
NHAPI

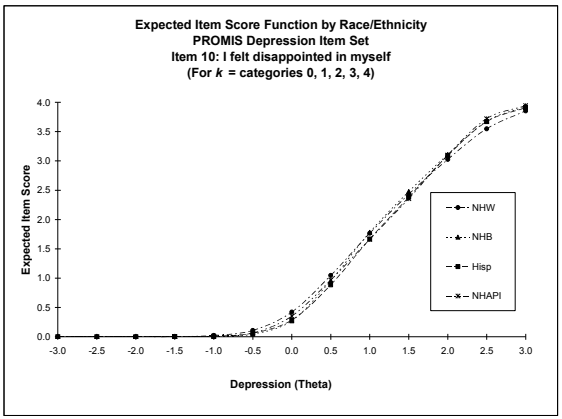
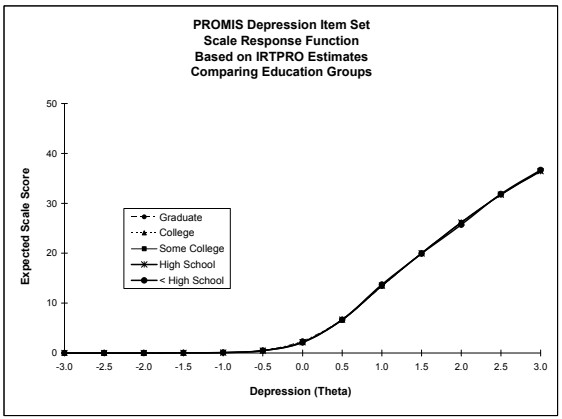
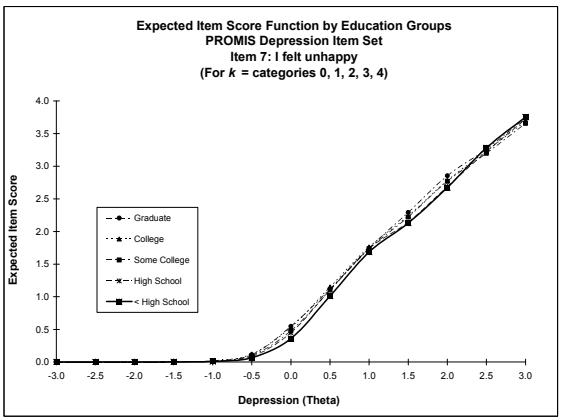
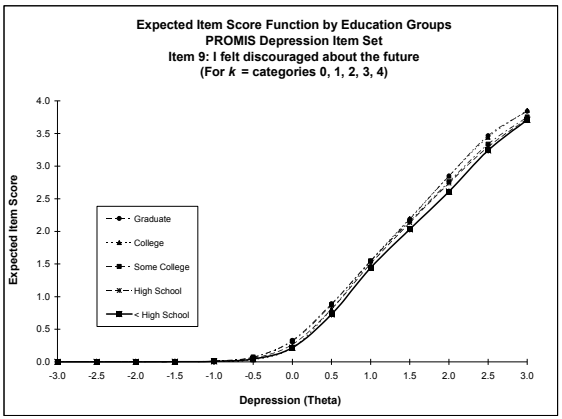


Figure 1: - cont.
PROMIS depression short form item set: Expected scale and item scores for education subgroups







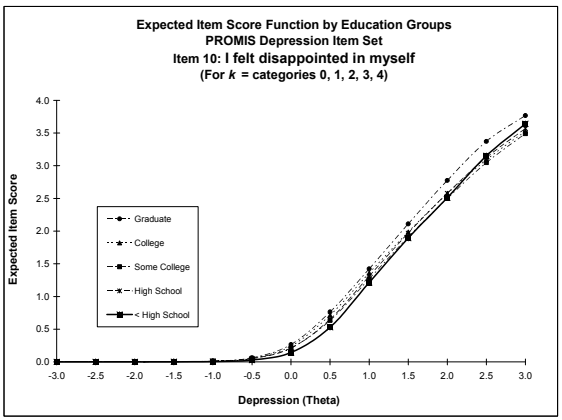
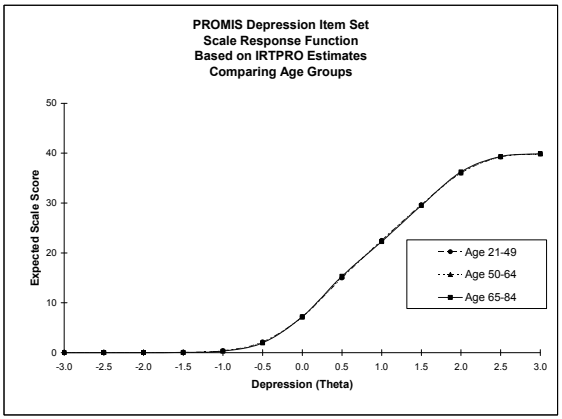
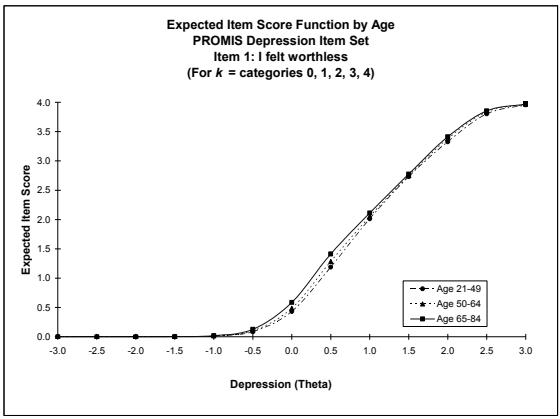
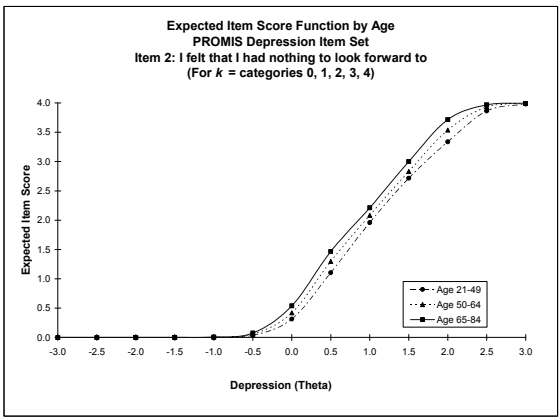


Figure 1: - cont.
PROMIS depression short form item set: Expected scale and item scores for age subgroups




Expected Item Score Function by Age
PROMIS Depression Item Set
Item 1: I felt worthless
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
Depression (Theta)
Age 21-49
Age 50-64
Age 65-84


Expected Item Score Function by Age
PROMIS Depression Item Set
Item 2: I felt that I had nothing to look forward to
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
Depression (Theta)
Age 21-49
Age 50-64
Age 65-84

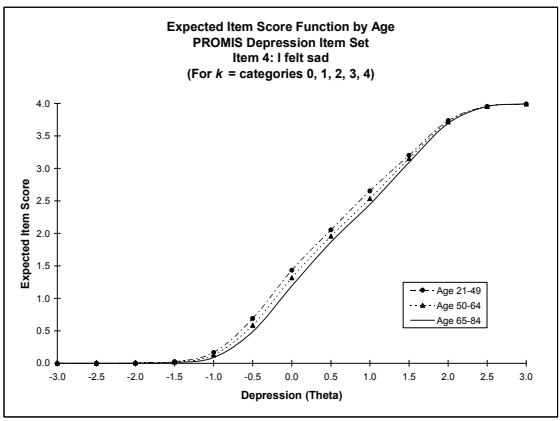
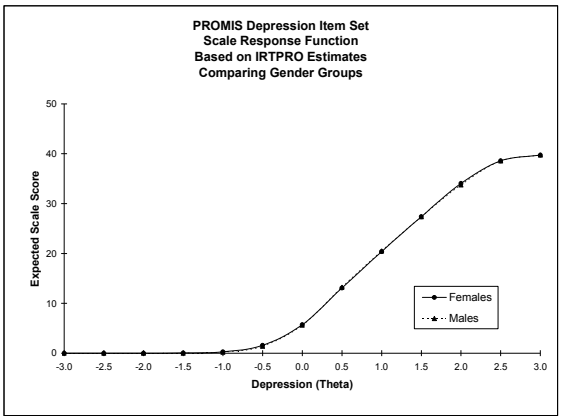
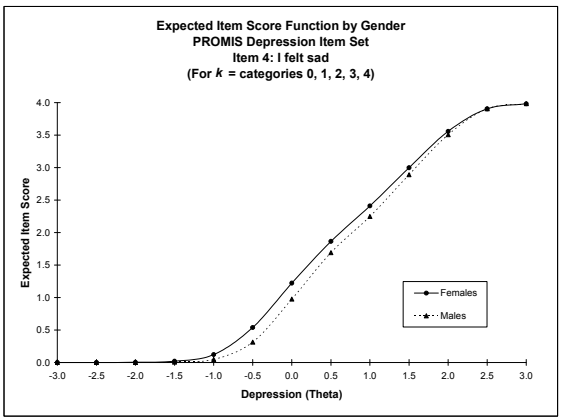

Expected Item Score Function by Age
PROMIS Depression Item Set
Item 4: I felt sad
(For k = categories 0, 1, 2, 3, 4)
Expected Item Score
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
-3.0 -2.5 -2.0 -1.5 -1.0 -0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
Depression (Theta)
Age 21-49
Age 50-64
Age 65-84

Figure 1: - cont.
PROMIS depression short form item set: Expected scale and item scores for gender subgroups





Sensitivity analyses

Sensitivity analyses were conducted to examine the effect of increasing the size of the anchor sets on the results. For all DIF analyses, the number of selected anchors was small. For the race/ethnicity analysis only two items showing no DIF were selected as anchor items: “I felt sad” and “I felt depressed”. Similarly for gender, three anchor items were selected: “I felt worthless”; “I felt I had nothing to look forward to”; and “I felt hopeless”. For the education groups, the following three anchors were selected: “I felt I had nothing to look forward to”; “I felt sad”; and “I felt depressed”. The only analysis with four or more anchor items was for age, with the following anchors: “I felt helpless”; “I felt like a failure”; “I felt depressed”; “I felt hopeless”; “I felt discouraged about the

future;" and "I felt disappointed in myself." Because of a small number of anchors selected for the majority of comparisons, the sensitivity DIF analysis was performed with four anchor items for each of the analyses for the race/ethnicity and gender demographic groups.

The DIF results did not change for gender; "I felt sad" showed DIF after the Bonferroni correction whether three or four anchors were included. The DIF results changed only minimally for race/ethnicity, all for the non-Hispanic Asians/Pacific Islanders compared to the non-Hispanic Whites as the reference group. The result for the item, "I felt worthless" changed from showing DIF after the Bonferroni correction to DIF not reaching that level. When four anchors were used, in addition to the original uniform DIF observed after the Bonferroni correction, non-uniform DIF was observed for the item, "I felt like a failure," but only before the Bonferroni correction. The item, "I felt hopeless" did not show DIF with four anchor items for Asians/Pacific Islanders.

Because local dependencies can result in over-identification of DIF, sensitivity analyses were performed for the depression analyses removing one item from a pair with high LD statistics across group comparisons: item 7 – "I felt unhappy" (which evidenced high LD statistics with the item 6 – "I felt depressed"). The results changed somewhat. Item parameters for "I felt worthless" and item 2 – "I felt that I had nothing to look forward to" comparing Whites with the Asians/Pacific Islanders were no longer significantly different after the Bonferroni correction. The type of DIF changed from uniform to non-uniform for the item, "I felt hopeless." Results for the item, "I felt discouraged about the future" were no longer significantly different after the Bonferroni correction for all comparisons. For the education comparison of respondents with graduate education to those with less than high school, for the item, "I felt worthless" the DIF results changed from significant after Bonferroni correction to just significant. There were no changes in DIF designations for age and gender comparisons.

Aggregate impact

As shown in Figure 1, there was no evident scale level impact. All group curves were overlapping for all comparisons.

Individual impact

Analyses were performed evaluating individual impact by comparing thetas estimated accounting for and not accounting for DIF. The analysis was limited to the race/ethnic and education subgroups. Individual impact for the race/ethnic groups was observed despite the high correlation of the two theta estimates (0.96). The changes were minimal (less than the absolute value of 0.5 standard deviations on the theta continuum) for 93 % of individuals. Although 356 individuals (7 % of the sample) were estimated with absolute change values greater than 0.5 standard deviations, only 36 of these changed greater than 1.0 standard deviation. Using an arbitrary cutoff point of theta $\geq$ 1.0 to classify respondents as depressed, there were 233 (4 %) of the total sample who changed from the designation of not depressed to depressed after the DIF adjustment and 86 (2 %) who

changed to not depressed. However, for only 2 patients (who changed to depressed) was the change in theta > 1.0 SD.

The variation in change among the race/ethnic groups was not large; for 90 % of non-Hispanic Blacks, the absolute difference was from 0.0 to 0.5 standard deviations; this was also true of 94 % of Hispanics, 94 % of non-Hispanic Whites and 95 % of non-Hispanic Asians/Pacific Islanders. The theta magnitude change from above to below the depression symptomatology threshold (theta ≥ 1.0) was observed for 20 (0.9 %) non-Hispanic Whites, 16 (1.4 %) non-Hispanic Blacks, 42 (4.0 %) Hispanics, and 8 (0.9 %) non-Hispanic Asians/Pacific Islanders. In contrast, 108 (4.8 %) Whites, 57 (5.1 %) Blacks, 22 (2.1%) Hispanics and 46 (5.1 %) Asians/Pacific Islanders changed from the non-symptomatic status to the designation indicative of depression symptomatology (change > 0.5 standard deviations).

The individual impact for the education groups was much smaller. The correlation of the two theta estimates was 1.0. All of the absolute values of the changes were less than 0.5 standard deviations. When considering the depression threshold designation, 39 (0.7 % of 5,386) patients changed designation in the direction of more symptomatology or above the threshold after the DIF adjustment. For no individual was the designation changed from the depression symptomatology to below the threshold.

Information

The item-level information functions were examined for the total sample (see Appendix, Figure 2.) As shown, the item, hopeless was estimated as most informative with the peak

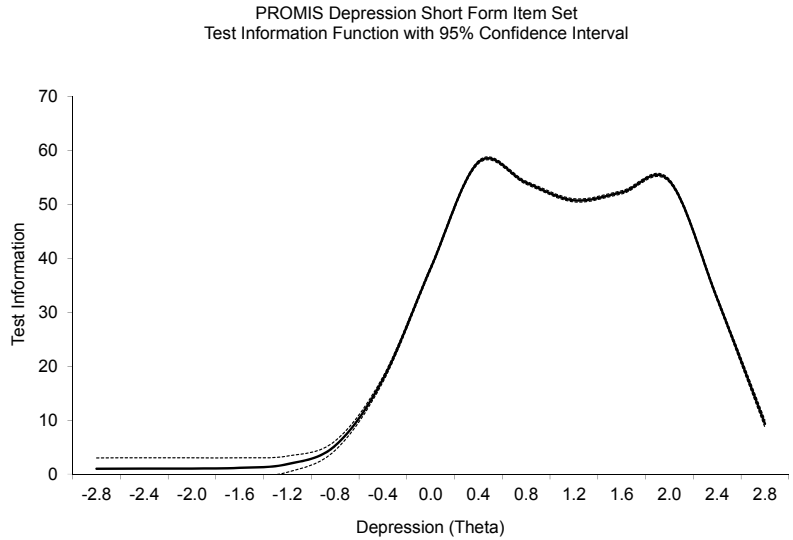


Figure 2:
PROMIS depression short form item set: Test information function (IRTPRO) (total sample)

information of 11.0 at theta level 0.4. The two items with the next highest peak information estimates were: "I felt that I had nothing to look forward to" and "I felt depressed" (information = 6.01 at theta = 0.8 and 6.0 at theta 0.4). The least informative items were: "I felt disappointed in myself" (information = 3.75 at theta = 0.4) and "I felt worthless" (information = 4.10 at theta = 0.8). Shown in Figure 2 is the scale-level information function. Most scale level information ranges in the middle and upper (depressed) tail of the distribution from theta level 0.4 to 2.0 with its peak of 57.9 at theta = 0.4. Peak item information is also provided in the middle upper (depressed) tail of the distribution ranging from theta = 0 to 0.8.

Discussion

The findings of DIF were examined in concert with the hypotheses tested. Only one item showed gender DIF. Conditional on depression, females were more likely to endorse the item, sad. This finding corresponded to the hypothesis that conditional on depression, women would express greater feelings of sadness. Other items hypothesized to show gender DIF: worthlessness, helplessness, depression, and unhappiness and nothing to look forward to were not found to evidence DIF after adjustment for multiple comparisons. However, the items helpless, depressed and unhappy were identified with DIF in the sensitivity analyses using IRTOLR. These items as well as helplessness did show significant DIF in the initial round of DIF testing, and were not included as anchor items. Thus, although they were not ultimately flagged with salient DIF, they did show DIF as hypothesized by the content experts.

It was hypothesized that conditional on depression, older people would be more likely to report feeling that there was nothing to look forward to and this hypothesis was corroborated by the findings. Conditional on depression, older respondents were more likely to respond in the depressed direction to this item. Age DIF was also found for the items, worthless and sad; however, in opposite directions, and no hypotheses were generated for these items.

For race/ethnicity, conditional on depression, it was hypothesized that minority groups as contrasted with the White majority would express more feelings of worthlessness, helplessness and failure. These three items were among the eight items with consistent DIF found by both IRT and OLR methods. Conditional on depression, Asians/Pacific Islanders (as contrasted with non-Hispanic Whites) evidenced a higher probability of responding in the depressed direction to the items: worthless, nothing to look forward to, helpless, failure, unhappy and disappointed. However, the sensitivity analyses modeling the local dependencies showed that two items: worthless and nothing to look forward to were not significant after Bonferroni adjustment. Non-Hispanic Blacks had a lower probability of endorsing the item, worthless, conditional on depression. Hispanics as contrasted with non-Hispanic Whites had a lower probability of item endorsement for the items: hopeless and disappointed; however, these items were not hypothesized to show DIF. Conditional on depression, the likelihood of endorsing the item unhappy was lower for the Hispanics vs. non-Hispanic Whites.

Previous studies also provided evidence of DIF in depressive symptom items across different racial/ethnic groups (e.g., Cole et al., 2000; Iwata & Buka, 2002; Iwata, Turner, & Lloyd, 2002; Kim, Chiriboga, & Jang, 2009; Yang et al., 2007), suggesting caution with respect to the interpretation of depressive symptoms among racial/ethnic minorities for clinical research and practice. For example, Iwata & Buka tested DIF for items included in the Center for Epidemiological Studies Depression (CES-D; Radloff, 1977) among White, Japanese, Native American, and Argentinean undergraduates. They reported that Japanese and Argentineans were more likely to inhibit endorsement of positive items such as hopeful, happy and enjoyed. As another example, Kim and colleagues (2009) also tested DIF for the CES-D items across three racial/ethnic groups of older adults (Mexican Americans, Blacks, and Whites) and provided evidence of the lack of measurement equivalence among Mexican Americans in comparison with Whites and Blacks. The comparison of non-Hispanic Whites to Mexican Americans resulted in the identification of 16 out of 20 items with DIF; two items were identified with DIF for the comparison of Whites and Blacks, both items related to interpersonal relations. Although the PROMIS Depression measure does not include somatic symptom items, the literature also suggests significant DIF observed for such items in the comparison of racial/ethnic minorities with non-minorities. Despite the negligible impact of DIF in the current analyses, given the findings of significant DIF in the comparisons of racial/ethnic groups for some items, cross-validation of results with different patient samples and additional ethnic groups is recommended in future research.

Conditional on depression, it was posited that those with lower education would express more feelings of being worthless and hopeless. The findings were of five items identified with DIF for education after adjustment for multiple comparisons (worthless, unhappy, hopeless, discouraged and disappointed). Only three were consistently identified by both the Wald test and the OLR procedure: worthless, hopeless and disappointed; two of these three were hypothesized to show DIF: worthless and hopeless. Conditional on depression, the item, worthless was more likely to be endorsed in the depressed direction by the patients with less than high school education vs. the patients with a graduate degree. However, in sensitivity analyses this item was not significant using the Bonferroni adjustment cutoff. The item, disappointed in myself showed DIF of higher magnitude for the graduate school vs. no high school groups. However, the NCDIF statistic was not above threshold, and the impact of DIF on the scale was trivial.

The analyses of language among Hispanic respondents did not identify any DIF, despite hypotheses that Spanish speakers would express greater feelings of worthlessness, helplessness and nothing to look forward to. Evidence from previous analyses of DIF identified an effect of acculturation on depressive symptom measures (Nguyen et al., 2007). Lower endorsement of somatic symptom items among low acculturated Hispanic women than in their high acculturated counterparts was identified. As stated earlier, the PROMIS Depression short form does not include somatic items; nonetheless future research should be performed examining potential DIF associated with acculturation and language.

Finally those with a diagnosis of cancer were posited to express greater feelings of worthlessness, helplessness, discouragement about the future and feeling like a failure, conditional on depression. No tests of DIF were performed for diagnosis.

Study limitations

The only analysis of language that was possible, given the subgroup sample sizes was Spanish. The examination of language DIF needs to be extended to other languages, including for example, Chinese, Portuguese, and Korean. Subgroups of Asians/Pacific Islanders and Hispanics have not been examined in the present DIF analyses. Thus, future research should consider examining potential differences across ethnic subgroups. Further investigation is needed to examine potential reasons for the DIF observed.

Summary

In summary, very little DIF of high magnitude or impact was observed. However, several items were identified that might require further study because there was a slightly higher magnitude of DIF and a correspondence of the DIF hypotheses to the findings of DIF. For gender, one item, sad, was both hypothesized and observed to show gender DIF in the direction of women reporting more sadness than men, conditional on depression. For age, one item showed slightly higher magnitude: nothing to look forward to; and this item was also hypothesized to show age DIF. Only one item showed DIF of higher magnitude (just above threshold) for Whites vs. non-Hispanic Asians/Pacific Islanders in the direction of higher likelihood of endorsement for Asians/Pacific Islanders: failure. This item was also hypothesized to show DIF for minority groups. Conditional on depression, the items, worthless and hopeless were more likely to be endorsed in the depressed direction by the patients with less than high school education vs. the patients with a graduate degree. These items were also hypothesized to show DIF in the direction of more feelings of worthlessness by groups with lower education. As noted, the magnitude and impact of DIF for all of these comparisons was relatively low.

These results could be useful for those using this scale in minority populations or to clinicians evaluating individuals, using the short form depression scale. Information provided was relatively high, particularly in the middle upper (depressed) tail of the distribution. Reliability estimates were high across all studied groups, regardless of estimation method. One potential caveat is that while aggregate impact was minimal, individual impact of relatively high magnitude ($\geq$1.0 standard deviation change in theta estimates before and after DIF adjustment) was observed for some, albeit a small number (36 or < 1 % of 5,324) of individuals for the race/ethnicity analysis, for 6 (0.6 %) Hispanics, 14 (1.3 %) non-Hispanic Blacks, 6 (0.7 %) non-Hispanic Asians/Pacific Islanders and 10 (0.4%) non-Hispanic Whites.

From a methodological perspective, the results illustrate some issues related to model assumptions and anchor item selection. Results can change if the numbers of anchor items are less than optimal, e.g. four; however, false DIF detection can result from inclu-

sion of items with DIF in the anchor. Thus, sensitivity analyses can inform about the extent to which results do not converge. In this case, most of the DIF results were similar. Local dependencies can result in discrimination parameter estimates that are too high; this was observed for the depression data set. However, sensitivity analyses, removing items with high LDs did not have an appreciable effect on the DIF results.

It is also noted that while the reliability estimates were high, the omega total value estimates ranged from 0.977 to 0.985, and were higher than were those for Cronbach's alpha, which ranged from 0.955 to 0.972. The relationship between alpha and omega total in the presence of unidimensionality is such that omega total is greater than or equal to alpha (Zinbarg, Revelle, Yovel, & Li, 2005). Thus, the higher value of omega total is congruent with the overwhelming evidence in support of unidimensionality.

Conclusions

The findings show superior psychometric properties of the PROMIS short form depression measure across several socio-demographic groups, and provide the first solid evidence regarding its performance among large samples of ethnically diverse groups. However, there was some evidence of DIF for some comparisons. There was a correspondence of the DIF hypotheses to the findings of DIF in a number of instances. While the magnitude and aggregate impact of DIF was small, in a few instances, individual impact was observed. Despite the low magnitude of DIF, clinicians or researchers working with patients from diverse cultural backgrounds and with lower educational levels should be alert to the potential for bias when items such as feeling worthless, hopeless and like a failure are used to measure depressive symptoms and diagnose clinical depression among racially/ethnically diverse groups and those with low education. Because the development of culturally appropriate screens for depression using the PROMIS Depression item bank is a high priority; this line of DIF research should be extended to other comparisons based on acculturation, chronic diseases and ethnicity.

Acknowledgements

Partial funding for these analyses was provided by the National Institute of Arthritis & Musculoskeletal & Skin Diseases, U01AR057971(PI: Potosky, Moinpour) and by the National Institute on Aging, 1P30AG028741-01A2 (PI, Siu). The authors thank Stephanie Silver, MPH for editorial assistance in the preparation of this manuscript.

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling, 16*, 397-438. doi:10.1080/10705510903008204

Azocar, F., Areán, P., Miranda, J., & Muñoz, R. F. (2001). Differential item functioning in a Spanish translation of the Beck Depression Inventory. *Journal of Clinical Psychology, 57*(3), 355-365. doi: 10.1002/jclp.1017

Baker, F. B. (1995). EQUATE 2.1: Computer program for equating two metrics in item response theory [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Benjamini, Y., & Hochberg, Y. (1995). Controlling for the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289-300. doi:10.2307/2346101 Key: citeulike:1042553

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. doi: 10.1037/0033-2909.107.2.238

Bjorner, J. B., Rose, M., Grandek, B., Stone, A. A., Junghaenel, D. U., & Ware, J. E. (2014). Difference in method of administration did not significantly impact item response: an IRT-based analysis from the Patient-Reported Outcomes Measurement Information System (PROMIS) initiative. *Quality of Life Research, 23*, 217-227. doi: 10.1007/s11136-013-0451-4

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 8*, 3-62.

Cai, L., Thissen, D., & du Toit SHC. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT Modeling [Computer software]. Chicago, IL: Scientific Software International, Inc.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage Publications.

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., … & Rose, M., on behalf of the PROMIS Cooperative Group. (2007). The patient-reported outcomes measurement information system (PROMIS): progress of an NIH roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3-S11. doi:10.1097/01.mlr.0000258615.42478.55.

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., & Sherbourne, C. D. (2004). The interview mode effect on the Center of Epidemiological Studies Depression (CES-D) scale: an item response theory analysis. *Medical Care, 42*(3), 281-289. doi: 10.1097/01.mlr.0000115632.78486.1f

Chen, W. H., Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289. doi: 10.2307/1165285

Choi, S.W., Gibbons, L. E., & Crane, P. K. (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*, 1-30. doi: 10.18637/jss.v039.i08

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research, 19*, 125-136. doi: 10.1007/s11136-009-9560-5

Cohen, A. S., Kim, S-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*, 15-26.doi: 10.1177/014662169602000102

Cohen, P., Cohen, J., Teresi, J., Marchi, P., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement, 14*, 183-196. doi: 10.1177/014662169001400207

Cole, S. R., Kawachi, I., Maller, S. R., & Berkman, L. F. (2000). Test of item-response bias in the CES-D scale: Experience from the New Haven EPESE Study. *Journal of Clinical Epidemiology, 53*, 285-9. doi:10.1016/S0895-4356(99)00151-1

Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*, 447-460. doi: 10.1007/s11136-009-9464-4

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S115-S123. doi: 10.1097/01.mlr.0000245183.28384.ed

Crane, P. K., van Belle, G., & Larson, E. B. (2004). Test bias in a cognitive test: differential item functioning in the CASI. *Statistics in Medicine, 23*, 241-256. doi: 10.1002/sim.1713.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.

DeWalt, D. A., Rothrock, N., Yount, S., & Stone, A. A. on behalf of the PROMIS cooperative group. (2007). Evaluation of item candidates: The PROMIS qualitative item review. *Medical Care, 45*(5 Suppl 1), S12-S21. doi: 10.1097/01.mlr.0000254567.79743.e2.

Fleer, P. F. (1993). A Monte Carlo assessment of a new measure of item and test bias. Illinois Institute of Technology. *Dissertation Abstracts International, 54*(04B), 2266.

Flowers, C. P., Oshima, T. C., & Raju, N. S. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*, 309-32.doi:10.1177/01466219922031437

Forkmann, T., Kroehne, U., Wirtz, M., Norra, C., Baumeister, H., Gaugett, S., Elhan, A. H., Tennant, A., & Becker, M. (2013). Adaptive screening for depression – Recalibration of an item bank for the assessment of depression in persons with mental and somatic diseases and evaluation in a simulated computer-adaptive test environment. *Journal of Psychosomatic Research, 75*, 437-443. doi:10.1016/j.jpsychores.2013.08.022

Grayson, D. A., Mackinnon, A., Jorm, A. F., Creasey, H., & Broe, G. A. (2000). Item bias in the Center for Epidemiologic Studies Depression Scale: effects of physical disorders and disability in an elderly community sample. *Journals of Gerontology: Psychological Sciences, 55B*(5), 273-282. doi: 10.1093/geronb/55.5.P273

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11 Suppl 3), S182-S188. doi: 10.1097/01.mlr.0000245443.86671.c4

Holland, P. H., & Wainer, H. (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement, 37*, 541-562. doi: 10.1177/0146621613491456

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

Iwata, N., Turner, R. J., & Lloyd, D. A. (2002). Race/ethnicity and depressive symptoms in community-dwelling young adults: a differential item functioning analysis. *Psychiatry Research, 110*, 281-289. doi:10.1016/S0165-1781(02)00102-6

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination: detecting differential item functioning using MIMIC modeling. *Medical Care, 44*(11 Suppl 3), S124-S133. doi: 10.1097/01.mlr.0000245250.50114

Kim, G., Chiriboga, D. A., & Jang, Y. (2009). Cultural equivalence in depressive symptoms in older White, Black, and Mexican-American adults. *Journal of the American Geriatrics Society, 57*(5), 790-796.

Kim, S., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*, 93-116. doi: 10.1111/j.1745-3984.2007.00029.x

Kim, S. H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22*, 345-355. doi: 10.1177/014662169802200403

Kim, Y., Pilkonis, P. A., Frank, E., Thase, M. E., & Reynolds, C. F. (2002). Differential functioning of the Beck Depression Inventory in late-life patients: use of item response theory. *Psychology and Aging, 17*(3), 379-391. doi: 10.1037/0882-7974.17.3.379

Kopf, J., Zeileis, A., & Stobl, C. (2015). Anchor selection strategies for DIF analysis: review, assessment and new approaches. *Educational and Psychological Measurement, 75*, 22-56. doi: 10.1177/0013164414529792

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling, 58.*

Langer, M. M. (2008). *A re-examination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation). University of North Carolina at Chapel Hill library, http://search.lib.unc.edu/search?R=UNCb5878458.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

McDonald, R.P. (1999). *Test theory: a unified treatment*. Mahwah, NJ: L. Erlbaum Associates.

Meade, A. W., Johnson, E. C., & Bradley, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*, 568-592. doi: 10.1037/0021-9010.93.3.568

Meade, A., Lautenschlager, G., & Johnson, E. (2007). A Monte Carlo examination of the sensitivity of the differential functioning of items and tests framework for tests of measurement invariance with Likert data. *Applied Psychological Measurement, 31*, 430-455. doi: 10.1177/0146621606297316

Meredith W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543. doi: 10.1007/BF02294825

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care,* 44, Suppl 3, S69-S77. doi: 10.1097/01.mlr.0000245438.73837.89

Millsap, R. E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement, 16*, 389-402. doi: 10.1177/014662169201600411

Mukherjee, S., Gibbons, L. E., Kristiansson, E., & Crane, P. K. (2013). Extension of an iterative hybrid ordinal logistic regression/item response theory approach to detect and account for differential item functioning in longitudinal data. *Psychological Test and Assessment Modeling, 55*, 127-147.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49*, 115-132. doi: 10.1007/BF02294210

Muthén, L. K., & Muthén, B. O. (2011). M-PLUS Users Guide. Sixth Edition. 1998-2011. Los Angeles, California: Muthén and Muthén.

Nguyen, H. T., Clark, M., & Ruiz, R. J. (2007). Effects of acculturation on the reporting of depressive symptoms among Hispanic pregnant women. *Nursing Research, 56*, 217-223. doi: 10.1097/01.NNR.0000270027.97983.96

Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: applications to the Mini-Mental State Examination. *Medical Care, 44*(11 Suppl 3), S134-S142. doi: 10.1097/01.mlr.0000 245251.83359.8c.

Oshima, T. C., Kushubar, S., Scott, J. C., & Raju, N. S. (2009). *DFIT for Window User's Manual: differential functioning of items and tests.* St. Paul, MN: Assessment Systems Corporation.

Pickard, A. S., Dalal, M. R., & Bushnell, D. M. (2006). A comparison of depressive symptoms in stroke and primary care: applying Rasch models to evaluate the Center for Epidemiologic Studies-Depression Scale. *Value in Health, 9*(1), 59-64. doi:10.1111/ j.1524-4733.2006.00082.x

Pilkonis, P. A., Choi, S.W., Reise, S. P., Stover, A. M., Riley, W. T., & Cella, D. (2011). Item banks for measuring emotional distress from the patient-reported outcomes measurement information system (PROMIS): Depression, Anxiety and Anger. *Assessment, 18*, 263-283. doi: 10.1177/1073191111411667

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

Raju, N. S. (1999). DFITP5: A Fortran program for calculating dichotomous DIF/DTF [Computer program]. Chicago: Illinois Institute of Technology.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M., L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the DFIT framework. *Applied Measurement in Education, 33*, 133-147. doi: 10.1177/0146621608319514

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.doi: 10.1177/014662169501900405.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Denmarks Paedagogiske Institut.

Reeve, B. (2000). *Item and scale level analysis of clinical and non-clinical sample responses to the MMPI-2 depression scales employing item response theory.* (Doctoral dissertation). The University of North Carolina at Chapel Hill, AAT 9968657.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., Thissen, D., … Cella, D. (2007). Psychometric Evaluation and Calibration of Health-Related Quality of Life Items Banks: plans for the Patient-Reported Outcome Measurement Information System (PROMIS). *Medical Care, 45*(5 Suppl 1), S22-S31.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research, 47*, 667-696. doi: 10.1080/00273171.2012.715555

Reise, S. P., Moore, T. M., Haviland, M. G. (2010). Bi-factor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544-559. doi: 10.1080/00223891.2010.496477

Reise, S., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*(Supp l1), 19-31. doi: 10.1007/s11136-007-9183-7

Rizopoulus, D. (2009). ltm: Latent Trait Models under IRT. http://cran.r project.org/web/packages/ltm/index.html .

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 34*, 100-114. doi: 10.1007/BF02290599

Schalet, B. D., Pilkonis, P. A., Yu, L., Dodds, N., Johnston, K. L., Yount, S., Riley, W., & Cella, D. (in press). Clinical validity of PROMIS depression, anxiety, and anger across diverse clinical samples. *Journal of Clinical Epidemiology*. doi: 10.1016/j.jclinepi.2015.08.036

Schmid, L., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61. doi: 10.1007/BF02289209

Seybert, J., & Stark, S. (2012). Iterative linking with the differential functioning of items and tests (DFIT) method: comparison of testwide and item parameter replication (IPR) critical values. *Applied Psychological Measurement, 36*(6), 494-515. doi: 10.1177/0146621612445182

Shealy, R. T., & Stout, W. F. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194. doi: 10.1007/BF02294572

Shealy, R. T., & Stout, W. F. (1993b). An item response theory model for test bias and differential item functioning. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (197-239). Hillsdale, NJ: Lawrence Erlbaum.

Shih, C. –L., & Wang, W. -C. (2009). Differential item functioning detection using multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement, 33*, 184-199. doi: 10.1177/0146621608321758.

Sijtsma K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120. doi: 10.1007/s11336-008-9101-0

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370 doi: 10.1111/j.1745-3984.1990.tb00754.x

Teresi, J. A., & Golden, R. (1994). Latent structure methods for estimating item bias, item validity and prevalence using cognitive and other geriatric screening measures. *Alzheimer Disease & Associated Disorders, 8*(Suppl), S291-S298.

Teresi, J.A. & Jones, R.N. (2016). Methodological Issues in Examining Measurement Equivalence in Patient Reported Outcomes Measures: Methods Overview to the Two-Part Series, "Measurement Equivalence of the Patient Reported Outcomes Measurement Information System (PROMIS) Short Form Measures" *Psychological Test and Assessment Modeling, 58*.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine, 19*, 1651-1683. doi: 10.1002/(SICI)1097-0258(20000615/30)19:11/12<1651::AID-SIM453>3.0.CO;2-H

Teresi, J., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. E., Crane, P. K., Jones, R. N., … Cella, D. (2009). 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, 51*, 148-180. NIHMSID#136951

Teresi, J.A., Ramirez, M., Lai, J-S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly, 50*, 538-612.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond group-mean differences: the concept of item bias. *Psychological Bulletin, 99*, 118-128. doi: 10.1037/0033-2909.99.1.118

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false discovery rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*, 77-83. doi: 10.3102/10769986027001077

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (123-135). Hillsdale, NJ: Lawrence Erlbaum, Inc.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differnces in trace lines. In H. Wainer, and H. Braum (Eds.), *Test Validity*. Hillsdale, NJ: Lawrence Erlbaum, Inc.

van de Vijver, F., & Leung, K. (1997). *Methods and Data Analyses for Cross-cultural Research.* Thousand Oaks, California: Sage Publications.

Wainer, H. (1993). Model-based standardization measurement of an item's differential impact. In P. W. Holland, & H. Wainer (Eds.). *Differential Item Functioning* (pp. 123-135). Hillsdale NJ: Lawrence Erlbaum, Inc.

Wang W. C, Shih C. L, & Sun G. W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement, 72*, 687-708. doi: 10.1177/0013164411426157

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with likelihood ratio test. *Applied Psychological Measurement, 27*, 479-498. doi: 10.1177/0146621603259902

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42-57. doi: 10.1177/0146621607314044.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532-547. doi: 10.1177/0013164412464875.

Yang, F. M., & Jones, R. N. (2007). Center of Epidemiologic Studies-Depression scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *Journal of Clinical Epidemiology, 60*, 1195-1200. doi: 10.1016/j.jclinepi.2007.02.008

Yost, K. J., Eton, D. T., Garcia, S. F., & Cella, D. (2011). Minimally important differences were estimated for six PROMIS cancer scales in advanced-stage cancer patients. *Journal of Clinical Epidemiology, 64*(5), 507-516. doi: 10.1016/j.jclinepi.2010.11.018

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β and McDonald's ω_H: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123-133. doi: 10.1007/s11336-003-0974-7

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, Canada: Directorate of Human Resources Research and

Evaluation, Department of National Defense. Retrieved from http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29.