Psychology Science Quarterly, Volume 51, 2009 (2), pp. 148 - 180

Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach

JEANNE A. TERESI[1,2], KATJA OCEPEK-WELIKSON[2], MARJORIE KLEINMAN[1], JOSEPH P. EIMICKE[2], PAUL K. CRANE[3,4]; RICHARD N. JONES[5,6], JIN-SHEI LAI[7], SEUNG W. CHOI[8], RON D. HAYS[9], BRYCE B. REEVE[10], STEVEN P. REISE[11], PAUL A.PILKONIS[12], DAVID CELLA[13]

Abstract

The aims of this paper are to present findings related to differential item functioning (DIF) in the Patient Reported Outcome Measurement Information System (PROMIS) depression item bank, and to discuss potential threats to the validity of results from studies of DIF. The 32 depression items studied were modified from several widely used instruments. DIF analyses of gender, age and education were performed using a sample of 735 individuals recruited by a survey polling firm. DIF hypotheses were generated by asking content experts to indicate whether or not they expected DIF to be present, and the direction of the DIF with respect to the studied comparison groups. Primary analyses were conducted using the graded item response model (for polytomous, ordered response category data) with likelihood ratio tests of DIF, accompanied by magnitude measures. Sensitivity analyses were performed using other item response models and approaches to DIF detection. Despite some caveats, the items that are recommended for exclusion or for separate calibration were "I felt like crying" and "I had trouble enjoying things that I used to enjoy." The item, "I felt I had no energy," was also flagged as evidencing DIF, and recommended for additional review. On the one hand, false DIF detection (Type 1 error) was controlled to the extent possible by ensuring model fit and purification. On the other hand, power for DIF detection might have been compromised by several factors, including sparse data and small sample sizes. Nonetheless, practical and not just statistical significance should be considered. In this case the overall magnitude and impact of DIF was small for the groups studied, although impact was relatively large for some individuals.

Key words: patient reported outcomes measurement information system; item response theory; differential item functioning; depression

[1] Columbia University Stroud Center; Faculty of Medicine, New York State Psychiatric Institute.
Correspondence should be addressed to: Jeanne A. Teresi, Ed.D., Ph.D., Research Division, HHAR, 5901 Palisade Avenue, Riverdale, New York 10471, USA; email: teresimeas@aol.com
[2] Research Division, Hebrew Home for the Aged at Riverdale, NY, USA
[3] Department of Rehabilitation Medicine, University of Washington, Seattle, WA, USA
[4] Department of Internal Medicine, University of Washington, Seattle, WA, USA
[5] Institute for Aging Research, Hebrew SeniorLife, Boston, Massachusetts, USA
[6] Division of Gerontology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, Massachussetts, USA
[7] Department of Medical Social Sciences, Northwestern University, Feinberg School of Medicine, Evanston, Il, USA
[8] Department of Medical Social Sciences, Northwestern University, Feinberg School of Medicine, Evanston, Il, USA
[9] University of California, Los Angeles; RAND Health Program, Los Angeles, California, USA
[10] National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA
[11] University of California, Los Angeles, Los Angeles, California, USA
[12] University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[13] Department of Medical Social Sciences, Northwestern University, Feinberg School of Medicine, Evanston, Il, USA

Background

Surveys that assess latent traits or states such as attitudes, affect, and health often use scales in order to increase the likelihood of accurate measurement. Conceptual and psychometric measurement equivalence of such scales are basic requirements for valid cross-cultural and demographic subgroup comparisons. Differential item functioning (DIF) analysis, commonly used to study the performance of items in scales, examines whether or not the likelihood of item (category) endorsement is equal across subgroups that are matched on the state or trait measured. Two basic types of DIF examined are uniform and non-uniform. Uniform DIF implies that one group is consistently more likely than another to endorse an item at each level of the trait or state, e.g., depression. Non-uniform DIF is observed when there is cross-over, so that at certain levels of the state or trait, one group is more likely to endorse the item, while at other levels, the other group is more likely to endorse the item (see also the Glossary).

DIF analyses of depression scales: The underlying state for the data presented in this paper is depression, and the items are scored in the impaired direction, reflecting depression symptomatology. Socio-cultural and health-related factors appear to affect the response patterns as well as the factorial composition of scales assessing depressive symptomatology (Mui, Burnette & Chen, 2001; Pedersen, Pallay & Rudolph, 2002). The variety of depression scales, the differences in the methodology, and the diversity in group variables examined for DIF across the reviewed studies make the synthesis of findings a difficult task. However, consistent findings across the articles suggest the presence of differential functioning in a substantial number of CES-D (Radloff, 1977) items as a function of a variety of sociodemographic and health-related variables. For example, the authors of several studies found that one or both of the interpersonal items, "people are unfriendly" and "people dislike me" showed DIF with respect to one or more of several variables: race, physical disorder, stroke, and interview mode (Chan, Orlando, Ghosh-Dastidar & Sherbourne, 2004; Cole, Kawachi, Maller & Berkman, 2000; Grayson, Mackinnon, Jorm, Creasey & Broe, 2000; Pickard, Dalal & Bushnell, 2006; Yang & Jones, 2007). Similarly, the affective CES-D item tapping sadness showed DIF based on physical disorder and interview mode (Grayson et al.; Chan et al.). DIF was also observed in the "crying" items (contained in the CES-D, and most other depression scales) 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 in the CES-D was found to be substantial in some studies (Chan, et al., 2004; Cole et al., 2000), but less so in others (Pickard et al. 2006). Low impact of DIF was also observed in the Hospital Anxiety and Depression Scale (Zigmond & Snaith, 1983) in one study of breast cancer patients (Osborne, Elsworth, Sprangers, Oort & Hopper, 2004). The impact of DIF in the Beck Depression Inventory (BDI) (Beck, Ward, Mendelsohn, Mock & Erbaugh, 1961) was demonstrated to be sizable, with artificially inflated scores for Latinos (in contrast to English speakers) (Azocar et al. 2001). Similarly, another analysis demonstrated that half of the items on the BDI scale accounted for 80% of the differential test (scale) functioning, and item response theory (IRT)-adjusted cutoff scores reduced considerably the false negative rate of clinically diagnosed patients with depression who would have been classified as non-depressed without DIF adjustment (Kim, Pilkonis, Frank, Thase & Reynolds, 2002). As is demonstrated by these studies, the findings of salient DIF in many

depression measures underscore the need for examination of DIF in items measuring depression. A more detailed review can be found in Teresi, Ramirez, Lai and Silver (2008).

Aims

The aims of this paper are to present findings related to DIF in the Patient Reported Outcome Measurement Information System (PROMIS) (Cella et al. 2007; Reeve et al. 2007) depression item bank, and to discuss strengths and limitations of approaches used in DIF detection analyses. Analyses of gender, age and education were performed. Sample sizes were insufficient to examine race/ethnicity.

Methods

Sample generation and description

The overall sample is discussed in Liu et al. (under review). However, a brief description of the subsamples used for the primary and sensitivity analyses is presented here. These data are from individuals who were administered the full bank of emotional distress items; data were collected from a survey panel by a polling firm, Polimetrix (www.polimetrix.com; www.pollingpoint.com).

The studied (also called the focal) group was females in the analyses of gender; the sample sizes for the groups were 379 females and 356 males. In the analyses of education, the studied group was low education through some college (n=518), and the reference group was college or advanced degree (n=217). The studied group for age was those 65 and over (n=201); the sample size for the younger reference group was 533. The sensitivity analyses sample sizes were 258 for the group aged 60 and over, and 476 for the group under age 60.

Measures

Depressive symptoms was a subdomain of emotional distress, and the 32 depression items studied were modified from several instruments, including two items from the Geriatric Depression Scale (Yesavage et al. 1982), one from the BDI (Beck et al. 1961), four from the CES-D (Radloff, 1977), and three from the Medical Outcomes Study (Stewart, Ware, Sherbourne & Wells, 1992). Other items came from an assortment of sources. It is noted that many of the items are quite similar to those from popular and older scales used cross-nationally, such as the Depression scale from the Geriatric Mental State or the Comprehensive Evaluation and Referral Examination (CARE) (Gurland et al. 1976; Copeland et al. 1976; Golden, Teresi & Gurland, 1984), and the most recent rendition of these instruments, the EURO-D (Prince et al. 1999). The timeframe for all items was the past 7 days. Items were administered using a five point response scale: 'never', 'rarely', 'sometimes', 'often' and 'always'. Because of sparse data, the categories, 'often' and 'always' were collapsed, resulting in four ordinal response categories for the preliminary analyses; however, the final analyses required collapsing 'sometimes', 'often' and 'always', resulting in three categories. Sensitivity

analyses of binary data were conducted, collapsing 'never' and 'rarely' vs. the other categories.

Procedures and statistical approach

Qualitative analyses and hypotheses generation: Extensive qualitative analyses, including focus groups and cognitive interviews were performed prior to data collection. Based on these data, the items were modified for use in PROMIS in order to refer to the same time frame, have the same response options, and target a 6th grade reading level (see DeWalt, Rothrock, Yount & Stone, 2007). Thirteen focus groups with 104 participants, largely from outpatient psychiatric clinics were convened (DeWalt et al.). Individuals were selected to be representative of a variety of chronic diseases, cultures and ages. Cognitive interviews were conducted; examined were the meaning of the item, the recall and decision process, including social desirability, and the response process. The protocol was based on that of Willis (2005). All questions were first completed by respondents using a paper-and-pencil format, followed by cognitive interviews with probes to elicit the information; five cognitive interviews were performed for each item (see DeWalt et al.).

DIF hypotheses were generated 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 race/ethnicity were also elicited, but subgroup sample sizes were not sufficient for DIF analyses.) A definition of DIF was provided, and the following instructions related to hypotheses generation were given. "Differential item functioning means that individuals in 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 same (e.g., mild) level of depression should have the same likelihood of reporting crying often. If it is hypothesized that this is not the case, it would be hypothesized that the item has gender DIF." Forms were developed for this purpose, and completed by 11 individuals (four clinical health psychologists, one clinical psychologist, two psychiatrists, one oncology nurse, and three "other" professionals). A summary table (available from the authors) was developed arraying the hypotheses and findings from the literature.

Quantitative analyses and tests of DIF hypotheses: Prior to any formal tests of DIF, a best practice recommended by Hambleton (2006) was used to examine the data at a basic, descriptive level. Ten equal intervals of the sum score were formed based on the focal group sums. The item means were examined for each group within each of the levels, and tested for significant group differences. Such "fat matching" (Dorans & Kulick, 2006) is not ideal; however sparse data precluded finer distinctions.

Item response theory (IRT) is often used in DIF analyses (see Hambleton, Swaminathan & Rogers, 1991; Lord, 1980; Lord & Novick, 1968). The method used for DIF detection that is described in this paper was the IRT log-likelihood ratio (IRTLR) approach (Thissen, 1991, 2001; Thissen, Steinberg & Gerard, 1986; Thissen Steinberg & Wainer, 1993), accompanied by magnitude measures (Teresi, Kleinman & Ocepek-Welikson, 2000). DIF magnitude was

assessed using the non-compensatory DIF (NCDIF) index (Raju, van der Linden & Fleer, 1995; Flowers, Oshima & Raju, 1999). Finally, scale level impact was assessed using expected scale scores, expressed as group differences in the total test (scale) response functions. These latter functions show the extent to which DIF cancels at the scale level (DIF cancellation). The findings presented here focus on IRTLR; however, other methods were also used in sensitivity analyses to examine DIF in this item set. These other methods include SIBTEST (Shealy & Stout, 1993a,b) for binary items and Poly-SIBTEST (Chang, Mazzeo & Roussos, 1996) for polytomous items. SIBTEST is non-parametric, conditioning on the observed rather than latent variable, and does not detect non-uniform DIF.

A second method used in sensitivity analyses was logistic regression (Swaminathan and Rogers, 1990) and ordinal logistic regression (OLR) (Zumbo, 1999; Crane, van Belle & Larson 2004), which typically condition 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 3 vs. 2; then uniform DIF is tested by examining the incremental effect of 2 vs. 1, with a chi-square (1 d.f.) test (Camilli & Shepard, 1994). A modification, IRTOLR (Crane et al. 2004; Crane, Gibbons, Jolley and van Belle, 2006) 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. Finally, also used was the multiple indicators, multiple causes (MIMIC) approach (Jöreskog & Goldberger, 1975; Muthén, 1984), which is a parametric model with conditioning on a latent variable; while related to IRT, the model comes from the tradition of factor analyses and structural equation modeling, and does not test for non-uniform DIF (see also Jones, 2006).

Description of the model

The following analyses were conducted using the graded (for polytomous, ordered response category) item response model (Samejima, 1969). Sensitivity analyses were performed using a two parameter logistic model (for items that were collapsed into two categories, non-symptomatic and symptomatic). The graded response model is given in the glossary under IRT.

The expectation is that respondents who are depressed would be more likely than those who are not depressed to respond in a symptomatic direction to an item measuring depression. Conversely, a person without depression is expected to have a lower probability (than a person with depression) of responding in a depressed direction to the item. The curve that relates the probability of an item response to the underlying state or trait, e.g., depression, measured by the item set is known as an item characteristic curve (ICC). This curve can be characterized by two parameters in some forms of the model: a discrimination parameter (denoted *a*) that is proportional to the slope of the curve, and a location (also called threshold, difficulty, or severity) parameter (denoted *b*) that is the point of inflection of the curve. (See also the Glossary for definitions.) According to the IRT model, an item shows DIF if people from different subgroups but at the same level of depression have unequal probabili-

ties of endorsement. Put another way, the absence of DIF is demonstrated by ICCs that are the same for each group of interest.

IRT log-likelihood ratio (IRTLR) modeling: IRTLR, the DIF detection procedure used in these analyses is based on a nested model comparison approach (Thissen et al. 1993). First, a compact (or more parsimonious) model is tested with all parameters constrained to be equal across groups for a studied item (together with the anchor items defined below and in the Glossary) (model 1), against an augmented model (model 2) with one or more parameters of the studied item freed to be estimated distinctly for the two groups. The procedure involves comparison of differences in log-likelihoods (-2LL) (distributed as chi-square) associated with nested models; the resulting statistic is evaluated for significance with degrees of freedom equal to the difference in the number of parameter estimates in the two models. For the graded response model, the degrees of freedom increase with the number of *b* (difficulty or severity) parameters estimated. (There is one less *b* estimated than there are response categories.) Severity (*b)* parameters are interpreted as uniform DIF only if the tests of the *a* parameters are not significant; in that case, tests of *b* parameters are performed, constraining the *a* parameters to be equal. The final p values are adjusted using Bonferroni (Bonferroni, 1936) or other methods such as Benjamini-Hochberg (B-H) (Benjamini & Hochberg, 1995; Thissen, Steinberg & Kuang, 2002).

Tests of model assumptions and fit: Important first steps (not presented here) in the analyses include examination of model assumptions such as unidimensionality (see Reise, Morizot & Hays, 2007). These analyses were conducted prior to release of these data sets for DIF analyses, and provided evidence of essential unidimensionality. A standardized residual measure of goodness-of-fit, defined as the difference between the observed and expected frequency divided by the square root of the expected frequency for each response pattern associated with a particular level of the underlying state or trait (denoted theta), measured by the scale was calculated. The standardized residual is distributed approximately normally with mean of 0 and σ^2 of 1. High values are indicative of poor fit.

Anchor items: If no prior information about DIF in the item set is available, initial DIF estimates can be obtained by treating each item as a "studied" item, while using the remainder as "anchor" items. Anchor items are assumed to be without DIF, and are used to estimate theta (depression state level), and to link the two groups compared in terms of depression state level. Anchor items are selected by first comparing a model with all parameters constrained to be equal between two comparisons groups, including the studied item, and a model with separate estimation of all parameters for the studied item. This process of log-likelihood comparisons is performed iteratively, and is described in detail in Orlando-Edelen, Thissen, Teresi, Kleinman, and Ocepek-Welikson (2006).

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. A method for quantification of the difference in the average expected item scores is the non-compensatory DIF index (Raju and colleagues,

1995) used in DFIT (Oshima, Kushubar, Scott, Raju, 2009; Raju, Fortmann-Johnson, Kim, Morris, Nering, & Oshima, 2009). While chi-square tests of significance are available, these were found to be too stringent, over identifying DIF. 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 are 0.006 for binary items, and 0.024 and 0.054 for polytomous items with three or four response options (after collapsing categories due to sparse data) (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 three response categories, the recommended cutoff of 0.024 would correspond to an average absolute difference greater than 0.155 (about 0.16 of a point) on a three point scale (see Raju, 1999; Meade, Lautenschlager & Johnson, 2007). Because of the sensitivity of cutoff thresholds to the distribution of parameter estimates, simulations to derive cutoffs based on empirical distributions have been incorporated into the latest versions of software such as DFIT (Raju, Fortmann-Johnson, Kim, Morris, Nering, & Oshima, 2009) and ordinal logistic regression (Choi, Gibbons & Crane, 2009). The issue is what difference is meaningful and makes a practical difference. Recent work on effect sizes is presented in Stark, Chernyshenko & Drasgow (2004); Steinberg & Thissen (2006); and Kim, Cohen, Alagoz & Kim (2007).

Evaluation of DIF impact

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.

Software and procecures

Software used was IRTLRDIF (Thissen, 2001) and MULTILOG (Thissen, 1991). Additionally, NCDIF (Raju et al. 1995; Flowers et al. 1999) was evaluated using DFITP5 (Raju, 1999). Prior to application of the DFIT software, estimates of the latent state or trait (theta) are usually calculated separately for each group, and equated together with the item parameters. Baker's (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 the same metric. If DIF is detected, the item showing DIF is excluded from the equating algorithm, and new DIF-free equating constants are computed, and purified iteratively.

Results

The results of preliminary analyses of group item means within sum score intervals provided information used in collapsing categories for use in more formal tests of DIF. The analyses showed that (a) the data were sparse and skewed; (b) the distributions were different for age groups; (c) a few items emerged as likely candidates for flagging, with findings consistent with the formal DIF analyses using the various methods.

Model fit: Examination of standardized residuals (not shown) showed that most items fit the IRT model for the age and education analyses (after collapsing categories from four to three). Collapsing categories to three resolved the problem of sparse data for most analyses; however cell sizes were relatively small (12-20 in the symptomatic category above "none/never") for the high education group for some items, e.g., "I felt I had no reason for living." The three category solution resulted in all items fitting for the low education group except for slight misfit in category 3 for items 4 and 24 (z=2.13). For high education, some misfit was evident with z scores ranging from 2 to 4. For age, no misfit was observed among older subjects (z=-1.01 to 1.47). For younger subjects, some misfit was observed, with z scores ranging from 2 to 3. For gender, all items fit the IRT model with four response categories for females (z=-1.20 to 1.47); however, some misfit was observed for males. Reduction to three categories reduced the sparse data and misfit; however, some items still evidenced relatively high misfit (z=3.00 to 5.00).

Gender: Shown in Table 1 are the final item parameters and DIF tests for gender. The most severe indicator of depression was "no reason for living"; among the least severe indicators were "I felt that I had no energy", and "felt lonely". Females were more depressed; the estimated mean was -0.27 for females, and -0.55 for males, indicating that the difference between the average depression levels for women and men was about one fourth of a standard deviation. As shown, eleven items evidenced gender DIF prior to adjustment for multiple comparisons, two with non-uniform DIF. Items with non-uniform DIF were: "felt helpless" and "sad". Uniform DIF was evident for the items: "crying", "nothing could cheer me up", "people did not understand me", "trouble feeling close to people", "depressed", "unhappy", "nothing interesting", "life was empty", "trouble enjoying things I used to do". After adjustment for multiple comparisons, using either the Benjamini-Hochberg (1995) or Bonferroni (1936) correction, only the item "I felt like crying" showed uniform DIF; the NCDIF index for this item was above the cutoff (0.074). The item is a more severe indicator for males; it takes higher levels of depression for men to endorse the item. The magnitude of DIF for this item is shown in Figure 1.

Education: The most severe indicators for the education analyses were: "no reason for living", "worthless", "helpless", "nothing cheers me up", "wanted to give up". The estimated mean for the depression state for the low education group was somewhat higher than that of the high education group (-0.38 vs -0.49). Four items were found to have education-related DIF prior to Bonferroni/Benjamini-Hochberg correction, two with non-uniform DIF, "felt hopeless", and "pessimistic". After adjustment, no items evidenced education-DIF (see Table 2). Overall, the magnitude of DIF was small (see Table 4), and only one item had NCDIF above the cutoff, "I felt that I had no energy". This item was a more severe indicator for those with higher education (see Figure 1).

Table 1:
Item parameters* and standard errors for the anchor items and studied items with DIF from the depression item bank: Comparison of gender groups

Content	Group	*a*	*B1*	*b2*	*a*DIF	*b*DIF‡
I felt that I had no energy	Female Male	1.69 (0.15)	-01.21 (0.09)	-0.12 (0.08)	NS, Anchor Item	
I felt worthless	Female Male	3.49 (0.29)	0.21 (0.05)	0.88 (0.07)	NS, Anchor Item	
I felt that I had nothing to look forward to	Female Male	2.96 (0.26)	0.07 (0.05)	0.73 (0.07)	NS, Anchor Item	
I felt helpless	Female Male	3.04 (0.34) 3.85 (0.53)	0.04 (0.07) 0.13 (0.07)	0.80 (0.09) 0.76 (0.09)	4.4 (0.036)	0.8 (0.670)
I withdrew from other people	Female Male	2.31 (0.19)	-0.25 (0.06)	0.50 (0.08)	NS, no DIF	
I felt that nothing could cheer me up	Female Male	3.27 (0.27) 3.27 (0.27)	0.12 (0.07) 0.02 (0.07)	0.94 (0.09) 0.65 (0.09)	0.3 (0.584)	8.8 (0.012)
I felt that other people did not understand me	Female Male	2.33 (0.17) 2.33 (0.17)	-0.60 (0.08) -0.74 (0.08)	0.20 (0.09) 0.04 (0.09)	<0.001 (>0.999)	6.3 (0.043)
I felt that I was not as good as other people	Female Male	2.41 (0.21)	-0.10 (0.06)	0.67 (0.08)	NS, Anchor Item	
I felt like crying	Female Male	1.94 (0.15) 1.94 (0.15)	-0.46 (0.09) 0.04 (0.10)	0.54 (0.10) 1.04 (0.13)	1.1 (0.317)	***27.6 (<0 .001)***
I felt sad	Female Male	2.50 (0.29) 3.39 (0.38)	-0.82 (0.08) -0.80 (0.07)	0.08 (0.09) 0.19 (0.08)	4.3 (0.038)	5.1 (0.078)
I felt that I wanted to give up on everything	Female Male	3.10 (0.29)	0.35 (0.06)	1.00 (0.08)	NS, Anchor Item	
I felt that I was to blame for things	Female Male	2.47 (0.20)	-0.29 (0.06)	0.58 (0.07)	NS, Anchor Item	
I felt like a failure	Female Male	3.37 (0.30)	-0.07 (0.05)	0.55 (0.06)	NS, no DIF	
I had trouble feeling close to people	Female Male	2.65 (0.19) 2.65 (0.19)	-0.39 (0.08) -0.48 (0.08)	0.42 (0.08) 0.20 (0.08)	0.7 (0.403)	7.9 (0.019)
I felt disappointed in myself	Female Male	2.76 (0.22)	-0.65 (0.05)	0.17 (0.06)	NS, Anchor Item	

I felt that I was not needed	Female	2.67 (0.22)	-0.05 (0.06)	0.72 (0.07)	NS, Anchor Item	
	Male					
I felt lonely	Female	2.36 (0.19)	-0.39 (0.06)	-0.37 (0.07)	NS, Anchor Item	
	Male					
I felt depressed	Female	3.44 (0.28)	-0.41 (0.04)	0.39 (0.06)	0.7 (0.403)	6.4 (0.041)
	Male					
I had trouble making decisions	Female	2.19 (0.18)	-0.45 (0.06)	0.60 (0.07)	NS, Anchor Item	
	Male					
I felt discouraged about the future	Female	2.54 (0.20)	-0.56 (0.06)	0.14 (0.06)	NS, Anchor Item	
	Male					
I found that things in my life were overwhelming	Female	2.61 (0.20)	-0.39 (0.06)	0.41 (0.07)	NS, Anchor Item	
	Male					
I felt unhappy	Female	3.71 (0.29)	-0.64 (0.06)	-0.26 (0.07)	1.4 (0.237)	12.7 (0.002)
	Male	3.71 (0.29)	-0.78 (0.06)	0.06 (0.07)		
I felt I had no reason for living	Female	2.50 (0.28)	0.82 (0.07)	1.50 (0.12)	NS, Anchor Item	
	Male					
I felt hopeless	Female	3.94 (0.39)	0.13 (0.05)	0.86 (0.06)	NS, Anchor Item	
	Male					
I felt ignored by people	Female	2.24 (0.19)	-0.35 (0.06)	0.54 (0.07)	NS, Anchor Item	
	Male					
I felt upset for no reason	Female	2.40 (0.21)	-0.01 (0.06)	0.81 (0.08)	NS, Anchor Item	
	Male					
I felt that nothing was interesting	Female	2.38 (0.16)	-0.12 (0.09)	0.76 (0.09)	<0.001 (>0.999)	7.9 (0.019)
	Male	2.38 (0.16)	-0.33 (0.08)	0.72 (0.10)		
I felt pessimistic	Female	2.43 (0.19)	-0.62 (0.06)	0.24 (0.06)	NS, Anchor Item	
	Male					
I felt that my life was empty	Female	2.98 (0.23)	0.09 (0.08)	0.64 (0.08)	0.8 (0.371)	6.6 (0.037)
	Male	2.98 (0.23)	-0.03 (0.07)	0.45 (0.09)		
I felt guilty	Female	2.12 (0.18)	-0.20 (0.06)	0.70 (0.08)	NS, Anchor Item	
	Male					
I felt emotionally exhausted	Female	2.77 (0.22)	-0.51 (0.05)	0.27 (0.06)	NS, Anchor Item	
	Male					
I had trouble enjoying things that I used to enjoy	Female	2.41 (0.18)	-0.30 (0.09)	0.58 (0.08)	0.4 (0.527)	11.9 (0.003)
	Male	2.41 (0.18)	-0.43 (0.08)	0.28 (0.09)		

* Parameter estimates are from the final MULTILOG run

‡ DIF significant after Bonferroni adjustment is bolded; NCDIF items are italicized.

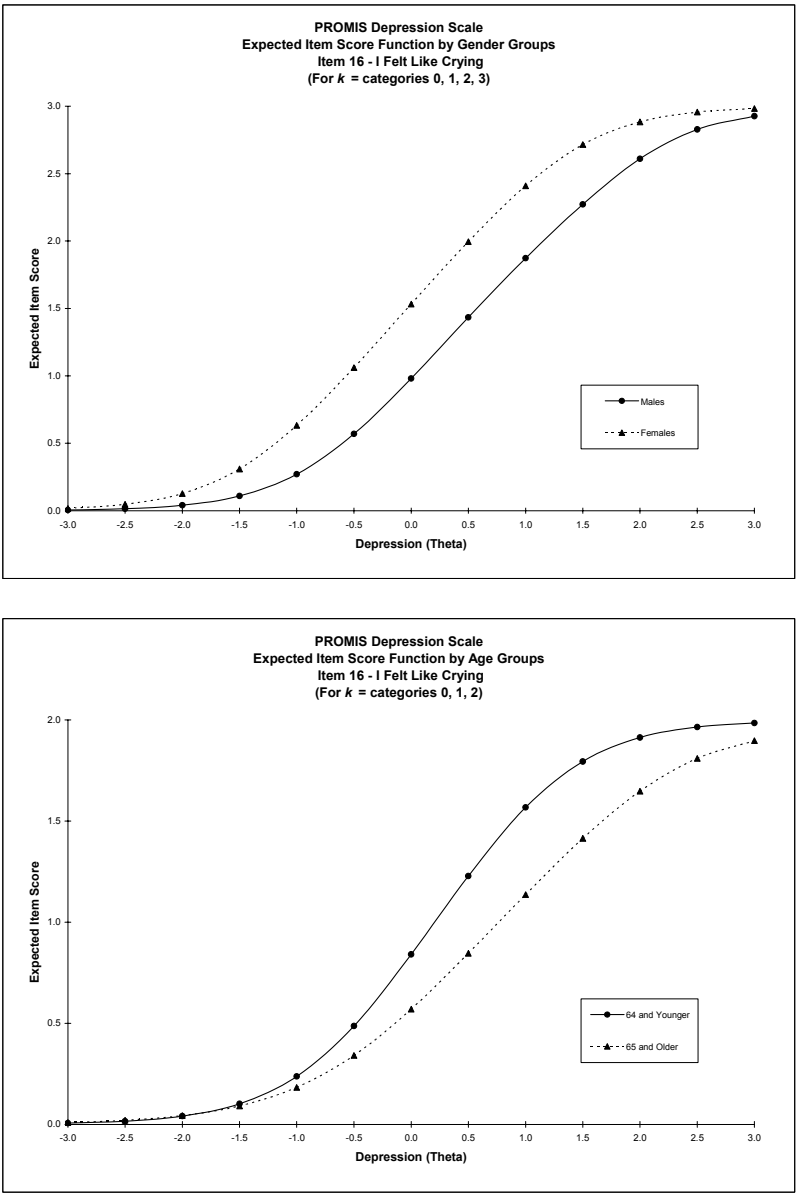
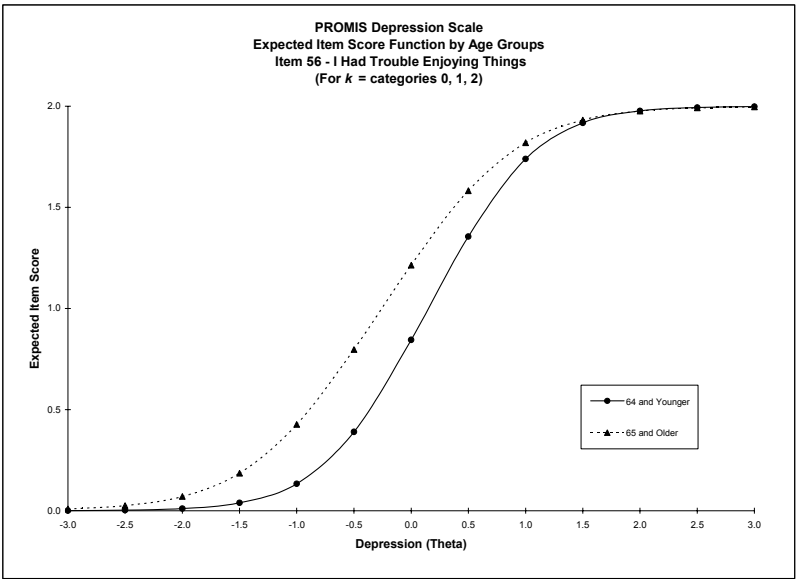


Figure 1:
Expected item score functions for gender, education and age groups for items with high magnitude of DIF: "I felt like crying (top); "I had trouble enjoying things" (right) and "I felt I had no energy" (right)


PROMIS Depression Scale
Expected Item Score Function by Age Groups
Item 56 - I Had Trouble Enjoying Things
(For k = categories 0, 1, 2)
Expected Item Score
Depression (Theta)
64 and Younger
65 and Older

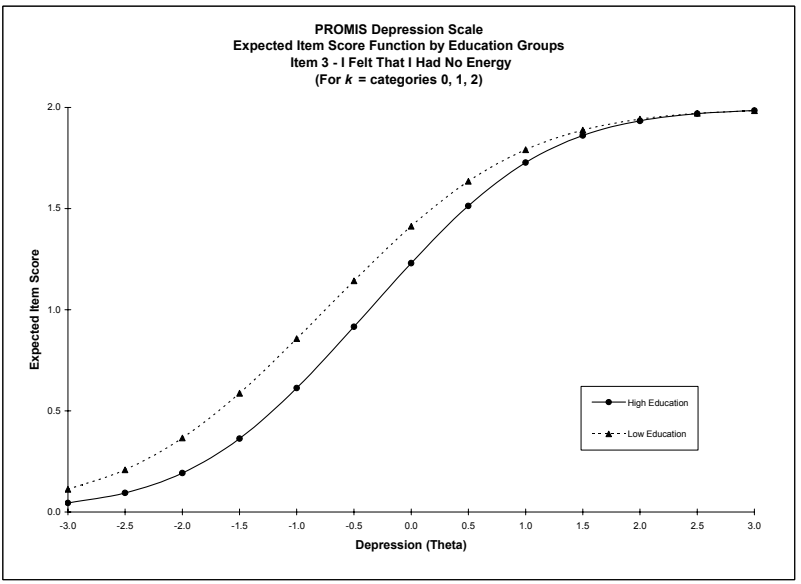

PROMIS Depression Scale
Expected Item Score Function by Education Groups
Item 3 - I Felt That I Had No Energy
(For k = categories 0, 1, 2)
Expected Item Score
Depression (Theta)
High Education
Low Education

Table 2:

Item parameters* and standard errors for the anchor items and studied items with DIF from the depression item bank: Comparison of education groups (Some college vs. College or Advanced degree)

Content	Group	*a*	*b1*	*b2*	*a*DIF	*b*DIF‡
I felt that I had no energy	Some College	1.66 (0.12)	-1.39 (0.11)	-0.28 (0.09)	2.2 (0.138)	*12.3 (0.002)*
	Coll Degree +	1.66 (0.12)	-1.07 (0.14)	-0.07 (0.14)		
I felt worthless	Some College / Coll Degree +	3.48 (0.30)	0.17 (0.05)	0.84 (0.08)	NS, Anchor Item	
I felt that I had nothing to look forward to	Some College / Coll Degree +	2.92 (0.26)	0.02 (0.05)	0.69 (0.07)	NS, Anchor Item	
I felt helpless	Some College / Coll Degree +	3.32 (0.28)	0.04 (0.05)	0.75 (0.07)	NS, Anchor Item	
I withdrew from other people	Some College / Coll Degree +	2.29 (0.19)	-0.30 (0.06)	0.46 (0.08)	NS, Anchor Item	
I felt that nothing could cheer me up	Some College / Coll Degree +	3.12 (0.29)	0.03 (0.05)	0.78 (0.07)	NS, Anchor Item	
I felt that other people did not understand me	Some College / Coll Degree +	2.27 (0.19)	-0.73 (0.06)	0.08 (0.07)	NS, Anchor Item	
I felt that I was not as good as other people	Some College / Coll Degree +	2.37 (0.21)	-0.15 (0.06)	0.63 (0.09)	NS, Anchor Item	
I felt like crying	Some College / Coll Degree +	1.90 (0.17)	-0.27 (0.07)	0.71 (0.10)	NS, Anchor Item	
I felt sad	Some College / Coll Degree +	2.79 (0.23)	-0.87 (0.05)	0.10 (0.06)	NS, Anchor Item	
I felt that I wanted to give up on everything	Some College / Coll Degree +	3.08 (0.29)	0.31 (0.06)	0.96 (0.08)	NS, Anchor Item	
I felt that I was to blame for things	Some College / Coll Degree +	2.45 (0.20)	-0.35 (0.06)	0.54 (0.07)	NS, Anchor Item	
I felt like a failure	Some College	3.44 (0.29)	-0.05 (0.06)	0.53 (0.07)	<0.01 (>0.999)	8.2 (0.017)
	Coll Degree +	3.44 (0.29)	-0.25 (0.08)	0.44 (0.12)		
I had trouble feeling close to people	Some College / Coll Degree +	2.57 (0.21)	-0.48 (0.06)	0.28 (0.07)	NS, Anchor Item	
I felt disappointed in myself	Some College / Coll Degree +	2.72 (0.21)	-0.71 (0.05)	0.12 (0.07)	NS, no DIF	

I felt that I was not needed	Some College	2.62 (0.22)	-0.09 (0.06)	0.68 (0.08)	NS, Anchor Item	
	Coll Degree +					
I felt lonely	Some College	2.33 (0.19)	-0.44 (0.06)	0.33 (0.07)	NS, Anchor Item	
	Coll Degree +					
I felt depressed	Some College	3.37 (0.28)	-0.47 (0.05)	0.35 (0.06)	NS, Anchor Item	
	Coll Degree +					
I had trouble making decisions	Some College	2.16 (0.18)	-0.50 (0.07)	0.56 (0.08)	NS, Anchor Item	
	Coll Degree +					
I felt discouraged about the future	Some College	2.48 (0.20)	-0.61 (0.06)	0.10 (0.07)	NS, Anchor Item	
	Coll Degree +					
I found that things in my life were overwhelming	Some College	2.61 (0.20)	-0.44 (0.06)	0.37 (0.07)	NS, Anchor Item	
	Coll Degree +					
I felt unhappy	Some College	3.55 (0.30)	-0.77 (0.05)	0.12 (0.05)	NS, Anchor Item	
	Coll Degree +					
I felt I had no reason for living	Some College	2.47 (0.28)	0.79 (0.08)	1.47 (0.12)	NS, Anchor Item	
	Coll Degree +					
I felt hopeless	Some College	3.68 (0.42)	0.14 (0.06)	0.86 (0.08)	4.4 (0.036)	6.3 (0.043)
	Coll Degree +	4.85 (1.02)	-0.04 (0.08)	0.71 (0.10)		
I felt ignored by people	Some College	2.21 (0.19)	-0.40 (0.06)	0.50 (0.08)	NS, Anchor Item	
	Coll Degree +					
I felt upset for no reason	Some College	2.35 (0.21)	-0.06 (0.06)	0.77 (0.08)	NS, Anchor Item	
	Coll Degree +					
I felt that nothing was interesting	Some College	2.32 (0.19)	-0.28 (0.06)	0.71 (0.09)	NS, Anchor Item	
	Coll Degree +					
I felt pessimistic	Some College	2.18 (0.21)	-0.70 (0.08)	0.20 (0.08)	9.7 (0.002)	0.4 (0.819)
	Coll Degree +	3.15 (0.47)	-0.64 (0.08)	0.16 (0.11)		
I felt that my life was empty	Some College	2.84 (0.25)	-0.01 (0.06)	0.52 (0.07)	NS, Anchor Item	
	Coll Degree +					
I felt guilty	Some College	2.11 (0.18)	-0.25 (0.06)	0.66 (0.09)	NS, Anchor Item	
	Coll Degree +					
I felt emotionally exhausted	Some College	2.73 (0.22)	-0.56 (0.05)	0.23 (0.06)	NS, Anchor Item	
	Coll Degree +					
I had trouble enjoying things that I used to enjoy	Some College	2.32 (0.20)	-0.42 (0.06)	0.40 (0.07)	NS, no DIF	
	Coll Degree +					

* Parameter estimates are from the final MULTILOG run

‡ DIF significant after Bonferroni adjustment is bolded; NCDIF items are italicized.

Table 3:

Item parameters* and standard errors for the anchor items and studied items with DIF from the depression item bank: Comparison of age groups (age 65 and over vs. age 64 and under)

Content	Group	*a*	*b1*	*b2*	*a*DIF	*b*DIF‡
I felt that I had no energy	65+	1.64 (0.12)	-1.55 (0.15)	-0.20 (0.15)	<0.1 (0.999)	8.2 (0.017)
	≤64	1.64 (0.12)	-1.14 (0.10)	-0.12 (0.09)		
I felt worthless	65+	3.48 (0.28)	0.00 (0.10)	0.77 (0.16)	0.2 (0.655)	8.9 (0.012)
	≤64	3.48 (0.28)	0.27 (0.06)	0.93 (0.07)		
I felt that I had nothing to look forward to	65+	3.03 (0.25)	-0.23 (0.10)	0.46 (0.12)	3.6 (0.058)	***17.4 (<0.001)***
	≤64	3.03 (0.25)	0.15 (0.06)	0.80 (0.07)		
I felt helpless	65+	4.59 (0.89)	-0.02 (0.09)	0.61 (0.13)	5.9 (0.015)	0.7 (0.705)
	≤64	3.02 (0.29)	0.10 (0.06)	0.85 (0.08)		
I withdrew from other people	65+	2.22 (0.18)	-0.27 (0.06)	0.51 (0.07)	NS, Anchor Item	
	≤64					
I felt that nothing could cheer me up	65+	3.14 (0.27)	-0.10 (0.11)	0.91 (0.16)	<0.01 (0.999)	7.6 (0.022)
	≤64	3.14 (0.27)	0.13 (0.06)	0.82 (0.07)		
I felt that other people did not understand me	65+	2.21 (0.19)	-0.72 (0.06)	0.12 (0.07)	NS, Anchor Item	
	≤64					
I felt that I was not as good as other people	65+	2.32 (0.21)	-0.12 (0.06)	0.69 (0.08)	NS, Anchor Item	
	≤64					
I felt like crying	65+	1.80 (0.14)	-0.08 (0.14)	1.20 (0.23)	1.6 (0.206)	*8.6 (0.014)*
	≤64	1.80 (0.14)	-0.29 (0.08)	0.70 (0.09)		
I felt sad	65+	2.68 (0.22)	-0.86 (0.05)	0.14 (0.07)	NS, no DIF	
	≤64					
I felt that I wanted to give up on everything	65+	3.10 (0.26)	0.10 (0.12)	0.88 (0.18)	2.0 (0.157)	9.5 (0.009)
	≤64	3.10 (0.26)	0.42 (0.06)	1.04 (0.08)		
I felt that I was to blame for things	65+	2.34 (0.20)	-0.32 (0.06)	0.60 (0.07)	NS, no DIF	
	≤64					
I felt like a failure	65+	3.26 (0.29)	-0.08 (0.05)	0.56 (0.06)	NS, Anchor Item	
	≤64					
I had trouble feeling close to people	65+	2.54 (0.19)	-0.58 (0.11)	0.12 (0.13)	<0.01 (0.999)	7.1 (0.029)
	≤64	2.54 (0.19)	-0.41 (0.07)	0.38 (0.06)		
I felt disappointed in myself	65+	2.64 (0.21)	-0.69 (0.05)	0.16 (0.07)	NS, Anchor Item	
	≤64					

I felt that I was not needed	65+	2.66 (0.20)	-0.28 (0.11)	0.54 (0.16)	1.0 (0.317)	8.7 (0.013)
	≤64	2.66 (0.20)	0.01 (0.06)	0.77 (0.08)		
I felt lonely	65+ / ≤64	2.26 (0.18)	-0.42 (0.06)	0.38 (0.07)	NS, Anchor Item	
I felt depressed	65+ / ≤64	3.25 (0.27)	-0.44 (0.05)	0.40 (0.06)	NS, Anchor Item	
I had trouble making decisions	65+ / ≤64	2.09 (0.18)	-0.48 (0.07)	0.61 (0.08)	NS, no DIF	
I felt discouraged about the future	65+	2.47 (0.18)	-0.62 (0.10)	-0.11 (0.11)	2.3 (0.129)	9.6 (0.008)
	≤64	2.47 (0.18)	-0.59 (0.07)	-0.20 (0.07)		
I found that things in my life were overwhelming	65+	2.50 (0.18)	-0.35 (0.13)	0.70 (0.15)	2.3 (0.129)	*7.9 (0.019)*
	≤64	2.50 (0.18)	-0.44 (0.07)	0.35 (0.07)		
I felt unhappy	65+ / ≤64	3.45 (0.29)	-0.76 (0.05)	0.17 (0.05)	NS, Anchor Item	
I felt I had no reason for living	65+ / ≤64	2.44 (0.27)	0.85 (0.08)	1.55 (0.12)	NS, Anchor Item	
I felt hopeless	65+ / ≤64	3.82 (0.38)	0.13 (0.05)	0.88 (0.06)	NS, Anchor Item	
I felt ignored by people	65+ / ≤64	2.12 (0.18)	-0.38 (0.06)	0.56 (0.08)	NS, no DIF	
I felt upset for no reason	65+ / ≤64	2.30 (0.20)	-0.02 (0.06)	0.83 (0.08)	NS, Anchor Item	
I felt that nothing was interesting	65+ / ≤64	2.26 (0.18)	-0.25 (0.06)	0.76 (0.08)	NS, Anchor Item	
I felt pessimistic	65+ / ≤64	2.33 (0.18)	-0.66 (0.06)	0.24 (0.07)	NS, Anchor Item	
I felt that my life was empty	65+	2.85 (0.22)	-0.13 (0.11)	0.61 (0.16)	0.3 (0.584)	6.4 (0.041)
	≤64	2.85 (0.22)	0.07 (0.06)	0.56 (0.07)		
I felt guilty	65+ / ≤64	2.04 (0.17)	-0.22 (0.07)	0.72 (0.08)	NS, Anchor Item	
I felt emotionally exhausted	65+ / ≤64	2.64 (0.21)	-0.55 (0.05)	0.27 (0.07)	NS, Anchor Item	
I had trouble enjoying things that I used to enjoy	65+	2.48 (0.18)	-0.80 (0.10)	0.06 (0.12)	2.7 (0.100)	***36.3 (<0.001)***
	64	2.48 (0.18)	-0.24 (0.07)	0.53 (0.07)		

* Parameter estimates are from the final MULTILOG run

‡ DIF significant after Bonferroni adjustment is bolded; NCDIF items are italicized.

Table 4:

Summary of DIF analyses of the depression item bank: Education, gender and age groups

Item	Item Name	Item Wording	Anchor Item			Type of DIF, if Present			DIF After Bonferroni/B-H Adjustment*			Magnitude (Expected Item Score Difference: NCDIF)		
			Sex	Educ	Age	Sex	Educ	Age	Sex	Educ	Age	Sex	Educ	Age
3	EDDEP03	I felt that I had no energy	✓				U	U					0.039	
4	EDDEP04	I felt worthless	✓	✓				U						
5	EDDEP05	I felt that I had nothing to look forward to	✓	✓				U			U			0.031
6	EDDEP06	I felt helpless		✓		NU		NU						
7	EDDEP07	I withdrew from other people		✓	✓									
9	EDDEP09	I felt that nothing could cheer me up		✓		U		U						
13	EDDEP13	I felt that other people did not understand me		✓	✓	U								
14	EDDEP14	I felt that I was not as good as other people	✓	✓	✓									
16	EDDEP16	I felt like crying		✓		U		U	U			0.074		0.065
17	EDDEP17	I felt sad		✓		NU								
19	EDDEP19	I felt that I wanted to give up on everything	✓	✓				U						
21	EDDEP21	I felt that I was to blame for things	✓	✓										
22	EDDEP22	I felt like a failure			✓		U							
23	EDDEP23	I had trouble feeling close to people		✓		U		U						
26	EDDEP26	I felt disappointed in myself	✓		✓									
27	EDDEP27	I felt that I was not needed	✓	✓				U						
28	EDDEP28	I felt lonely	✓	✓	✓									
29	EDDEP29	I felt depressed		✓	✓	U								
30	EDDEP30	I had trouble making decisions	✓	✓										
31	EDDEP31	I felt discouraged about the future	✓	✓				U						
35	EDDEP35	I found that things in my life were overwhelming	✓	✓				U						0.026
36	EDDEP36	I felt unhappy		✓	✓	U								
39	EDDEP39	I felt I had no reason for living	✓	✓	✓									
41	EDDEP41	I felt hopeless	✓		✓		NU							
42	EDDEP42	I felt ignored by people	✓	✓										
44	EDDEP44	I felt upset for no reason	✓	✓	✓									
45	EDDEP45	I felt that nothing was interesting		✓	✓	U								
46	EDDEP46	I felt pessimistic	✓		✓		NU							
48	EDDEP48	I felt that my life was empty		✓		U		U						
50	EDDEP50	I felt guilty	✓	✓	✓									
54	EDDEP54	I felt emotionally exhausted	✓	✓	✓									
56	EDDEP56	I had trouble enjoying things that I used to enjoy				U		U			U			0.080

* DIF was significant after Bonferroni/Benjamini-Hochberg adjustments. Both adjustments resulted in the same selection of items.

Age: Shown in Table 3 are the analyses of age. The original analyses of age with four response categories produced sparse data and very high *a* parameter estimates, resulting in false non-uniform DIF detection, and the identification of only three anchor items. In order to reduce sparse data, the top three categories were collapsed, yielding three response categories. The most severe indicator was "no reason for living"; among the least severe was "no energy". Comparison of the distributions indicated less depression for the older than for the younger cohort (estimated μ = -0.84 for older respondents vs. -0.23 for the younger group). The results of these analyses produced 15 anchor items. Before adjustment for multiple comparisons, 13 items showed DIF, one with non-uniform DIF; after Bonferroni/B-H correction, two items showed uniform DIF: "I felt I had nothing to look forward to" (NCDIF=0.031), and "I had trouble enjoying the things I used to enjoy" (NCDIF=0.080). Both of these items were more severe indicators for the younger cohort. Two other items had NCDIF values above the cutoff: "I felt like crying" (NCDIF=0.065) and the item, "I found that things in my life were overwhelming" (NCDIF=0.026). Both items evidenced uniform DIF prior to adjustment for multiple comparisons, and were more severe indicators for the older cohort.

Sensitivity analyses

Comparisons using different categorizations: It is possible that lack of model fit and sparse data may have resulted in false DIF detection in the earlier analyses with four response categories; thus as stated above, the primary analyses were performed using three response categories. Additionally, because some depression scales use binary versions of many of the items examined, and in order to obtain more robust results, IRTLR analysis was also performed using a binary version of the items: not present vs. symptomatic (combining the categories above none and rarely into the value, 1). This reduced the number of parameters estimated, and also reduced the sparse data.

Consistent with the primary analyses, only one item showed significant gender DIF in the sensitivity analyses, the crying item. The NCDIF index for the crying item ranged across sensitivity analyses from 0.043 to 0.091. This represents a relatively small effect size or absolute difference ranging from 0.21 to 0.30 on a two to four point scale, depending on the analyses.

For education, after adjustment for multiple comparisons, no items evidenced DIF; this is the same result found in the primary analyses. The NCDIF index was low for all items, consistently showing low magnitude of DIF. However, in the primary analyses with three categories, one item, "I felt that I had no energy", although not significant after adjustment for multiple comparisons, did evidence NCDIF over the threshold for education comparisons.

For age, the results for the analysis examining the focal group, aged 60 years and over instead of 65 and over showed that the item, "nothing to look forward to" showed uniform DIF, and the item, "trouble enjoying things I used to" showed non-uniform DIF after Bonferroni/B-H correction. The latter finding was consistent with the findings for the polytomous version, in which the item was found to show uniform DIF of relatively high magnitude. The NCDIF ranged from 0.046 to 0.089 across sensitivity analyses for this item.

The various analyses of age produced similar parameter estimates. The item, "I had trouble enjoying the things I used to do" evidenced uniform DIF; across most of the range of the depression scores, the probability of endorsement was higher for the older than the younger cohort. Regardless of analyses, this item showed significant, relatively higher magnitude DIF.

A concern is that sparse data may have produced spurious (large and inconsistent) *a* parameters. The correlations among *a* parameters estimates for the final models, and those used in the sensitivity analyses ranged from 0.905 to 0.998 for age, from 0.895 to 0.995 for education, and from 0.934 to 0.992 for gender, providing some evidence for the consistency of estimates for the final models.

Comparison with other methods: The consistent finding across all methods is that there is gender DIF associated with the item, "I felt like crying". Higher conditional endorsement was observed for women; the item was a more severe indicator of depression for men than for women. This finding was 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 not observed to be significant with IRTLR after adjustment, but the hypothesis was confirmed by two analyses (IRTOLR and DFIT). This item was found by all methods to have DIF, either for gender, age and/or education.

Another item hypothesized by content experts to possibly show gender and age DIF that was confirmed by three methods (IRTOLR, Poly-SIBTEST and DFIT) to show age, gender or education DIF was the item, "I felt that I had no energy". Conditional on depression, those 65 and over were less likely to report no energy (IRTORL, IRTLR before Bonferroni adjustment); and those with lower education were more likely to endorse the item (POLYSIB); the βuni from the SIBTEST analyses was -0.26. This result was consistent with the primary analyses where the NCDIF index was above the threshold (0.039). No confirmatory literature was available.

Impact of DIF: Figure 2, showing the test response functions, is a summary of the findings across the methods regarding the impact of DIF associated with the PROMIS depression items. Impact, examined at the aggregate level, was found to be minimal in the IRTOLR and

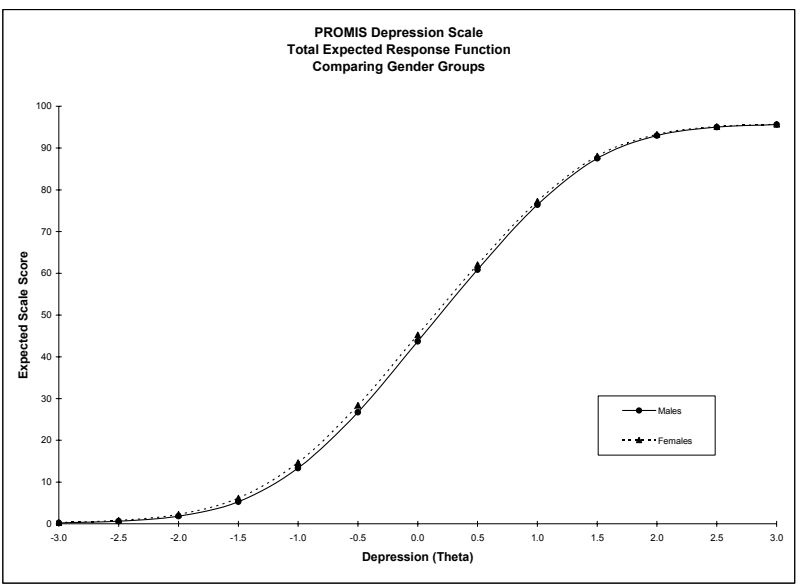
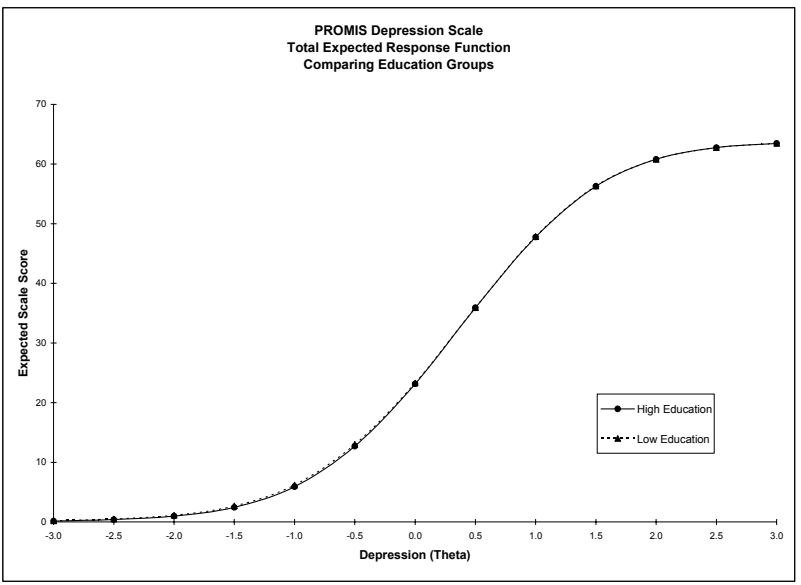
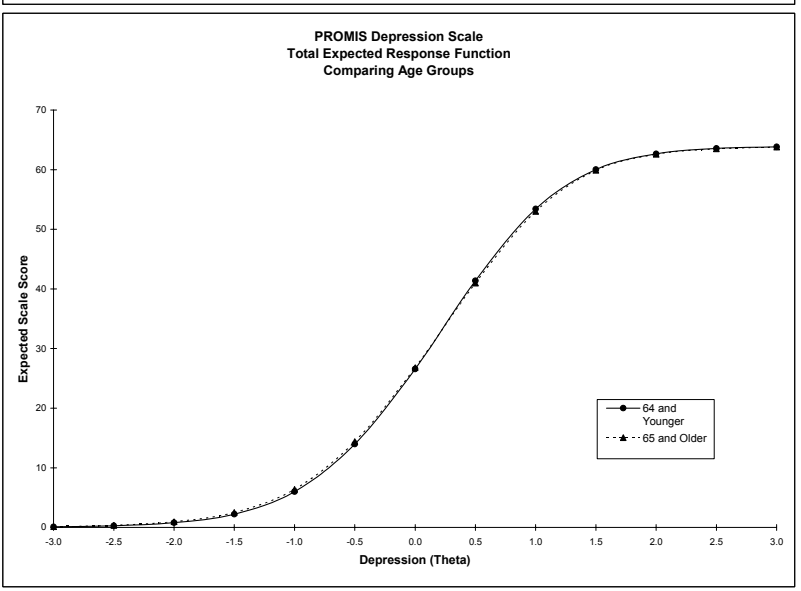


Figure 2: Test response functions for gender (top), education and age (right)

MIMIC analyses when mean scale or latent state scores were examined with and without adjustment for DIF. As shown in Figure 2, this result is confirmed using the expected scale scores that are based on the IRTLR/MULTILOG result. The Differential Test Function (DTF) (Raju et al. 1995) values (a density-weighted summary of differences between groups in test response functions) were small, and not significant: 0.071 for gender, 0.047 for education and 0.129 for age.





Analyses at the individual level using IRTOLR showed DIF impact for some people; this result was confirmed using IRTLR. An examination of the differences in thetas with and without adjustment for DIF showed that 22.4% of subjects changed by at least 0.5 theta (about one half standard deviation), of which 5.8% changed by the equivalent of one standard deviation in the analyses of gender; for education the figures are 53.5% and 25.1%, and for age 3.8% changed by at least 0.5 standard deviations. The impact is in the direction of false positives for depression. For example, using a cutoff of theta ≥ 1, comparable to about a standard deviation above the mean, 9.5% would be classified as depressed prior to DIF adjustment, but not after adjustment in the analyses of gender; the comparable figures for education and age are 13.6% and 3.4%, respectively. Thus, the impact at the individual level was large for at least 100 people. Examination of the characteristics of those with very large changes in theta (≥ 1.25) shows that for the gender analyses all were females with lower education (86.1%); in the analyses of education, all except two were of lower education. Thus a common component across the analyses is lower education. The discrepancy between the aggregated and the individual impact result is in part because the expected scale scores reflect DIF cancellation, in that items with DIF in one direction may cancel items with DIF in another direction, producing low overall impact. However, specific individuals may still be affected.

Summary: Despite caveats discussed below, based on review of (a) hypotheses (b) findings from the literature (c) the collective results from the various analyses, the items that were recommended for exclusion from calibration or treatment using some other technique that accounts for DIF include the following two items, "I felt like crying" and "I had trouble enjoying things that I used to enjoy". The item, "I felt I had no energy" was also flagged as an item that showed DIF across several methods and comparisons, and was recommended for further review.

Discussion

The findings were of relatively few items with significant or salient DIF in the depression item bank. Only two items were recommended for exclusion from calibrations, and one was recommended for further review. These items were hypothesized to show DIF and found in the literature to evidence DIF. Variants of the crying item have been identified in most studies as evidencing DIF (Cole et al. 2000; Gelin & Zumbo, 2003; Grayson et al. 2000; Pickard et al. 2006; Reeve, 2000; Teresi & Golden, 1994; Yang & Jones, 2007). Steinberg and Thissen (2006) in discussing DIF in the context of personality inventories and other depression surveys showed that uniform gender DIF of relatively large effect size has been observed for the crying item. While it is possible to use other methods, such as separate group calibrations for items with DIF, expert review of content area is necessary to determine whether an item is sufficiently clinically salient to remain in the item pool or bank. Procedures for accounting for DIF in the context of CAT require further development.

Limitations: Limitations of the study include the inability (due to small sample sizes) to examine DIF by ethnicity or language. Smaller sample sizes may also have affected the power to detect DIF for the analyses of education and age. However, sensitivity analyses, conducted using other models, e.g., MIMIC, did not yield substantively different results;

MIMIC has been found in simulation studies to be more powerful than IRTLR for uniform DIF detection under conditions of smaller sample sizes (Woods, 2009b).

A caveat is that, to the extent that the findings are not robust given the various features of the data described above, these impact results could be incorrect. It is noted that many of the analyses of impact from the literature (see below) concluded that the impact of DIF on depression scale scores was not trivial, and the impact on some individuals may be large. The findings from these analyses were similar. While the aggregated impact was low, with evidence of DIF cancellation, relatively large individual impact (defined as a large change in the depression estimate after DIF adjustment) was observed for about 14% of the sample for at least one analysis. This underscores the need for removal or separate calibration of items with a high magnitude of DIF.

These findings should be interpreted in the context of factors that may affect DIF (see Teresi, 2006). Several features of the data have been found to be problematic in terms of DIF detection for one or more methods. These include the presence of sparse and skewed data, usually from small subgroup sample sizes. For example, simulation studies have shown that skewed data (with floor effects) resulted in reduced power for DIF detection using ordinal logistic regression (Scott, Fayers, Aaronson, Bottomley, de Graeff et al., 2009). An attempt was made to remedy this by collapsing categories and performing sensitivity analyses using binary items. Gelin and Zumbo (2003), using an ordinal logistic regression approach to examine DIF in the CES-D items found that the endorsement proportions and the way in which items were scored: ordinally, binary or in terms of a persistency threshold (frequency of at least 3-7 days) affected DIF results, with higher magnitude of DIF for the binary and ordinal methods. In the analyses presented in this paper, the results were similar, regardless of categorization method.

Parametric models are more powerful for DIF detection with small subgroup sample sizes such as those observed in this study. IRTLR has been found in simulation studies, conducted using several polytomous response IRT models, to result in false DIF detection in the presence of group differences in the state/trait distributions when large sample sizes are studied (Bolt, 2002). While group differences in distributions were observed, particularly for age groups, simulation studies have not observed type 1 error inflation in studies of smaller subgroups such as those present in PROMIS.

IRTLR has also been found to be affected by lack of purification, magnitude of DIF present, and degree of DIF cancellation in simulations of binary data, resulting in both over and under-identification of items with DIF (Finch, 2005; Navas-Ara & Gómez-Benito, 2002; Wang & Yeh, 2003; Finch & French, 2007). An assumption in the use of IRTLR log-likelihood tests for DIF-detection is that the conditioning variable is DIF-free. The authors of one recent simulation (Finch & French) recommend first using SIBTEST or LR to purify the matching variable. In the case of ordinal data, OLR would be most efficient, followed by IRTLR. Purification was performed for the data set reported in this paper; however, predetermined anchor items were not available. Although Poly-SIBTEST was one of the methods used in sensitivity analyses, it was not used as a first-stage screen to identify potential anchor items as suggested by Finch and French, but rather to examine the convergence of findings across methods. It is noted that the recommendations of these authors were based on the results of simulations of binary data, and may not hold for polytomous data, such as those reported in this paper. The selection of anchor items has been discussed in earlier literature (e.g., Cohen, Kim and Wollack, 1996; Thissen, Steinberg and Wainer, 1993); more

recent research has focused on anchor items and purification approaches in the context of IRTLR (Orlando-Edelen, Thissen, Teresi, Kleinman and Ocepek-Welikson, 2006; Woods 2009a), MIMIC (Shih & Wang, 2009; Wang, Shih, Yang, in press) and the more general multigroup confirmatory factor analysis approach to examining invariance (French & Finch, 2008). These studies generally support the use of anchor items or the selection of invariant referent items. One study showed that the inclusion of at least 4 anchor items was preferable, in the context of power for IRTLR DIF detection (Wang and Yeh, 2003); a similar result was observed for MIMIC (Shih and Wang, 2009). On the positive side, a sufficient number (at least 10 in these analyses) of anchor items were identified for each IRTLR analysis, mitigating the impact of DIF on initial estimates.

False DIF detection (type 1 error inflation) can also result from model mis-specification for polytomous data (Bolt, 2002). A recent examination of the results of model misspecification of the likelihood ratio test used in IRTLR and other nested models such as logistic regression examined models that included 5, 10 and 30 items; data were generated using the one, two and three parameter logistic IRT model, with a normally distributed latent variable. Thus, generalization is to binary items. Lack of model fit was found to affect the G^2 used in nested models to test for DIF in IRTLR. If the least restricted model does not fit the data, this misspecification will result in incorrect statistical inferences (Maydeu-Olivares & Cai, 2006) and inflated type 1 error (Yuan & Bentler, 2004). Model fit was examined in the analyses reported here, and some misfit was observed for the high education group, and for males. Further collapsing resolved the misfit problems for education groups, and reduced, but did not eliminate misfit for males.

Finally, a good practice recommended by several authors (e.g., Crane et al. 2004; Hambleton, 2006; Millsap, 2006; Teresi & Fleishman, 2007) is to apply magnitude measures to identify salient DIF. For example, one recent simulation study of logistic regression (French & Maller, 2007) found that use of effect sizes under several conditions reduced false DIF detection, albeit at the expense of reduced power. An issue is what flagging rules to use in DIF detection (Hidalgo & López-Pina, 2004). Simulation studies of flagging rules or cutoff thresholds indicative of magnitude have resulted in differing suggested values, leading to a recent recommendation (Meade et al. 2007) to derive empirically cutoff values appropriate for the data set used. While such magnitude measures were applied in these analyses, cutoff values were not data-specific. PROMIS investigators have examined different criteria for flagging DIF (Crane et al. 2007), and are currently developing the capability to derive empirical thresholds using Monte Carlo simulations (Choi, Gibbons & Crane, 2009).

Despite these limitations, several strengths of the study include the extensive qualitative analyses performed that led to item revisions, the generation of DIF hypotheses and the use of purification of the conditioning depression variable. Additionally, model assumptions were tested, and sparse data controlled to the extent possible by collapsing categories. Finally, extensive sensitivity analyses were performed, and multiple methods were used in combination with DIF magnitude measures in order to investigate DIF and converge on valid, consistent findings, as suggested by Hambleton (2006).

Conclusions

Little DIF was found in the depression item bank for the groups studied. The extensive qualitative analyses that preceded this effort may have mitigated the extent of DIF in the item bank. On the one hand, false DIF detection (Type 1 error) was controlled to the extent possible by ensuring model fit and purification. On the other hand, power for DIF detection might have been compromised by several factors, including sparse data and small sample sizes. Nonetheless, practical and not just statistical significance should be considered. In this case the overall magnitude and impact of DIF was small for the groups studied; although impact for some individuals was relatively large, supporting the removal or separate calibration of a few items. This is a particularly important consideration in the context of item banks, because individuals may receive only a subset of items, with the potential for magnification of the impact of DIF for some people. Future analyses of the item bank should be performed examining ethnicity and language.

A question arises as to the practical implication of DIF for selection and prediction. A discussion of the relationship of measurement invariance to fair selection and prediction invariance is beyond the scope of this article, but is discussed in several seminal works, e.g., Meredith (1993), Millsap (1997), and more recently in Millsap (2007), Meredith and Teresi (2006), and Borsboom, Romeijn and Wicherts (2008). As shown by Millsap (2007), and illustrated in part by the results presented here, measures may show no aggregate prediction bias, but can produce systematic selection errors at the individual level due to measurement bias.

Item banks are being used increasingly to assess health and psychological domains; in that context it is critical that decisions and resource allocation based on these assessments result from a valid measurement process. The methods described in this paper are key steps in the development and evaluation of item banks, and of short-form measures that may be constructed from such banks. These methods may also be applied to examine the performance of existing measures. Individual differences, reflected in cultural, gender, educational or ethnic diversity, must be considered in the development and evaluation of measures. Analysis of DIF in the PROMIS depression item bank is an important step toward the goal of increasing the likelihood of measurement equivalence across diverse groups.

Glossary of terms

Anchor items

Anchor items are those items found (through an iterative process or prior analyses) to be free of DIF. These items serve to form a conditioning variable used to link groups in the final DIF analyses. (See also the discussion of purification, below.)

Differential Item Functioning (DIF)

In the context of item response theory, DIF is observed when the probability of item response differs across comparison groups such as gender, country or language, after conditioning on (controlling for) level of the state or trait measured, such as depression.

Uniform DIF: Uniform DIF occurs if the probability of response is consistently higher (or lower) for one of the comparison groups across all levels of the state or trait.

Non-Uniform DIF: Non-uniform DIF is observed when the probability of response is in a different direction for the groups compared at different levels of the state or trait. For example, the response probability might be higher for females than for males at higher levels of the measure of the depression state, and lower for females than for males at lower levels of depression. For some DIF detection methods, e.g., logistic regression, non-uniform DIF is defined as a significant group by depression interaction.

Magnitude: The magnitude of DIF relates to the degree of DIF present in an item. In the context of IRT, a measure of magnitude is non-compensatory DIF (NCDIF) described for binary items as the unsigned probability difference (Camilli & Shepard, 1994), and later expanded to polytomous items by Raju and colleagues (1995). This index reflects the group difference in expected item scores (see Expected Item Scores). In essence this method provides an estimate of what expected score would obtain for an individual if s/he was scored based on the parameters and depression estimates for group X, and then based on the depression and parameter estimates for group Y. NCDIF is the average squared difference in expected item scores for a given individual as a member of the focal group, and as a member of the reference group. Theoretical work in this area was provided by Chang and Mazzeo (1994). (For computational details, see Collins, Raju & Edwards, 2000; Morales, Flowers, Gutiérrez, Kleinman & Teresi, 2006; Teresi et al. 2007).

Specific cutoff values are used to indicate salient DIF. While chi-square tests of significance are available, these were found to be too stringent, over identifying DIF. Cutoff values established based on simulations (Fleer, 1993; Flowers, Oshima & Raju, 1999) provide an estimate of the magnitude of item-level DIF. The cutoff values are controversial; for example, for polytomous items with five response options the recommended cutoff in the manual is 0.096 (Raju, 1999). However, simulation studies have suggested the use of different cutoff values for five response categories: 0.032 for smaller sample sizes of 300 per group (Bolt, 2002) or 0.016 (Flowers, 1995). The most recent recommendations (Meade et al. 2007) suggest using 0.0115 for a liberal test of DIF for five response category items, and 0.009 for a conservative test for sample sizes $\leq$ 500/group. Recently, Oshima, Raju and Nanda (2006) have recommended other cutoff values for binary items, and empirically derived cutoffs based on the data set have been incorporated into DFIT8 (Oshima, Kushubar, Scott, Raju, 2009).

Impact: (See also Expected Scale Score and Differential Test Functioning) Impact refers to the influence of DIF on the scale score. There are various approaches to examining impact, depending on the DIF detection method. In the context of IRTLR, differences in "test" response functions can be constructed by summing the expected item scores to obtain an expected scale score. Plots (for each group) of the expected scale score against the measure of the state or trait (e.g., depression) provides a graphic depiction of the difference in the areas between the curves, and shows the relative impact of DIF. The Differential Test Functioning (DTF) index (Raju et al. 1995) is a summary measure of these differences that incorporate such a weight, and reflects the aggregated net impact. The DTF is the sum of the item-level compensatory DIF indices, and as such reflects the results of DIF cancellation. (See also Stark et al. 2004.)

Individual impact can be assessed through an examination of changes in depression estimates (thetas) with and without adjustment for DIF. The unadjusted thetas are produced from a model with all item parameters set equal for the two groups. The adjusted thetas are

produced from a model with parameters that showed DIF based on the IRTLRDIF results estimated separately (freed) for the groups.

DIF cancellation

DIF is said to cancel if the net impact of DIF is trivial. For example, if the differences between expected scale scores (defined below) for the groups compared are negligible, resulting in small areas between the curves relating expected scale scores to the measure of the latent state, depression, then DIF is said to cancel. Because the expected scale score is on the raw metric of the scale score, at each level of depression disorder, it is possible to locate the average scale score associated with that degree of symptomatology.

Expected Item Scores (EIS) item level magnitude (effect size) measures

An EIS is the sum of the weighted (by the response category value) probabilities of scoring in each of the possible item categories. Used by Wainer, Sireci and Thissen (1991), this effect size measure is gaining in popularity. (See also Collins et al. 2000; Orlando-Edelen et al. 2006; Steinberg & Thissen, 2006; Teresi et al. 2007.)

Expected Scale Score (ESS) (test response function) scale level impact measures (see Impact, above)

The expected scale score is the sum of the expected item scores. The test response function (Lord & Novick, 1968) relates average expected scale scores to theta (the estimate of depression). Note that these scores are typically not weighted by the response frequency; however, such a weight can be applied so that the results reflect the relative frequencies in the sample.

Item Response Theory (IRT)

Several forms of item response theory models are available for binary, categorical and ordinal data. Because the data presented here were ordinal, with up to five ordered response categories, a graded response model (Samejima, 1969) was applied to the data using MULTILOG (Thissen, 2001). In this model (which reduces to the 2 parameter logistic item response model with binary data), we assume ordered responses, x=k and k=1,2,...m. The discrimination parameter or slope can be defined as a_i, and difficulty parameters for response k as b_{ik}.

$$P(x=k) = P^*(k) - P^*(k+1) = 1 / [1 + \exp[-Da_i(\theta - b_{ik-1})] - 1 / [1 + \exp[-Da_i(\theta - b_{ik})].$$

P*(k) is the ICC describing the probability that a response is in category k or higher, for each value of θ (see Thissen, 2001; Orlando-Edelen et al. 2006). The model assumes an

average discrimination across response categories. Note that the scaling parameter, D, is used in some IRT programs, but not in MULTILOG, the program used in these analyses.

Purification

Item sets that are used to construct preliminary estimates of the attribute assessed, e.g., depression, include items with DIF. Thus, estimation of a person's standing on the attribute may be incorrect, using this contaminated estimate. Purification is the process of iteratively testing items for DIF so that final estimation of the trait can be made after taking this item-level DIF into account. Because simulation studies have shown that most methods of DIF detection are adversely affected by lack of purification, the process is critical, particularly for IRTLR. Using this method, anchor items are selected that are free of DIF. For most models, these anchor items and the studied item form the conditioning set of items used in the DIF detection process. During this iterative process items may change in terms of DIF status, a result of the use of a less than optimal (contaminated) conditioning variable at various steps in the analyses. The final estimates of the attribute use all items, however, only after parameters have been appropriately set as freely or equally estimated, depending on whether the items showed DIF or not.

Acknowledgements

These analyses were conducted on behalf of the Statistical Coordinating Center to the Patient Reported Outcomes Measurement Information System (PROMIS), a United States National Institutes of Health roadmap project. PROMIS was funded by cooperative agreements to a Statistical Coordinating Center (Evanston Northwestern Healthcare, PI: David Cella, PhD, U01AR52177) and six Primary Research Sites (Duke University, PI: Kevin Weinfurt, PhD, U01AR52186; University of North Carolina, PI: Darren DeWalt, MD, MPH, U01AR52181; University of Pittsburgh, PI: Paul A. Pilkonis, PhD, U01AR52155; Stanford University, PI: James Fries, MD, U01AR52158; Stony Brook University, PI: Arthur Stone, PhD, U01AR52170; and University of Washington, PI: Dagmar Amtmann, PhD, U01AR52171). NIH Science Officers on this project are Deborah Ader, Ph.D., Susan Czajkowski, PhD, Lawrence Fine, MD, DrPH, Louis Quatrano, PhD, Bryce Reeve, PhD, William Riley, PhD, and Susana Serrate-Sztein, PhD. This manuscript was reviewed by the PROMIS Publications Subcommittee prior to external peer review. See the web site at www.nihpromis.org for additional information.

Funding for analyses was provided in part by the National Institute on Aging (NIA), Resource Center for Minority Aging Research at Columbia University, PI: Rafael Lantigua, Co-Director, Jeanne Teresi (AG15294), and by the NIA project, AG025308, Understanding Disparities in Mental Status Assessment, PI: Richard Jones. This paper was prepared for the International Conference on Survey Methods in Multinational, Multiregional and Multicultural Contexts, Berlin, Germany, June 25-28, 2008.

References

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.

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.

Beck, A. T., Ward, C. H., Mendelsohn, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Arch. General Psychiatry, 4*, 561-571.

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.

Bolt, D. M., (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141.

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.

Borsboom, D., Romeijn, J-W. & Wicherts, J.M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*, 75-98.

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., Ader, D., Fries, J. F., Bruce, 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.

Chan, K. S., Orlando, M., Ghosh-Dastidar, B., Duan, N., & 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.

Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response functions in polytomously scored item response models. *Psychometrika, 39*, 391-404.

Chang, H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *Journal of Educational Measurement, 59*, 333-353.

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2009). Development of freeware for an iterative hybrid ordinal logistic regression/IRT DIF: A Monte Carlo simulation approach for determining cutoff values. Paper presented at the National Council on Measurement in Education. April.

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.

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.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential item functioning in a satisfaction scale. *Journal of Applied Psychology, 85*, 451-461.

Copeland, J. R. M., Kelleher, M. J., Kellett, J. M., Gourlay, A. J., Gurland, B. J., Fleiss, J. L., & Sharpe, L. A. (1976). Semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule I: Development and reliability. *Psychological Medicine, 6*, 439-449.

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.

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., Hays, R., & Teresi, J. A. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Quality of Life Research, 16*, 69-84.

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.

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.

Dorans, N. J., & Kulick, E. (2006). Differential item functioning on the MMSE: An application of the Mantel Haenszel and Standardization procedures. *Medical Care, 44*(11 Suppl 3), S107-S114.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST and the IRT likelihood ratio test. *Applied Psychological Measurement, 29*, 278-295.

Finch, W. H., & French, B. F. (2007). Detection of crossing differential item functioning. A Comparison of Four Methods. *Educational and Psychological Measurement, 67*, 565-582.

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. (1995). A Monte Carlo assessment of DFIT with dichotomously-scored unidimensional tests. (Dissertation, Georgia State University, Atlanta, GA).

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.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling, 15,* 96-113.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373-393.

Gelin, M. N., & Zumbo, B. D. (2003). Differential item functioning results may change depending on how an item is scored: An illustration with the center for epidemiologic studies depression scale. *Educational and Psychological Measurement, 63*(1), 65-74.

Golden, R. R., Teresi, J. A., & Gurland, B. J. (1984). Development of indicator scales for the Comprehensive Assessment and Referral Evaluation (CARE) interview schedule. *Journal of Gerontology, 39*, 138-146.

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.

Gurland, B. J., Fleiss, J. L., Goldberg, K., Sharpe, L., Copeland, J. R. M., Kelleher, M. J., & Kellet, J. M. (1976). A semi-structured clinical interview for the assessment of diagnosis and mental state in the elderly: The Geriatric Mental State Schedule II: A factor analysis. *Psychological Medicine, 6*, 451-459.

Hambleton, R. K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11 Suppl 3), S182-S188.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications, Inc.

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915.

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.

Jöreskog, K., & Goldberger, A. (1975). Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Society, 10*, 631-639.

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.

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.

Liu, H-H., Cella, D., Gershon, R., Shen, J., Morales, L. S., Riley, W. & Hays, R. D. Representativeness of the PROMIS Internet Panel. Under Review, *Journal of Clinical Epidemiology.*

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

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley Publishing Co.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2 (dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41*, 55-64.

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.

Meredith W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58,* 525-543.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44* (11 Suppl 3), S69-S77.

Millsap, R. E. (2006). Comments on the methods for the investigation of measurement bias in the Mini-Mental State Examination. *Medical Care, 44*(11 Suppl 3), S171-S175.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72,* 461-473.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2,* 248-260.

Morales, L. S., Flowers, C., Gutiérrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the Mini-Mental Status Exam assessed using the differential item and test functioning (DFIT) framework. *Medical Care, 44*(11 Suppl 3), S143-151.

Mui, A. C., Burnette, D., & Chen, L. M. (2001). Cross-cultural assessment of geriatric depression: A Review of the CES-D and GDS. Measurement in older ethnically diverse population. *Journal of Mental Health and Aging, 7*(1), 137-164.

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

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment, 18*, 9-15.

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.

Osborne, R. H., Elsworth, G. R., Sprangers, M. A. G., Oort, F. J., & Hopper, J. L. (2004). The value of the Hospital Anxiety and Depression Scale (HADS) for comparing women with early onset breast cancer with population-based reference women. *Quality of Life Research, 13*, 191-206.

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.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance of the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17.

Pedersen, R. D., Pallay, A. G., & Rudolph, R. L. (2002). Can improvement in well-being and functioning be distinguished from depression improvement in antidepressant clinical trials? *Quality Life Research, 11*, 9-17.

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.

Prince, M., Reischies, F., Beekman, A. T., Fuhrer, R., Jonker, C., Kivela, S. L., Lawlor, B., Lobo, A., Magnusson, H., Fichter, M. M., Van Oyen, H., Roelands, M., Skoog, I., Turrina, C., & Copeland, J. R. (1999). Development of the EURO-D scale, a European Union initiative to compare symptoms of depression in 14 European centres. *British Journal of Psychiatry, 174*, 330-338.

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.

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 Polytomous DFIT framework. *Applied Psychological Measurement, 33,* 133-147.

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.

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., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise S. P., Jai, J.-S., & 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., Morizot, J., & Hays, R. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*(Suppl 1), 19-31.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph, 17*.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Petersen, M. A., Sprangers, M. A. G. on behalf of the EORTC Quality of Life Group and the Quality of Life Cross-Cultural Meta-Analysis Group. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of Clinical Epidemiology. 62,* 288-295.

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.

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.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402-415.

Stewart, A. L., Ware, J. E., Sherbourne, C. D., & Wells, K. B. (1992). Psychological distress/well-being and cognitive functioning measures. In A. L. Stewart & J. E. Ware (Eds.), *Measuring functioning and well-being. The Medical Outcomes Study approach* (102-142). Durham, NC: Duke University Press.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370

Teresi, J. A. (2006). Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. *Medical Care, 44*(11 Suppl 3), S152-170.

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research, 16*(Supp. 1), 33-42.

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., 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.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Cook, K. F., Crane, P. K, Gibbons, L. E., Morales L. S, Orlando-Edelen, M., & Cella, D. (2007). Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measure physical functioning ability and general distress. *Quality Life Research, 16*(Suppl 1), 43-68.

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. (1991). *MULTILOG™ User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory.* Chicago: Scientific Software, Inc.

Thissen, D. (2001). IRTLRDIF v2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning [Computer software]. Retrieved from http://www.unc.edu/~dthissen/dl.html.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group-mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

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.

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.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28*, 197-219.

Wang, W-C., Shih, C-L., Yang, C-C. (in press). The MIMIC method with scale purification procedure for detecting differential item functioning, *Educational and Psychological Measurement.*

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.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage Publications.

Woods, C. M. (2008). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement, 33*, 42-57.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44,* 1-27.

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.

Yuan, K., & Bentler, P. M. (2004). On chi-square differences and z tests in mean and covariance structure analysis when the base model is misspecified. *Educational and Psychological Measurement, 64*, 737-757.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., Leirer, V. O. (1982). Development and validation of a geriatric depression screening scale: A preliminary report. *Journal of Psychiatric Research, 17*(1), 37-49.

Zigmond, A. S., & Snaith, R. P. (1983). The Hospital Anxiety and Depression Scale. *Acta. Psychiat. Scand., 67*, 361-370.

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.