Psychological Test and Assessment Modeling, Volume 58, 2016 (2), 255-307

Measurement equivalence of the Patient Reported Outcomes Measurement Information System® (PROMIS®) Applied Cognition – General Concerns, short forms in ethnically diverse groups

Robert Fieo[1], Katja Ocepek-Welikson[2], Marjorie Kleinman[3], Joseph P. Eimicke[2,4], Paul K. Crane[5], David Cella[6] & Jeanne A. Teresi[2,4,7]

Abstract

Aims: The goals of these analyses were to examine the psychometric properties and measurement equivalence of a self-reported cognition measure, the Patient Reported Outcome Measurement Information System® (PROMIS®) Applied Cognition – General Concerns short form. These items are also found in the PROMIS Cognitive Function (version 2) item bank. This scale consists of eight items related to subjective cognitive concerns. Differential item functioning (DIF) analyses of gender, education, race, age, and (Spanish) language were performed using an ethnically diverse sample (n = 5,477) of individuals with cancer. This is the first analysis examining DIF in this item set across ethnic and racial groups.

Methods: DIF hypotheses were derived by asking content experts to indicate whether they posited DIF for each item and to specify the direction. The principal DIF analytic model was item response theory (IRT) using the graded response model for polytomous data, with accompanying Wald tests and measures of magnitude. Sensitivity analyses were conducted using ordinal logistic regression (OLR) with a latent conditioning variable. IRT-based reliability, precision and information indices were estimated.

---

[1] *Correspondence concerning this article should be addressed to:* Robert Fieo, Assistant Professor, University of Florida, College of Medicine, Department of Geriatric Research, 2004 Mowry Road, Gainesville, FL 32611, USA; fieo@ufl.edu

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

[3] New York State Psychiatric Institute, Division of Child and Adolescent Psychiatry

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

[5] University of Washington, Department of Medicine

[6] Northwestern University Feinberg School of Medicine, Department of Medical Social Sciences

[7] Columbia University Stroud Center at New York State Psychiatric Institute

Results: DIF was identified consistently only for the item, brain not working as well as usual. After correction for multiple comparisons, this item showed significant DIF for both the primary and sensitivity analyses. Black respondents and Hispanics in comparison to White non-Hispanic respondents evidenced a lower conditional probability of endorsing the item, brain not working as well as usual. The same pattern was observed for the education grouping variable: as compared to those with a graduate degree, conditioning on overall level of subjective cognitive concerns, those with less than high school education also had a lower probability of endorsing this item. DIF was observed for age for two items after correction for multiple comparisons for both the IRT and OLR-based models: "I have had to work really hard to pay attention or I would make a mistake" and "I have had trouble shifting back and forth between different activities that require thinking". For both items, conditional on cognitive complaints, older respondents had a higher likelihood than younger respondents of endorsing the item in the cognitive complaints direction. The magnitude and impact of DIF was minimal.

The scale showed high precision along much of the subjective cognitive concerns continuum; the overall IRT-based reliability estimate for the total sample was 0.88 and the estimates for subgroups ranged from 0.87 to 0.92.

Conclusion: Little DIF of high magnitude or impact was observed in the PROMIS Applied Cognition – General Concerns short form item set. One item, "It has seemed like my brain was not working as well as usual" might be singled out for further study. However, in general the short form item set was highly reliable, informative, and invariant across differing race/ethnic, educational, age, gender, and language groups.

Key words: PROMIS®, cognitive concerns, item response theory, differential item functioning, race, ethnicity

Background

Conceptual equivalence of measures implies that questions are understood in the same way by all respondents (Collins, 2003). Differences in race/ethnicity, culture, socio-economic status, education, and gender can lead to systematic measurement error in interpreting survey responses to standardized questionnaires (Warnecke et al., 1997). Differential item functioning (DIF) analysis in the context of item response theory (IRT) examines whether or not the likelihood of item (category) endorsement is equal across subgroups, conditional on the construct or trait level. For example, DIF is present if different groups of individuals (e.g., males and females) at the same levels of the latent construct exhibit different probabilities of individual item responses (Hulin, 1987).

This paper presents the dimensionality, reliability, information functions, and DIF of the Patient Reported Outcome Measurement Information System® (PROMIS®) Applied Cognition - General Concerns, 8 item short form. This is a measure of self-reported cognitive concerns or complaints, and both terms are used interchangeably to describe the construct assessed. Qualitative methods were used to generate DIF hypotheses for subgroups.

Acknowledgment of the salience of subjective cognitive complaints is relatively new within the field of neurology, and more generally cognitive aging. Early studies of sub-

jective cognitive decline focused on memory, e.g., Gurland et al., 1999. Recent findings suggest that subjective complaints are associated with increased risk of dementia (Jessen et al., 2014; Reisberg, Shulman, Torossian, Leng, & Zhu, 2010) and biomarkers of Alzheimer's Disease (Barnes et al., 2006; Sperling et al., 2011) among those presenting with otherwise-normal cognitive function. Subjective cognitive complaints are a key feature of mild cognitive impairment (MCI). However, to date, there is little evidence extant regarding the psychometric performance of such measures, and particularly of their measurement equivalence across subgroups. Moreover, subjective cognitive impairment may be common among people with cancer, especially those undergoing chemotherapy, and this is an important element of health-related quality-of-life for such individuals..

Racial and ethnic differences have been observed in informant-reported cognitive function. For example, examining diagnosis of cognitive impairment no dementia (CIND; based on neuropsychological testing), informant reports of cognitive decline were found to be associated with an increased odds of CIND among Whites, but not African Americans (Potter et al., 2009). Differences have also been observed among Hispanic and non-Hispanic White respondents in self-reported cognitive function. For example, 16.9 % of Hispanic or Latino respondents said that they had experienced confusion or memory loss (CML), which was significantly higher than the 12.1 % among Whites (Centers for Disease Control and Prevention, 2013). Differences in self-reported cognition may also occur by gender. Among older adults, reports of subjective memory have been shown to differ between men and women, with women reporting significantly more memory complaints (Gagnon et al., 1994). Further, in a sample of young adults, males and females tended to assess their divergent thinking (i.e., creativity) across traditionally stereotypic lines (Kaufman, 2006); females rated themselves higher on verbal skills, while males rated themselves higher on general analytic thinking. It is also possible, however, that these results reflect DIF, which is to say, for example, when controlling for the overall level of cognitive complaints, females were more likely to endorse higher verbal skills and males to endorse general analytic thinking. DIF analyses are needed to differentiate between true differences and those attributable to DIF.

Previous psychometric investigations of the PROMIS 8-item Applied Cognition - General Concerns short form have been limited to reliability and convergent validity in a community-dwelling sample of adults (Saffer, Lanting, Koehle, Klonsky, & Iverson, 2015). In that study, participants were 156 adult and older adult (mean age = 52.5, SD = 13.6) medical outpatient members of a multi-disciplinary healthcare center in British Columbia, Canada. Over half the participants were women (55.8 %), married (68.6 %), employed full-time (50.6 %), and had obtained at least a Bachelor's level education (55.1 %). The vast majority of participants (98.7 %) reported English as their dominant language. The Cronbach's alpha internal consistency estimate was high ($\alpha = 0.95$). Becker, Stuifbergen, and Morrison (2012) examined convergent validity with a neuropsychological battery comprised of five tests. The sample ($n = 29$) was of multiple sclerosis patients (69 % non-progressive). The majority (90 %) was female, and highly educated (72 % college graduates or higher), with a mean age of 50 (SD = 7.5). The sample was primarily White (90 %). The test battery included: Controlled Oral Word Association Test (COWAT; Benton, Sivan, Hamsher, Varney, & Spreen, 1983) assessing verbal fluency

and word finding; California Verbal Learning Test (CVLT-II; Delis, Kramer, Kaplan, & Ober, 2000) assessing verbal memory; Brief Visuospatial Memory Test (BVMT; Benedict 1997) assessing nonverbal learning and memory; the Paced Auditory Serial Addition Test (PASAT; Gronwall, 1977) assessing auditory processing speed, flexibility, and calculations; and the Digit Symbol Modalities Test (Smith, 1982) assessing complex scanning and visual tracking. The strongest correlations (r = 0.30) emerged for the PASAT (2-second version) and the BVMT. Test/retest reliability was conducted after a two month delay (r = 0.80). Finally, paired t-test analysis was used to assess statistically significant change from pre to post test, after an eight week cognitive intervention. The observed effect size was large (Cohen's d = 1.25).

As shown in this review, very little analyses of DIF in subjective cognitive assessment measures have been performed. One early analysis (Teresi et al., 2000) examined DIF in five subjective cognition items embedded within a cognitive screening measure. Samples of 866 Latinos, 619 African-Americans, and 360 non-Latino Whites were used to examine item performance. Among the self-report items, one item related to remembering telephone numbers was found to show DIF for Latino's in the direction of a higher probability of difficulty for this group in comparison to the others. An item related to self-reported difficulty remembering names of family or close friends or words was found to be a poor performing item in terms of item discrimination parameters. Little DIF analyses have been performed on the PROMIS Applied Cognition – General Concerns short forms, and virtually no literature exists examining racial and ethnic groups.

Aims

The aim of this paper is to examine the psychometric properties and measurement equivalence of the 8-item PROMIS Applied Cognition - General Concerns scale in an ethnically diverse sample. DIF was examined across race/ethnicity, education, age, gender, and language (Spanish and English) groups.

Methods

Sample generation and description

These data are from individuals with cancer who were selected from cancer registries. The analytic sample sizes for gender were 2,196 males and 3,245 females. The studied group was males in the analysis of gender. The analyses of race/ethnicity included five subgroups, with the reference group designated as non-Hispanic Whites (n = 2,272); the studied groups were: non-Hispanic Blacks (n = 1,121), Hispanics (n = 1,045), and Asians/Pacific Islanders (n = 902). Respondents (n = 133) who indicated multiple ethnic groups were not included in the analysis. The age groups studied were: 21 to 49 (n = 1,199), 50 to 64 (n = 2,008), and 65 to 84 (n = 2,234). The reference group was the 21 to 49 cohort. The respondents were grouped in five education categories: less than high school (n = 968), high school graduate (n = 1,051), some college (n = 1,762), college degree (n = 984), and post graduate degree (n = 641), the latter of which was used as the

reference group. Finally, there were 705 Hispanic respondents interviewed in English (the reference group) and 335 interviewed in Spanish (the studied group). Details of the sample characteristics are provided in an overview article by Jensen, et al. (2016) in this series.

Measure

The PROMIS Applied Cognition – General Concerns scale can be used as an outcome measure in clinical research. The scale consists of eight items measuring self-reported cognitive troubles or deficits. Items were drawn from the PROMIS item bank (Cella et al., 2007), an item repository that can be used by researchers to generate short forms or be administered as computerized adaptive tests. Based on the World Health Organization framework of physical, mental, and social health, nearly 7,000 items available from patient-reported outcome measures in areas such as pain, emotional distress, and physical functioning were reviewed (Becker et al, 2012). The final cognition item bank consists of 34 subjective concerns about one's cognitive ability. This bank includes questions pertaining to the broad domains of memory (e.g., My memory is as good as usual…) and executive function/control (e.g., I have had trouble shifting back and forth between different activities that require thinking…). A domain team was convened with a focus on representing a brief range of the trait or construct represented in the item bank. Domain experts reviewed short forms to give input on the relevance of each item.

The applied cognition – general concerns short form items include, for example: "I have had trouble forming thoughts", "I have had trouble concentrating", and "It has seemed like my brain was not working as well as usual". Each item asks participants to report deficits "within the last 7 days" using five response options: *never, rarely (once), sometimes (2 or three times), often (about once a day), very often (several times a day)*. Based on face validity (depending on which executive function model is referenced) this instrument may be best classified as a self-reported assessment of working memory and executive control because the item content relates to keeping track and forming thoughts which may assess maintenance of content in short-term working memory or episodic buffers. The item, slow thinking may also be related to maintenance in that slower processing speed leaves more time for working-memory contents to decay, thus reducing effective capacity (Salthouse, 1996). The items, pay attention and trouble concentrating reference the executive monitoring system (Shallice, Burgess, & Robertson, 1996). Finally, the item, shifting back and forth relates to the neuropsychological tasks of set shifting, thought to capture one's cognitive flexibility in switching between different tasks or mental states (Miyake et al., 2000).

Psychometric properties and clinical input were both used in the decision making process related to selection of short-form items. Content experts reviewed the items and rankings (based on IRT-based information) and made cuts of 4, 6, and 8 items. The 4 and 6 item short forms are subsets of the 8 item short form.

Procedures and statistical approach

Qualitative analysis and DIF hypothesis generation

Fair and accurate measurement requires that test scores have the same meaning across all relevant groups (Reise & Waller, 2009). 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, education, and diagnosis of health conditions (e.g., cancer).

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., trouble forming thoughts) should depend only on the level of the trait (state), e.g., perceived cognition, 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 perceived cognitive impairment should have the same likelihood of reporting difficulty with memory. If it is theorized that this might not be the case, it would be hypothesized that the item has gender DIF.

Each of the cognitive concerns items was reviewed qualitatively by nine content experts regarding potential sources of DIF. Three of the members of the panel were clinical or counseling psychologists, three were public health professionals, two were gerontologists, and one a geriatrician. They provided hypotheses in terms of presence and direction of DIF.

Quantitative analyses

Descriptive analyses: Item frequencies were evaluated within each subgroup and for the total sample to detect problems relating to skew and empty cell or sparse data (see Hambleton, 2006).

Model assumptions and fit

Unidimensionality: Unidimensionality was assessed with exploratory (principal components estimation) and confirmatory factor analysis. This merged application (Asparouhov & Muthén, 2009) was performed with MPlus software (Muthén & Muthén, 2011), fitting a unidimensional model with polychoric correlations allowing for cross-loadings. The exploratory analyses included tests of scree. The confirmatory process included tests of fit, e.g., Meade, Johnson, and Bradley, 2008; Muthén, 1982, with a focus on the Comparative Fit Index (CFI; Bentler, 1990). However, to avoid complete reliance on model fit indices, such as the CFI, confirmation of the unidimensional model was performed using a bi-factor model (see Cook, Kallen, & Amtmann, 2009). Bi-factor analysis fits a model with a general factor and group factors that capture specific remaining common variance across item subsets uncorrelated with the general factor (Primi, Rocha da Silva,

Rodrigues, Muniz, & Almeida, 2013; Reise, Morizot, & Hays, 2007). Loadings from a traditional unidimensional model (one-factor solution) were compared to those from the bi-factor model, obtained using the Schmid-Leiman (S-L; Schmid-Leiman, 1957; R "psych" package; Rizopoulus, 2009) solution. The procedure required that all items load on the general factor, with the loadings on the group factors adhering to the Schmid-Leiman solution.

The explained common variance (ECV) establishes whether the observed variance/ covariance matrix is close to unidimensionality (Sijtsma, 2009), and reflects the percent of observed variance explained (Reise, 2012). The first random half of a split sample was used to perform exploratory principal component analysis (PCA) and to fit a unidimensional confirmatory factor analysis (CFA) model.

Local independence: Local independence occurs when the respondent's answer to one item has a bearing on the answer to another item. Local independence can affect the estimation of precision-related test information (e.g., inflating reliability estimates); it may also affect discrimination parameters (Embretson & Reise, 2000), and can result in false (positive) DIF detection (Houts & Edwards, 2013). Previous research has shown that many contemporary tests contain item dependencies, and not accounting for these dependencies leads to misleading estimates of item, test, and ability parameters (Zenisky, Hambleton, & Sireci, 2001). The local independence assumption was tested using the generalized and standardized local dependency chi-square statistics (Chen & Thissen, 1997) supported by IRTPRO, version 2.1 (Cai, Thissen, & du Toit, 2011). Values greater than 10 are flagged for review. The procedure included sensitivity analysis whereby one item each from two pairs with elevated LD was removed.

IRT-model fit: Model fit was investigated using the root mean square error of approximation (RMSEA) from IRTPRO (Cai et al., 2011). The criterion for acceptable fit was a value < 0.10.

Anchor items and linking

In this step of the analyses the comparison groups were linked on cognitive complaints and the mean and variance were estimated for the target groups under investigation. The reference group mean was set to 0 and the variance to 1. There are multiple methods that can be employed to derive anchors, a set of DIF-free items (Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006; Wang, Shih, & Sun, 2012; Woods, 2009). The method used here follows an iterative purification process in which a set of "purified" anchor items that do not evidence DIF were identified. A variant of what has been termed the iterative backward all-other test method (Kopf, Zeileis, & Strobl, 2015) was used, which examines *p*-values to remove items with DIF from the anchor. In this procedure the χ^2 statistics resulting from two models were compared, the first with all parameters fixed to be equal for comparison groups, and the second freeing all parameters for the item under investigation. The derived log-likelihood ratio $\chi 2$ statistic was then evaluated for significance. It has been suggested that a minimum of four anchor items be used in establishing the particular latent trait under investigation (Cohen, Cohen, Teresi,

Marchi, & Velez, 1990); additionally, the use of four as contrasted with fewer anchor items has been shown to increase the power for DIF detection (Shih & Wang, 2009).

Sensitivity analyses for anchor item selection: Two sets of sensitivity analyses were performed to examine the effects of local dependencies and the number of anchor items on the results of the DIF analyses. First the number of anchor items was increased to four in instances in which fewer than four were originally identified. This was accomplished by comparing the goodness-of-fit statistics resulting from two nested models used in DIF detection. Second, the rank-order method was used to identify additional items with lower levels of DIF. In this case, the result was the same as the former method because all items had the same number of response categories and hence degrees of freedom. When less than four anchors were available, the analysis was repeated with four anchor items. The items were selected from the top of the hierarchical list of highest to lowest *p*-values associated with the log-likelihood ratio tests described above.

Model for DIF detection

The graded response model (GRM; Samejima, 1969) was used to estimate parameters. The item characteristic curve (ICC) describes the relationship between item response and the underlying attribute measured, e.g., self-perceived cognition difficulties. There are two properties of the ICC for the graded response model: the item difficulty or location parameters (denoted *b*), and the discrimination (denoted *a*) which reflects the steepness of the curve or the degree to which the item is related to the underlying attribute measured. DIF is observed if there are group differences in the ICCs, reflecting unequal probabilities of response, given equal levels of the trait.

DIF detection tests: The Wald test was the primary method used to detect DIF, assessing group differences in IRT parameters. In this process, a model was established in which all parameters were constrained to be equal across comparison groups for anchor items, while the target item parameters were freed to be estimated separately for study groups. A simultaneous joint test of differences was assessed for the *a* and *b* parameters, which includes step down tests for group differences in the discrimination parameter, and conditional tests of the difficulty parameters. Adjustments were made for multiple comparisons.

Sensitivity analyses for DIF detection: An additional DIF assessment model is based on an iterative ordinal logistic regression IRT framework (Crane et al., 2007; Crane, Gibbons, Jolley, & van Belle, 2006; Crane, van Belle, & Larson, 2004) using lordif software (Choi, Gibbons, & Crane, 2011). This method has been used to examine cognitive assessment measures (Crane et al., 2004; Crane et al., 2006; Crane Gibbons, Jolley, & van Belle, 2006; Gibbons et al., 2009). Lordif performs an ordinal (common odds-ratio) logistic regression DIF analysis using IRT theta (θ) estimates as the conditioning variable. The GRM or the generalized partial credit model (GPCM) is used for IRT trait estimation. Items flagged for DIF are treated as unique items for each group to be calibrated separately, and group-specific item parameters are obtained. Items without DIF serve as anchors for IRT calibration. The procedure runs iteratively until the same set of items is flagged over two consecutive iterations, unless anchor items are specified in advance. A

discussion of cutoff values for DIF detection in the context of anchor items can be found in Mukherjee, Gibbons, Kristiansson, and Crane (2013). DIF was identified if the likelihood ratio (LR) $\chi 2$ p-value was less than 0.01, and the McFadden (1974) R^2 was greater than 0.02. (The threshold using the β change criteria was ≥ 0.1; pseudo $R^2 \geq 0.02$).

Details of these methods are discussed in the overview article in this series (Teresi & Jones, 2016). An important point is that while many items may be flagged for significant DIF using the OLR method, interpretation of the findings of DIF must be made only after considering the magnitude of DIF.

Evaluation of DIF magnitude, effect sizes and impact

The expected item and scale scores were examined to determine the magnitude and impact of DIF, respectively (see Figure 1 for examples).

DIF magnitude: The expected score reflects the sum of weighted response probabilities for each item. This information is used to quantify the difference in the average expected item scores using the non-compensatory DIF (NCDIF) index (Raju, van der Linden, & Fleer, 1995), which is part of DFIT (Oshima, Kushubar, Scott, & Raju, 2009; Raju, 1999; Raju, et al., 2009). Additional effect size metrics, T statistics (Wainer, 1993) modified to accommodate polytomous responses (Kim, Cohen, Alagoz, & Kim, 2007) were examined. Further information on these methods is given in this series (Kleinman & Teresi, 2016).

DFIT software was applied after latent trait estimates were derived separately for each group and then equated together with item parameters using EQUATE software (Baker, 1995). When DIF was observed the item was removed from the equating algorithm, thus incorporating new DIF-free equating constants. This iterative purification of equating constants has been shown to reduce type 1 error (Seybert & Stark, 2012).

Cutoff values based on simulation studies (Fleer, 1993; Flowers, Oshima, & Raju, 1999) were used to estimate item-level DIF. Given the five category polytomous response data, a cutoff of 0.096 was applied (Raju, 1999). This cutoff corresponds to an average absolute difference of 0.310, about a third of a point difference on a five point scale (see Raju, 1999; Meade, Lautenschlager, & Johnson, 2007).

Evaluation of DIF impact: Aggregate DIF impact was assessed with expected scale score functions; group differences in these functions provide an overall aggregated measure of DIF impact. DIF-adjusted and unadjusted estimates of the latent cognition complaints construct were compared to determine DIF impact at the individual level. Estimates were adjusted for all items evidencing DIF after the Bonferroni correction. By fixing and freeing parameters and comparing results with and without DIF adjustment, the individual impact was estimated by calculating the number of individual θ estimates that differ by more than 0.5 and 1.0 standard deviations. Additionally, a threshold marker (a cutoff of θ equal to 1) defining individuals as cognitively compromised or not was examined.

Crane and colleagues (2007) used a similar method in calculating the difference between naïve scores that ignore DIF and scores that account for DIF to examine cumulative impact of DIF on individual participants. The distribution of these difference scores is

then examined; for individual-level DIF impact, a box-and-whiskers plot of the difference scores is constructed. (This is shown on the left side of the graphic in Appendix, Figure A3.) The interquartile range is represented in the shaded box and is the middle 50 % of the difference scores. The median of the difference scores is the bolded line (for most panels this value is around zero). The graphic on the right side shows the plot of the difference scores (ordinate) against the initial θ score on the x axis. Positive values on the right panel indicate that accounting for DIF resulted in somewhat lower cognitive concerns scores than those not accounting for DIF. In the third panel showing non-Hispanic Whites vs. Asians/Pacific Islanders, the positive scores indicate that White respondents tended to have lower scores after DIF adjustment across mid to higher ranges of θ. The negative scores indicate that Asian/Pacific Islanders at mid to higher levels of cognitive concerns tend to have higher scores after DIF adjustment. A dotted line shows the mean difference between the initial and DIF-adjusted θ estimates (which in this case is close to zero). In the graphic in the first panel, the individual differences are small, ranging from -0.03 to about 0.03. "Salient" changes refer to changes exceeding the median standard error of the initial score. Differences larger than that value are termed salient individual-level DIF impact. (See Appendix Figure 3A depicting graphics from lordif [Choi et al., 2011], an R software module.)

Evaluation of reliability and information

McDonald's Omega Total (ω_t; McDonald, 1999) was estimated based on the proportion of total common variance explained. Internal consistency was also estimated with Cronbach's alpha (Cronbach, 1951; Cronbach & Meehl, 1955) as well as ordinal alpha based on polychoric correlations (Zumbo, Gadermann, & Zeisser, 2007). An IRT-based reliability statistic was calculated as well, allowing for precision to be estimated at multiple points on the trait (θ) continuum.

Results

Qualitative analysis

Table 1 shows the hypotheses generated for the cognition items. It was hypothesized that conditional on cognitive complaints women would more likely report trouble with forming thoughts and concentrating as contrasted with males. The majority of raters did not posit race/ethnicity DIF hypotheses except for one item where some raters were in agreement that Latinos, in contrast to majority group members would be more likely to report that "my brain was not working as well as usual". Language DIF was posited for one item also suggesting that Spanish speakers would be more likely (conditional on cognitive complaints) to report that they "have had to work really hard to pay attention or I would make a mistake" in comparison to the reference group. Similarly, with respect to education DIF, several expert panelists hypothesized that conditional on cognitive complaints, individuals with higher levels of education would be more likely to endorse responses indicating higher dysfunction with regard to forming thoughts and brain not

Table 1:
DIF hypotheses generated by nine content experts for applied cognition - general concerns items

Item Stem	Gender	Age	Race Ethnicity	Language	Education	Diagnosis
I have had trouble forming thoughts (8a)	3[a] *Women higher impairment (3)*[b]	4 *Older higher impairment (2)*		2	4 *Higher education higher impairment (2)*	4 Depression /anxiety higher impairment (1); Ill higher impairment (1); Cancer higher impairment (1)
My thinking has been slow (4a, 6a, 8a)		4 *Older higher impairment*				3
My thinking has been foggy	2	2 *Older higher impairment*		5	3	2
I have had trouble concentrating (6a, 8a)	3 *Women higher impairment (3)*	3 *Older higher impairment (2)*			2	5 *Cancer higher impairment (2);* terminally ill/pain higher impairment (1); Depression /anxiety higher impairment (1)
I have had to work really hard to pay attention or I would make a mistake (6a, 8a)		2		2 Non-English higher impairment; Spanish higher impairment	4 *Lower education higher impairment (2*)	
It has seemed like my brain was not working as well as usual (4a, 6a, 8a)		4 *Older higher impairment (3)*	2 *Latinos higher impairment*		2 *Higher education higher impairment*	
I have had to work harder than usual to keep track of what I was doing (4a, 6a, 8a)		3 *Older higher impairment (2)*			2	
I have had trouble shifting back and forth between different activities that require thinking (4a, 6a, 8a)		2			2	

[a] Number indicates total number of hypotheses; [b] Number of directional hypotheses; Italicized entries are those with 2 or more ratings in the same direction.

Note: The following short-form 8a item was not included in the analyses: "My problems with memory, concentration, or making mental mistakes have interfered with the quality of my life."

working as well as usual; and that those with lower levels of education would be more likely to endorse the item "I have had to work really hard to pay attention or I would make a mistake". Age-DIF hypotheses were posited for all items; for six out of the eight items, it was hypothesized that conditional on overall cognitive complaints, older individuals would endorse responses that indicate higher levels of cognitive dysfunction in contrast to younger individuals. Directions were not provided for the hypotheses for two items: had to work hard to pay attention and had trouble shifting back and forth. Raters posited directional DIF hypotheses for two items suggesting that (conditional on cognitive complaints) individuals diagnosed with cancer or those terminally ill will be more likely to report trouble with forming thoughts or concentrating than those in the reference group (see Table 1).

Quantitative results

Item and raw score distributions

The distribution as a whole was skewed toward no difficulty with cognition. Thirty four percent of the respondents (1,847 of 5,477) reported no problems; additionally, 48 to 54 % of respondents reported that they never experienced the problems queried by individual items. Only 6 % of respondents received a sum score of 24 to 32 (the maximum), a level that indicates on average having difficulties often or very often.

Test of model assumptions and fit

Unidimensionality: The results present strong evidence that essential unidimensionality was met for all subgroups (Table 2). The scree plot for the total sample provides a graphical representation of the unidimensionality (Appendix Figure A1). For all comparison demographic subgroups the ratio of component 1 to 2 was large (21.0 to 32.0), with the first component accounting for 87 % to 92 % of the variance. A bifactor model from Mplus was used to examine dimensionality further using the second random half of the sample. The results summarized in Table 3 show that the loadings on the single common factor were very high (range of 0.94 to 0.97) and similar in magnitude to those on the general factor in the bifactor model. The high loadings imply intra-item correlations ranging from 0.85 to 0.93. The range of differences between the values of the loadings from the single common factor and that of the general factor was from 0 to 0.04, while the loadings on the group factors were low (0.13 to 0.36), thus providing additional support for unidimensionality. The communality values were also large, ranging from 0.89 to 0.93.

Tests of model fit and unidimensionality: The range of CFI values from the unidimensional CFA analyses from Mplus was from 0.994 to 0.999 (see Appendix Table A1). The ECV, estimated with Pearson correlations ranged from 81.17 to 86.35 (Table 4). The IRT model fit statistic: the RMSEA from IRTPRO for the IRT models ranged from 0.05 to 0.10 across DIF grouping variables, indicating good to acceptable fit (see Appendix, Table A1).

Table 2:
PROMIS applied cognition - general concerns item set: Tests of dimensionality from principal components analysis (eigenvalues by subgroup)

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio Component 1/ Component 2
Total Sample (n = 5,477)					
Eigenvalues	7.251	0.254	0.123	0.104	28.5
Explained Variance	90.6 %	3.2 %	1.5 %	1.3 %	
Random First Half Sample (n = 2,739)					
Eigenvalues	7.282	0.243	0.120	0.097	30.0
Explained Variance	91.0 %	3.0 %	1.5 %	1.2 %	
Females (n = 3,245)					
Eigenvalues	7.221	0.268	0.134	0.110	26.9
Explained Variance	90.3 %	3.4 %	1.7 %	1.4 %	
Males (n = 2,196)					
Eigenvalues	7.268	0.246	0.116	0.096	29.5
Explained Variance	90.9 %	3.1 %	1.5 %	1.2 %	
Age 21 to 49 (n = 1,199)					
Eigenvalues	7.268	0.254	0.122	0.101	28.6
Explained Variance	90.9 %	3.2 %	1.5 %	1.3 %	
Age 50 to 64 (n = 2,008)					
Eigenvalues	7.285	0.231	0.123	0.103	31.5
Explained Variance	91.1 %	2.9 %	1.5 %	1.3 %	
Age 65 to 84 (n = 2,234)					
Eigenvalues	7.155	0.297	0.139	0.112	24.1
Explained Variance	89.4 %	3.7 %	1.7 %	1.4 %	
Race/Ethnicity: Non-Hispanic White (n = 2,272)					
Eigenvalues	7.325	0.229	0.116	0.095	32.0
Explained Variance	91.6 %	2.9 %	1.5 %	1.2 %	
Race/Ethnicity: Non-Hispanic Black (n = 1,121)					
Eigenvalues	7.240	0.272	0.131	0.100	26.6
Explained Variance	90.5 %	3.4 %	1.6 %	1.3 %	
Race/Ethnicity: Hispanic (n = 1,045)					
Eigenvalues	7.186	0.280	0.130	0.113	25.7
Explained Variance	89.8 %	3.5 %	1.6 %	1.4 %	

Statistic	Component 1	Component 2	Component 3	Component 4	Ratio Component 1/ Component 2
Race/Ethnicity: Non-Hispanic Asians/Pacific Islanders (n = 902)					
Eigenvalues	7.171	0.277	0.148	0.128	25.9
Explained Variance	89.6 %	3.5 %	1.9 %	1.6 %	
Education: Less Than High School (n = 968)					
Eigenvalues	7.126	0.289	0.149	0.111	24.7
Explained Variance	89.1 %	3.6 %	1.9 %	1.4 %	
Education: High School (n = 1,051)					
Eigenvalues	7.211	0.266	0.145	0.105	27.1
Explained Variance	90.1 %	3.3 %	1.8 %	1.3 %	
Education: Some College (n = 1,762)					
Eigenvalues	7.327	0.238	0.112	0.087	30.8
Explained Variance	91.6 %	3.0 %	1.4 %	1.1 %	
Education: College Degree (n = 984)					
Eigenvalues	7.242	0.265	0.123	0.115	27.3
Explained Variance	90.5 %	3.3 %	1.5 %	1.4 %	
Education: Graduate Degree (n = 641)					
Eigenvalues	7.297	0.241	0.121	0.104	30.3
Explained Variance	91.2 %	3.0 %	1.5 %	1.3 %	
Hispanics Interviewed in English (n = 705)					
Eigenvalues	7.267	0.266	0.118	0.106	27.3
Explained Variance	90.8 %	3.3 %	1.5 %	1.3 %	
Hispanics Interviewed in Spanish (n = 335)					
Eigenvalues	6.990	0.333	0.170	0.149	21.0
Explained Variance	87.4 %	4.2 %	2.1 %	1.9 %	

Table 3:

PROMIS applied cognition - general concerns item set: Item loadings (λ) from the unidimensional confirmatory factor analysis (Mplus) for the first half of the random sample (n = 2,739), Schmid-Leiman bi-factor model with two and three group factors (performed with R for the second random half of the sample) and Mplus bi-factor two group solution for the second random half of the sample (n = 2,738)

Item description	One Fact.* λ (s.e.)	Schmid-Leiman Bi-Factor Three and Two Group Factor Solutions					Mplus Bi-Factor Two Group Factor Solution (Based on S-L** Result)						
		G λ	F1 λ	F2 λ	F3 λ	h^2	G λ	F1 λ	F2 λ	h^2	G λ (s.e.)	F1 λ (s.e.)	F2 λ (s.e.)
I have had trouble forming thoughts	0.94 (0.003)	0.89		0.32		0.89	0.91		0.24	0.89	0.90 (0.005)		0.27 (0.015)
My thinking has been slow	0.95 (0.003)	0.91		0.32		0.93	0.93		0.25	0.93	0.92 (0.004)		0.36 (0.015)
My thinking has been foggy	0.95 (0.003)	0.92		0.23		0.91	0.93			0.91	0.94 (0.003)		0.13 (0.009)
I have had trouble concentrating	0.95 (0.003)	0.92				0.90	0.93			0.89	0.96 (0.002)		
I have had to work really hard to pay attention or I would make a mistake	0.95 (0.003)	0.92	0.23			0.91	0.92	0.24		0.91	0.94 (0.003)	0.15 (0.011)	
It has seemed like my brain was not working as well as usual	0.96 (0.002)	0.93				0.91	0.93			0.91	0.96 (0.002)		
I have had to work harder than usual to keep track of what I was doing	0.97 (0.002)	0.93	0.26			0.93	0.93	0.25		0.93	0.94 (0.003)	0.20 (0.013)	
I have had trouble shifting back and forth between different activities that require thinking	0.96 (0.002)	0.93	0.26			0.92	0.93	0.25		0.92	0.93 (0.004)	0.31 (0.015)	

* Geomin (oblique) rotation

** Schmid-Leiman bi-factor model, three group factors solution with no loadings on the 3rd group factor; 2 group factor solution did not converge in MPlus

Note: Comparative fit index (CFI) for the Mplus one-factor solution is 0.996 and for the bi-factor solution is 0.999

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

Local independence: In general, the local dependence values (not shown) were in the acceptable range. However, two sets of items showed elevated values for LD statistics: Item 1 – trouble forming thoughts paired with Item 2 – thinking has been slow (28.2 for non-Hispanic Black respondents and 22.1 for respondents with less than high school education) and Item 7 – has to work harder to keep track paired with Item 8 – trouble shifting activities (20.0 for non-Hispanic Black respondents and 27.6 for respondents with less than high school education).

Reliability estimates

The estimates of internal consistency were high; Cronbach's alphas ranged from 0.967 to 0.977. The ordinal alpha using polychoric correlations ranged from 0.979 to 0.987; the omega total values (Table 4) ranged from 0.980 to 0.987. The IRT-generated reliability estimates at points along the latent construct (θ) inform about the measurement precision.

Table 4:
PROMIS applied cognition - general concerns item set. Reliability statistics Alpha, Omega Total and explained common variance (ECV) for the total sample and demographic subgroups ("Psych" R package)

	Cronbach's Alpha	**Ordinal Alpha**	**McDonald's Omega**	**ECV**
Total Sample	0.975	0.985	0.985	85.113
Random Second Half of the Sample	0.974	0.985	0.985	84.581
Age 21 to 49 years	0.977	0.986	0.986	85.942
Age 50 to 64 years	0.976	0.986	0.986	85.612
Age 65 to 84 years	0.971	0.983	0.983	83.145
Male	0.974	0.986	0.986	84.673
Female	0.975	0.985	0.985	85.050
Non-Hispanic White	0.977	0.987	0.987	86.176
Non-Hispanic Black	0.975	0.985	0.985	84.891
Hispanic	0.974	0.984	0.984	84.563
Non-Hispanic Asian/Pacific Islander	0.971	0.983	0.984	83.346
Less Than High School	0.972	0.982	0.983	83.781
High School Degree	0.974	0.984	0.985	84.707
Some College	0.977	0.987	0.987	86.350
College Graduate	0.973	0.985	0.985	84.399
Graduate Degree	0.974	0.986	0.986	84.893
Hispanics Interviewed in English	0.976	0.986	0.986	85.799
Hispanics Interviewed in Spanish	0.967	0.979	0.980	81.174

Table 5:
PROMIS applied cognition - general concerns item set: Item response theory (IRT) reliability estimates at varying levels of the attribute (θ) estimate based on results of the IRT analysis (IRTPRO) for total sample and demographic subgroups

Cognition (Theta)	IRT Reliability																
	Total	F	M	Age 21-49	Age 50-64	Age 65-84	NHW	NHB	Hisp.	NH API	<HS	HS	Some Coll.	Coll.	Grad.	Lang. Engl.	Lang. Span.
-1.2	0.56	0.62	0.52	0.66	0.56	0.55	0.53	0.56	0.63	0.58	0.66	0.57	0.56	0.54	0.53	0.62	0.68
-0.8	0.77	0.86	0.65	0.90	0.78	0.71	0.70	0.75	0.86	0.79	0.89	0.78	0.78	0.70	0.65	0.86	0.88
-0.4	0.95	0.97	0.90	0.98	0.95	0.92	0.93	0.94	0.97	0.95	0.98	0.95	0.96	0.92	0.89	0.97	0.97
0.0	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98
0.4	0.98	0.98	0.98	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
0.8	0.98	0.98	0.99	0.98	0.98	0.98	0.99	0.99	0.98	0.98	0.98	0.99	0.98	0.98	0.99	0.98	0.98
1.2	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.98	0.98	0.98	0.99	0.99	0.98	0.98	0.99	0.98
1.6	0.99	0.98	0.99	0.98	0.99	0.98	0.99	0.99	0.98	0.98	0.98	0.99	0.99	0.98	0.99	0.99	0.98
2.0	0.98	0.97	0.98	0.94	0.97	0.98	0.98	0.97	0.97	0.98	0.95	0.98	0.97	0.98	0.98	0.96	0.98
2.4	0.89	0.86	0.91	0.75	0.88	0.95	0.89	0.87	0.86	0.95	0.79	0.90	0.86	0.95	0.95	0.83	0.94
2.8	0.65	0.63	0.68	0.55	0.65	0.81	0.65	0.65	0.63	0.78	0.59	0.67	0.62	0.81	0.79	0.59	0.81
Overall (Average)	**0.88**	**0.89**	**0.87**	**0.88**	**0.88**	**0.89**	**0.87**	**0.88**	**0.89**	**0.90**	**0.89**	**0.89**	**0.88**	**0.89**	**0.88**	**0.89**	**0.92**

Note: Reliability estimates are calculated for θ levels for which there are respondents. NHW, non-Hispanic White; NHB, non-Hispanic Black; Hisp., Hispanic; NHAPI, non-Hispanic Asian/ Pacific Islander; HS, high school; Coll., college; Grad., graduate school; Lang., language

The estimates, limited to θ levels where respondents were observed were high: 0.88 for the total sample and from 0.87 to 0.92 for individual subgroups (see Table 5). The estimates were low at θ level -1.2 (0.52 for males to 0.68 for respondents interviewed in Spanish). For the total sample, estimates were in the upper 0.90's at θ levels from -0.4 to 2.0.

Shown in Table 6 are the graded response item parameters and their standard errors for the total sample. For all items, the *a* (discrimination) parameters are high, ranging from 4.35 for Item 1 – "I have had trouble forming thoughts" to 6.26 for Item 7 – "I have had to work harder than usual to keep track of what I was doing". Similar patterns hold for all subgroups, although there was some variation (see Appendix Table A2). The *a* parameters ranged from 3.45 (Item 1 – trouble forming thoughts for respondents interviewed in Spanish) to 7.35 (Item 7 – harder to keep track of what I was doing for non-Hispanic Whites).

Table 6:
PROMIS applied cognition - general concerns item set: Item response theory (IRT) item parameters and standard error estimates (using IRTPRO) for the total sample (n = 5,459)

Item description	***a***	**s.e. of *a***	***b1***	**s.e.**	***b2***	**s.e.**	***b3***	**s.e.**	***b4***	**s.e.**
I have had trouble forming thoughts	4.35	0.11	0.08	0.02	0.75	0.02	1.36	0.02	1.85	0.03
My thinking has been slow	4.96	0.12	-0.04	0.02	0.60	0.02	1.26	0.02	1.74	0.03
My thinking has been foggy	5.50	0.15	0.10	0.02	0.73	0.02	1.36	0.02	1.85	0.03
I have had trouble concentrating	5.65	0.15	-0.04	0.02	0.62	0.02	1.29	0.02	1.74	0.03
I have had to work really hard to pay attention or I would make a mistake	5.63	0.15	0.07	0.02	0.72	0.02	1.31	0.02	1.81	0.03
It has seemed like my brain was not working as well as usual	6.01	0.17	0.01	0.02	0.61	0.02	1.20	0.02	1.67	0.02
I have had to work harder than usual to keep track of what I was doing	6.26	0.18	0.05	0.02	0.65	0.02	1.20	0.02	1.65	0.02
I have had trouble shifting back and forth between different activities that require thinking	5.90	0.17	0.09	0.02	0.71	0.02	1.27	0.02	1.70	0.03

DIF results

Appendix Tables A3-A7 show detailed DIF results for race/ethnicity, education, age, gender, and interview language. Tables 7-10 are summaries of DIF results. Table 7 shows the results for race/ethnicity. Only one item, "It has seemed like my brain was not working as well as usual", showed DIF by both DIF detection methods: the Wald tests and ordinal logistic regression after Bonferroni correction. This item also evidenced T statistics above threshold for Hispanic and non-Hispanic Black respondents compared to the non-Hispanic White respondents; however, NCDIF magnitude estimates were below threshold for all items and all comparison groups. Conditional on cognitive complaints, Hispanic and Black respondents had a lower probability (higher *b* parameters) of endorsing the item in the cognitive complaints direction as compared to non-Hispanic White respondents (see Appendix Table A3). The magnitude of DIF is also reflected in the degree of non-overlap in the expected item score function curves in Figure 1.

The item brain not working as well as usual was also flagged for DIF in education group comparisons by both the Wald and OLR-based tests after Bonferroni correction; however, the magnitude statistics were all under the thresholds (see Table 8). The DIF statistic was significant for the respondents with less than high school education as compared to those with a graduate degree; conditional on cognitive complaints (θ), those with less than high school education had a lower probability of endorsing the item in the cognitive complaints direction than those with a graduate school education (Appendix Table A4).

Table 9 presents DIF results for both gender and age group comparisons. No item was flagged for DIF by both methods for gender comparisons (see also Appendix Table A6). For age group comparisons two items showed DIF by both the Wald and OLR-based tests: "I have had to work really hard to pay attention or I would make a mistake;" and "I have had trouble shifting back and forth between different activities that require thinking". The latter item, trouble shifting between activities was significant for both age comparisons (50 to 64 years and 65 to 85 years) vs. the youngest group (aged 21 to 49). Conditional on cognitive complaints (θ), people in both older age groups had a higher likelihood of endorsing the item in the cognitive complaints direction (lower *b* parameters). The item, working hard to pay attention showed significant DIF in the oldest (65 to 84) vs. the youngest (21 to 49) age group comparison. Older respondents had a higher likelihood of endorsing that item in the cognitive complaints direction, conditional on the cognition (θ) estimate (See Appendix Table A5). No magnitude results were above the thresholds.

No item showed DIF by both methods for the Spanish or English language of interview comparisons. The results are summarized in Table 10 and Appendix Table A7.

Table 7:
PROMIS applied cognition - general concerns item set: Differential item function (DIF) results. Race/ethnicity subgroup comparisons

Item description	IRTPRO			lordif			Magnitude (NCDIF)			Effect Size T1		
	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI	White vs. Black	White vs. Hisp.	White vs. NHAPI
I have had trouble forming thoughts				U*; NU*		U*; NU	0.0006	0.0003	0.0005	-0.0135	-0.0055	0.0185
My thinking has been slow				U*; NU*	U	U*; NU	0.0004	0.0003	0.0012	0.0073	0.0064	-0.0204
My thinking has been foggy		U	NU	U*; NU*	U*; NU	U*	0.0021	0.0036	0.0019	-0.0360	0.0277	0.0306
I have had trouble concentrating			NU	U*; NU*		U*; NU*	0.0005	0.0018	0.0045	0.0026	-0.0335	0.0481
I have had to work really hard to pay attention or I would make a mistake				U*; NU*		U*; NU	0.0007	0.0014	0.0011	-0.0004	-0.0245	-0.0180
It has seemed like my brain was not working as well as usual	U*	U*	U	U*; NU*	U*	U*; NU*	0.0201	0.0205	0.0060	0.1084†	0.1148†	0.0565
I have had to work harder than usual to keep track of what I was doing				U*; NU*		U*; NU*	0.0012	0.0013	0.0031	0.0241	-0.0130	-0.0327
I have had trouble shifting back and forth between different activities that require thinking			U	U*; NU*		U*; NU*	0.0009	0.0054	0.0074	-0.0123	-0.0568	-0.0650

*Asterisks indicate significance after adjustment for multiple comparisons. All non compensatory DIF (NCDIF) values were smaller than the threshold (0.0960) † Indicates value above threshold of 0.10.
NU = Non-uniform DIF involving the discrimination parameters; U = Uniform DIF involving the location parameters.
For the lordif analyses, uniform and non-uniform DIF were determined using likelihood ratio chi-square tests. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF of high magnitude was not detected using the pseudo R^2 measures of Cox & Snell (1989), Nagelkerke (1991), and McFadden (1974) or with the change in β criterion. The threshold for β change was ≥ 0.1; pseudo R^2 was ≥ 0.02.

Table 8:
PROMIS applied cognition - general concerns item set: Differential item function (DIF) results. Education subgroups comparisons

Item description	IRTPRO				lordif				Magnitude (NCDIF)				Effect Size T1			
	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS	GD vs. CD	GD vs. Some Coll.	GD vs. HS	GD vs. No HS
I have had trouble forming thoughts			U			U	U		0.0006	0.0030	0.0088	0.0056	0.0111	-0.0421	-0.0717	-0.0612
My thinking has been slow									0.0029	0.0002	0.0011	0.0005	0.0413	-0.0013	-0.0053	-0.0062
My thinking has been foggy						U			0.0011	0.0015	0.0011	0.0043	-0.0037	-0.0249	-0.0218	-0.0375
I have had trouble concentrating									0.0004	0.0001	0.0005	0.0010	-0.0098	0.0055	-0.0100	-0.0121
I have had to work really hard to pay attention or I would make a mistake			U			U*			0.0047	0.0033	0.0050	0.0045	-0.0468	-0.0433	-0.0488	-0.0553
It has seemed like my brain was not working as well as usual				U*; NU			U*	U*	0.0009	0.0046	0.0110	0.0385	0.0015	0.0483	0.0750	0.1560 †
I have had to work harder than usual to keep track of what I was doing							U	U	0.0005	0.0021	0.0041	0.0046	-0.0011	0.0226	0.0310	0.0464
I have had trouble shifting back and forth between different activities that require thinking									0.0017	0.0006	0.0026	0.0018	0.0060	0.0152	0.0095	-0.0259

*Asterisks indicate significance after adjustment for multiple comparisons. All NCDIF values were smaller than the threshold (0.0960) † Indicates value above threshold of 0.10. GD, graduate degree; CD, college degree

NU = Non-uniform DIF involving the discrimination parameters; U = Uniform DIF involving the location parameters.

For the lordif analyses, uniform and non-uniform DIF were determined using likelihood ratio chi-square tests. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF of high magnitude was not detected using the pseudo $R2$ measures of Cox & Snell, Nagelkerke, and McFadden or with the change in β criterion. The threshold for β change was ≥ 0.1; pseudo $R2$ was ≥ 0.02.

Table 9:

PROMIS applied cognition - general concerns item set: Differential item function (DIF) results. Gender and age subgroups comparisons

Item description	IRTPRO			lordif			Magnitude (NCDIF)			Effect Size T1		
	Gender	Age		Gender	Age		Gender	Age		Gender	Age	
		21-49 vs. 50-64	21-49 vs. 65-84		21-49 vs. 50-64	21-49 vs. 65-84		21-49 vs. 50-64	21-49 vs. 65-84		21-49 vs. 50-64	21-49 vs. 65-84
I have had trouble forming thoughts				U*		NU*	0.0020	0.0013	0.0008	0.0328	0.0227	0.0168
My thinking has been slow				U*		U; NU*	0.0003	0.0016	0.0006	0.0081	0.0223	0.0006
My thinking has been foggy				U*		U; NU*	0.0001	0.0024	0.0004	-0.0040	0.0197	0.0124
I have had trouble concentrating				U*	NU	NU*	0.0004	0.0008	0.0055	0.0133	0.0132	0.0546
I have had to work really hard to pay attention or I would make a mistake		U	U*	U*	U	U*	0.0013	0.0019	0.0053	-0.0247	-0.0274	-0.0342
It has seemed like my brain was not working as well as usual		U	U	U*; NU		U; NU	0.0009	0.0017	0.0023	0.0196	0.0093	0.0052
I have had to work harder than usual to keep track of what I was doing		U		U*; NU		U*	0.0001	0.0017	0.0048	-0.0047	-0.0016	-0.0134
I have had trouble shifting back and forth between different activities that require thinking		U*; NU*	U*	U*; NU	U*	U*; NU	0.0022	0.0041	0.0062	-0.0310	-0.0429	-0.0541

*Asterisks indicate significance after adjustment for multiple comparisons. All NCDIF values were smaller than the threshold (0.0960) † Indicates value above threshold of 0.10.

NU = Non-uniform DIF involving the discrimination parameters; U = Uniform DIF involving the location parameters.

For the lordif analyses, uniform and non-uniform DIF were determined using likelihood ratio chi-square tests. Uniform DIF is obtained by comparing the log likelihood values from models one and two. Non-uniform DIF is obtained by comparing the log likelihood values from models two and three. DIF of high magnitude was not detected using the pseudo R^2 measures of Cox & Snell, Nagelkerke, and McFadden or with the change in β criterion. The threshold for β change was ≥ 0.1; pseudo R^2 was ≥ 0.02.

Table 10:
PROMIS applied cognition - general concerns item set: Differential item function (DIF) results Language subgroups comparison, English vs. Spanish interview, for Hispanics only

Item description	IRTPRO	lordif	Magnitude (NCDIF)	Effect Size T1
I have had trouble forming thoughts			0.0031	-0.0477
My thinking has been slow			0.0009	-0.0214
My thinking has been foggy	U	U*	0.0271	0.1309†
I have had trouble concentrating			0.0064	-0.0652
I have had to work really hard to pay attention or I would make a mistake			0.0071	-0.0545
It has seemed like my brain was not working as well as usual		U*	0.0237	0.1190†
I have had to work harder than usual to keep track of what I was doing			0.0038	0.0189
I have had trouble shifting back and forth between different activities that require thinking	NU		0.0114	-0.0209

*Asterisks indicate significance after adjustment for multiple comparisons. All NCDIF values were smaller than the threshold (0.0960) † Indicates value above threshold of 0.10.

NU = Non-uniform DIF involving the discrimination parameters; U = Uniform DIF involving the location parameters.

For the lordif analyses, uniform and non-uniform DIF were determined using likelihood ratio chi-square tests.

Uniform DIF is obtained by comparing the log likelihood values from models one and two.

Non-uniform DIF is obtained by comparing the log likelihood values from models two and three.

DIF of high magnitude was not detected using the pseudo R^2 measures of Cox & Snell, Nagelkerke, and McFadden or with the change in β criterion.

The threshold for β change was ≥ 0.1; pseudo R^2 was ≥ 0.02.

Sensitivity analysis

For the age analysis only three items were originally selected as anchor items: “I have had trouble forming thoughts”, “My thinking has been slow” and “My thinking has been foggy”. In the sensitivity analysis, the item, “It has seemed like my brain was not working as well as usual” was added to the anchor set. There was no change in the item DIF designations.

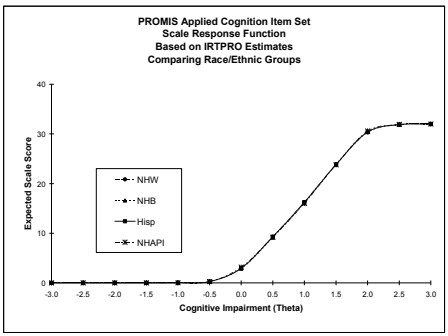
The second set of sensitivity analyses was performed to correct for high local dependency among the items by excluding one of the items in a pair with the highest LD statistic.

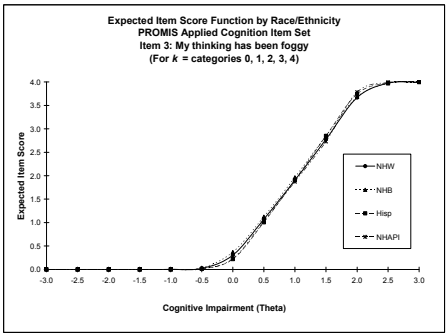
In the race/ethnicity DIF analysis, the item, “I have had trouble forming thoughts” was excluded (high LD was present for the item paired with the item: thinking has been slow). An additional item became significant after Bonferroni correction: “My thinking has been foggy” in the Hispanic vs. non-Hispanic White respondent comparison and the item, “I have had trouble concentrating” changed to significant, but only before the correction for multiple comparisons for the Asians/Pacific Islanders vs. White respondents comparison. In the analyses of education DIF, the item, “I have had to work harder to keep track of what I was doing” evidenced high LD values when paired with the item, “I have had trouble shifting back and forth between different activities that require thinking”. After excluding the item, harder to work harder to keep track, no additional items with DIF after Bonferroni correction were identified; however two items with DIF no longer evidenced DIF: trouble forming thoughts and “I have had to work really hard to pay attention or I would make a mistake”.

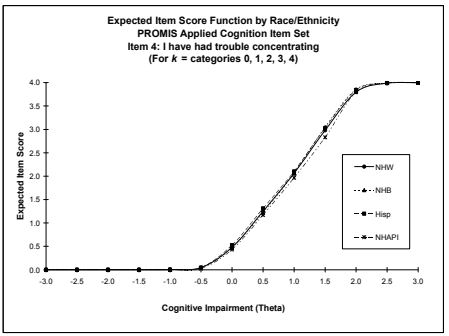
Aggregate impact

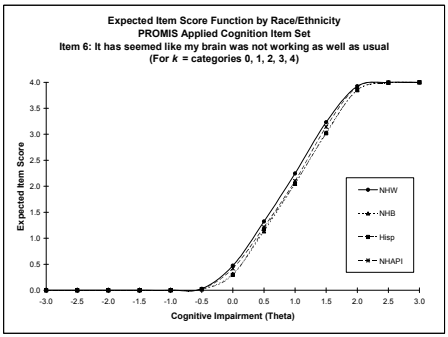
As shown in Figure 1, there was no evident scale level impact. All group curves were overlapping for all comparisons.

Figure 1:
PROMIS applied cognition - general concerns item set: Expected scale and item score functions for race/ethnicity subgroups









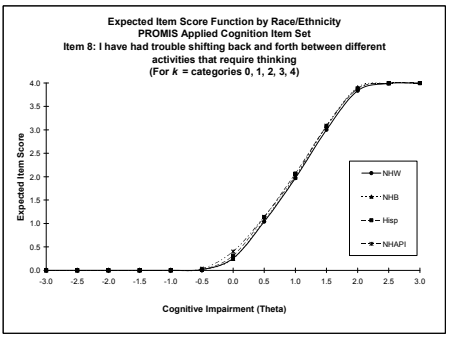
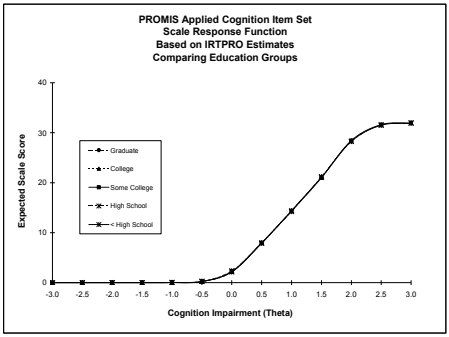
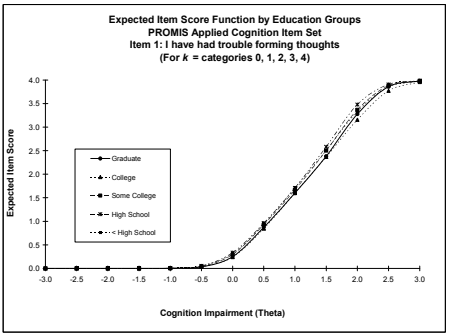
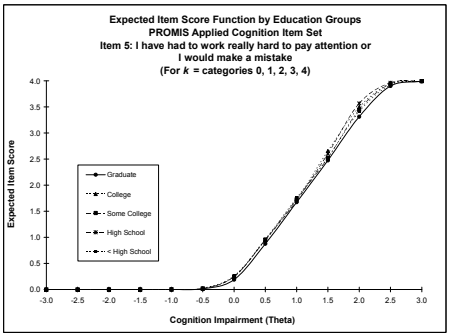


Figure 1 - cont.:
PROMIS applied cognition - general concerns item set: Expected scale and item score functions for education subgroups







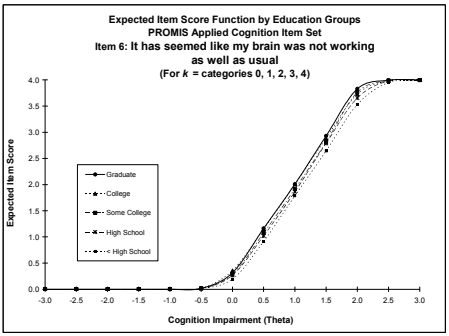
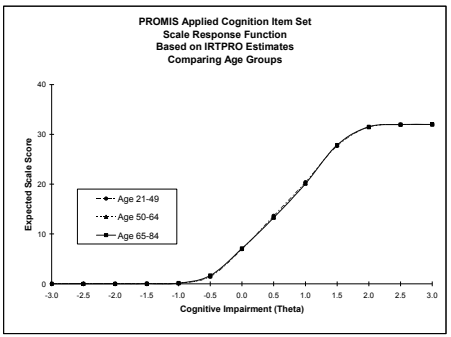
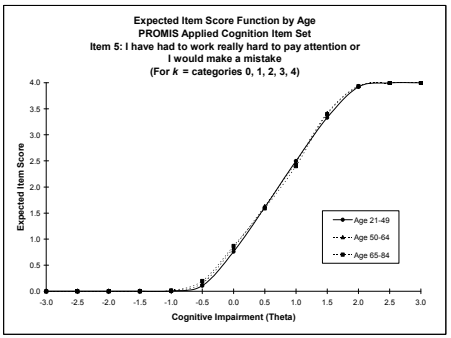
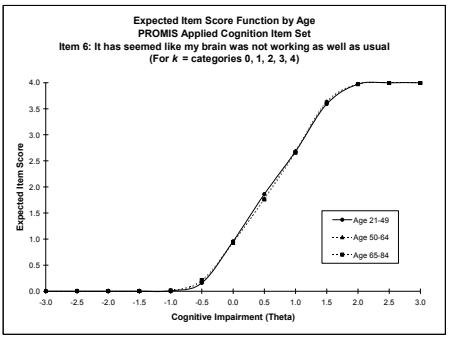
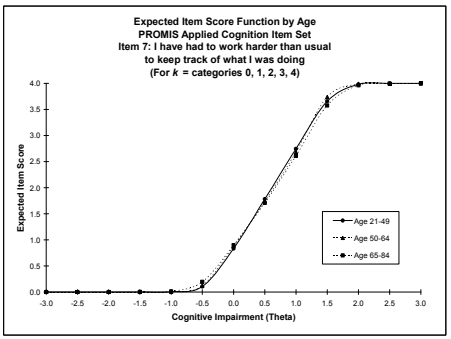


Figure 1 - cont.:
PROMIS applied cognition - general concerns item set: Expected scale and item score functions for age subgroups









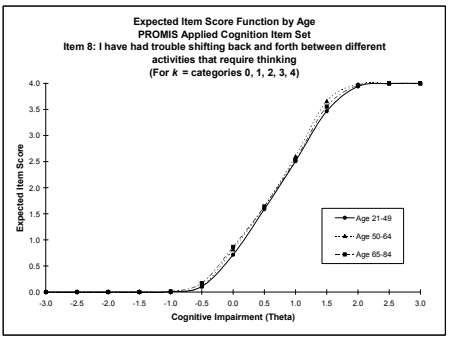


Figure 1 - cont.:
PROMIS applied cognition - general concerns item set: Expected scale and item score functions for gender subgroups

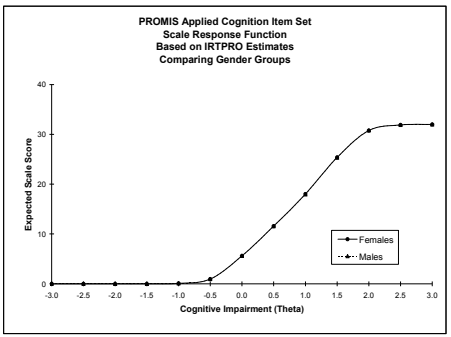
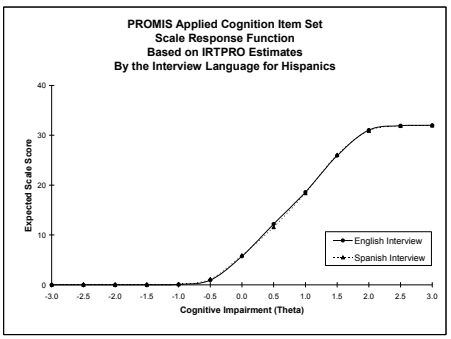
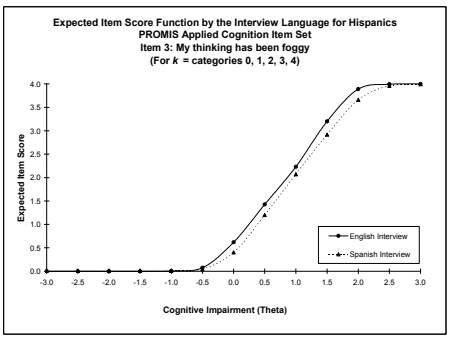
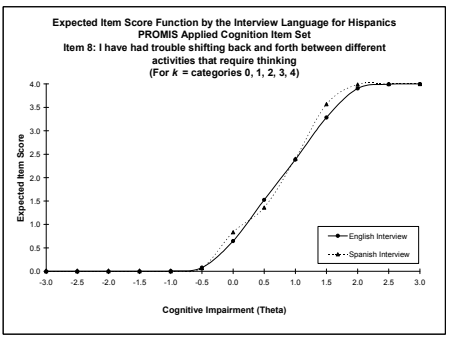


Figure 1 - cont.:
PROMIS applied cognition - general concerns item set: Expected scale and item score functions for language of the interview subgroups, Hispanics only







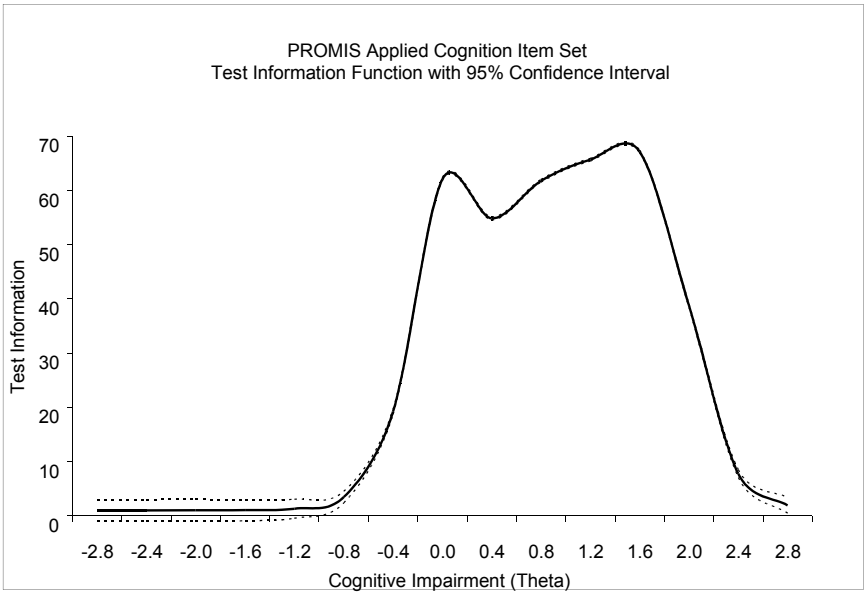
Individual impact

Analyses were performed evaluating individual impact by comparing θs estimated accounting and not accounting for DIF. The analysis was limited to the race/ethnicity, education, and age subgroups because there was no DIF observed in the primary analyses for the gender subgroups and only minor DIF, non-significant after the Bonferroni correction, for the language subgroups. Individual impact for all comparative subgroups was minimal. The correlation of the two θ estimates was 1.0 for all three sets. There were only minor shifts in the θ estimates for all groups in both directions, some higher after the DIF adjustment and some lower. None was greater than 0.5 standard deviations. Using a cutoff point of $\theta \geq 1.0$ to classify respondents as cognitively challenged or not, there were no changes in this designation when comparing the two θ estimates. As shown in the graphics in Appendix Figure A3, the individual difference scores between unadjusted and DIF-adjusted scores are very small, ranging from -0.03 to 0.03 across most analyses.

Information

The item-level and scale information functions were examined for the total sample (see Figure 2 and Appendix Figure A2). Most scale information is supplied in the θ range from 0 to 1.6 with the peak of 67.2 at θ level = 1.6. The information function is also

Figure 2:
PROMIS applied cognition - general concerns item set: Test information function (IRTPRO)
Total sample



slightly bimodal dipping to 54.9 at $\theta = 0.4$. Precision, expressed as the standard error of measurement is the inverse square root of information. The observed values (relatively low standard error of measurement about 0.12 to 0.14 at the peaks) support the high precision of the scale. The item, "I have had to work harder than usual to keep track of what I was doing" was the most informative with the peak information = 10.6 at θ level 1.2 followed by the item, "It has seemed like my brain was not working as well as usual" with the peak information = 9.7 at θ level 1.2. The item, "I had trouble forming thoughts" with peak information = 5.4 at θ of 1.6 was the least informative. (It is noted that the values in Appendix Figure A2 are slightly different from those cited here because the curves in the graphs have been smoothed.)

Discussion

Measurement of self-reported cognition can be valuable, particularly in clinical settings, for example, serving as a resource-effective (e.g., time-based) method for ascertaining the impact of drug treatments. Perhaps more relevant is the association of subjective cognitive complaints to the diagnosis of mild cognitive impairment (MCI). The most problematic feature of this classification is the large number of people who adapt or otherwise revert back to normal function. It is well known that not all patients with MCI deteriorate. In fact some patients appear to improve over time (Ingles, Fisk, Merry, & Rockwood, 2003; Wolf et al., 1998). In one study, over almost 3 years, 19.5 % of those classified as MCI recovered, and an additional 61 % neither improved nor deteriorated (Wolf et al., 1998). The diagnosis of MCI incorporates concerns regarding a change in cognition that includes self-report and/or proxy reports of cognition (Albert et al., 2011). MCI was reclassified as mild neurocognitive disorder (mNCD) in the latest DSM revision (American Psychiatric Association, 2013). Perhaps this diagnostic reclassification should coincide with more vigorous scrutiny of the self-reported features of the disorder. This would entail an increase in evaluation of instrument performance relating to precision and accuracy, and moving beyond classical dimensions of assessments. It is unknown if the PROMIS Applied Cognition – General Concerns short form will be useful in the classification of non-amnestic MCI.

Descriptive measures of DIF magnitude are meaningful components of analysis and interpretation; because sample size can influence statistical significance thresholds, magnitude assessment is a required step in DIF analyses. A framework for empirical decisions relating to item performance must include an evaluation of magnitude. Despite finding significant differences in parameter estimates for many items, these results did not meet thresholds for meaningful DIF. For example, using the lordif methodology, DIF was not detected using the pseudo-R^2 measures of Cox and Snell (1989), Nagelkerke (1991) and McFadden (1974) or with the change in β criterion. (The threshold for β change was ≥ 0.1; pseudo R^2 was ≥ 0.02.)

The most robust evidence for item-level DIF across subgroups was observed for the item, "It has seemed like my brain was not working as well as usual". Further, within groups (i.e., age), DIF was salient for "I have had to work really hard to pay attention or

I would make a mistake;" and "I have had trouble shifting back and forth between different activities that require thinking". Given these findings, the items should be examined in other cross-validation samples.

Cognitive change in cancer patients has been reported after treatment. For example, Shilling, Jenkins, Morris, Deutsch, and Bloomfield (2005) observed that subjects receiving chemotherapy (n = 50) for breast cancer had more than double the odds (OR = 2.25) of declining on measures of working memory than did the healthy control group (n = 43). This is particularly relevant to our study of items pertaining to forming thoughts and concentrating. Having said this, not all forms of treatment will impact cognition. Joly et al. (2006) found that treatment for prostate cancer using androgen did not impact cognition, including self-reported/subjective cognitive function as compared with controls. Perhaps the elevated local dependence for selected items suggests that despite assessing several different neuropsychological constructs of function, such as speed (my thinking has been slow) verses attentional control (hard to pay attention); individuals may have difficulty differentiating in their self-reports, domains that are more clearly separable using neuropsychological tests.

Based on the cognitive aging literature, it was hypothesized that DIF might be observed for age, with older respondents reporting greater complaints, conditional on cognition. The content experts also hypothesized that conditional on cognition older respondents would report higher impairment for six out of eight items. The two items that showed DIF for both the Wald and OLR-based tests: "I have had to work really hard to pay attention or I would make a mistake" and "I have had trouble shifting back and forth between different activities that require thinking", both in the direction of more cognitive concerns for older respondents were also hypothesized to show DIF; however the direction was not specified. Set shifting tasks of executive function often indicate age-related impairment. These findings are in line with the observation that executive functions are sensitive to age-related decline (Salthouse, Atkinson, & Berish, 2003). However, self-reported cognition can show somewhat modest convergent validity with neuropsychological measures in clinical and nonclinical populations (Becker, et al., 2012; Johnco, Wothrich, & Rapee, 2014), and both may be useful in assessing overall cognitive function. A parallel observation can be found in physical decline: there is evidence that self-reported disability (e.g., getting around the house or walking upstairs) and performance based measures (e.g., walk-time) are comparable to each other, but usually measure different aspects of functioning (Coman & Richardson, 2006). Combining information from self-report and performance measures has been shown to increase prognostic value for physical function, particularly in high-functioning older adults (Reuben et al., 2004).

DIF was hypothesized for the item brain not working as well as usual, in the direction of higher self-reported impairment for Latinos and respondents with higher education. After the Bonferroni correction, this item showed significant DIF in both primary and sensitivity analyses for both the race/ethnicity and education comparisons. The hypothesis was confirmed in the education DIF analysis; however, in the race/ethnicity comparisons Hispanics and non-Hispanic Black respondents were less likely to endorse the item in the cognitive difficulties direction compared to the non-Hispanic White respondents. Hypothesized DIF for the items, trouble forming thoughts (higher education higher impair-

ment) and had to work hard to pay attention (lower education higher impairment) were not confirmed. In the language analysis, no DIF was found even though it was hypothesized that conditional on cognition, non-English and Spanish speakers would report higher impairment in having to work hard to pay attention.

Limitations

Two possible limitations relate to the lack of ability to distinguish among Hispanic and Asian/Pacific Islander ethnic groups, and the local dependencies observed. Two items that might be singled out for further study include one of each item pair showing elevated local dependence: forming thoughts and shifting back and forth. Both evidenced slightly lower metrics relating to information and discriminatory power respectively, as compared to their paired locally dependent items. Overall, the forming thoughts item evidenced the lowest information, and the shifting item was found to have DIF in age group comparisons. The item, brain not working as well as usual might also be a candidate for further study because this item was the most problematic in terms of DIF.

Conclusions

In general, the psychometric properties of the PROMIS Applied Cognition-General Concerns scale, version 1 of the Cognitive Function item bank, were good to excellent in terms of reliability, information and measurement equivalence across groups. Although DIF was observed in several items, the magnitude was low, and the impact of DIF on the scale was trivial. Future work is needed examining this measure across different populations.

Acknowledgements

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

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., … Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dementia. 7*(3), 270-279. doi: 10.1016/j.jalz.2011.03.008.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders*. (5th ed.). Washington, DC: Author.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397-438. doi: 10.1016/j.jrp.2014.07.001.

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.

Barnes, J., Whitwell, J. L., Frost, C., Josephs, K. A., Rossor, M., & Fox, N. C. (2006). Measurements of the amygdala and hippocampus in pathologically confirmed Alzheimer disease and frontotemporal lobar degeneration. *Archives of Neurology, 63*, 1434–1439. doi:10.1001/archneur.63.10.1434.

Becker, H., Stuifbergen, A., & Morrison, J. (2012). Promising new approaches to assess cognitive functioning in people with multiple sclerosis. *International Journal of MS Care*, *14*(2), 71-76. doi: 10.7224/1537-2073-14.2.71

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

Benedict, R. (1997). *Brief Visuospatial Memory Test – Revised*. Odessa, Fl.: Psychological Assessment Resources, Inc.

Benton, A. L., Sivan, A. B., Hamsher, K. D., Varney, N. R., & Spreen, O. (1994*). Contributions to Neuropsychological Assessment. A Clinical Manual*. (2nd ed.). New York, NY: Oxford University Press. doi: 10.1136/jnnp.59.3.346

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

Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., … PROMIS Cooperative Group (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress on an NIH-Roadmap cooperative group during its first two years. *Medical Care*, *45*, S3–11. doi: 10.1097/01.mlr.0000258615.42478.55

Centers for Disease Control and Prevention. (2013). Self-reported increased confusion or memory loss and associated functional difficulties among adults aged ≥ 60 years - 21 states, 2011. *Morbidity and Mortality Weekly Report (MMWR)*, *62*, 345-350.

Chen, W. H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265-289. doi: 10.3102/10769986022003265

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*, 1-30. doi: 10.18637/jss.v039.i08

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

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229-238.

Coman, L., & Richardson, J. (2006). Relationship between self-report and performance measures of function: A systematic review. *Canadian Journal on Aging*, *25*(3), 253-270.

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

Cox, D. R. & Snell, E. J. (1989). *The analysis of binary data* (2nd Ed.). London: Chapman and Hall.

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*, S115-S123. doi: 10.1097/01.mlr.0000245183.28384.ed

Crane, P. K., Gibbons, L. E., Jolley, L., van Belle, G., Selleri, R., Dalmonte, E., & De Ronchi, D. (2006). Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. *International Psychogeriatrics*, *18*, 505–515. doi: 10.1017/S1041610205002978

Crane, P. K., Gibbons, L. E., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., … 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. doi: 10.1007/s11136-007-9185-5

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334. doi: 10.1007/BF02310555

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302. http://dx.doi.org/10.1037/h0040957

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning Test* (2nd ed). San Antonio, TX: Psychological Corporation.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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

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

Gagnon, M., Dartigues, J. F., Mazaux, J. M., Dequae, L., Letenneur, L., Giroire, J. M., & Barberger-Gateau, P. (1994). Self-reported memory complaints and memory performance in elderly French community residents: Results of the PAQUID research program. *Neuroepidemiology*, *13*(4),145–154. doi: 10.1159/000110373

Gibbons, L. E., McCurry, S., Rhoads, K., Maski, K., White, L., Borenstein, A. R., … Crane, P. K. (2009). Japanese–English language equivalence of the cognitive abilities screening instrument among Japanese-Americans. *International Psychogeriatrics, 21*, 129–137. doi: 10.1017/S1041610208007862

Gronwall, D. M. A. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills, 44*, 367-373. doi: 10.2466/pms.1977.44.2.367

Gurland, B. J., Wilder, D.E., Lantigua, R., Stern, Y., Chen, J., Killeffer, E. H., & Mayeux, R. (1999). Rates of dementia in three ethnoracial groups. *International Journal of Geriatric Psychiatry*, *14*, 481–493. doi 10.1002/(SICI)1099-1166(199906)14:63.0.CO;2-5

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

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

Hulin, C. L. (1987). A psychometric theory of evaluations of item scale translations. *Journal of Cross-Cultural Psychology*, *18*, 115–142. doi:10.1177/0022002187018002001

Ingles, J. L., Fisk, J. D., Merry, H. R., & Rockwood, K. (2003). Five-year outcomes for dementia defined solely by neuropsychological test performance. *Neuroepidemiology*, *22*,172–178. doi:10.1159/000069891

Jensen, R. E., Moinpour, C. M., Keegan, T. H. M., Cress, R.D., Wu, X.-C., Paddock, L. A., … Potosky, A. L. (2016). The Measuring Your Health Study: Leveraging community-based cancer registry recruitment to establish a large, diverse cohort of cancer survivors for analyses of measurement equivalence and validity of the Patient Reported Outcomes Measurement Information System® (PROMIS®) short form items. *Psychological Test and Assessment Modeling, 58*, 99-117.

Jessen, F., Amariglio, R. E., van Boxtel, M., Breteler, M., Ceccaldi, M., Chételat, G., ...Subjective Cognitive Decline Initiative (SCD-I) Working Group. (2014). A conceptual framework for research on subjective cognitive decline in preclinical Alzheimer's disease. *Alzheimer's and Dementia, 10*, 844-852. doi: 10.1016/j.jalz.2014.01.001

Johnco, C., Wuthrich, V. M., & Rapee, R. M. (2014). The influence of cognitive flexibility on treatment outcome and cognitive restructuring skill acquisition during cognitive behavioural treatment for anxiety and depression in older adults: Results of a pilot study. *Behaviour Research and Therapy, 57*, 55-64. doi: 10.1016/j.brat.2014.04.005

Joly, F., Alibhai, S. M., Galica, J., Park, A., Yi, Q. L., Wagner, L., & Tannock, I. F. (2006). Impact of androgen deprivation therapy on physical and cognitive function, as well as quality of life of patients with nonmetastatic prostate cancer. *Journal of Urology, 176*, 2443-2447. doi: http://dx.doi.org/10.1016/j.juro.2006.07.151

Kaufman, J. C. (2006). Self-reported differences in creativity by gender and ethnicity. *Journal of Applied Cognitive Psychology*, *20*, 1065-108. doi: 10.1002/acp.1255

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

Kleinman, M., & Teresi, J. A. (2016). Differential item functioning magnitude and impact measures from item response theory models. *Psychological Test and Assessment Modeling, 58*, 79-98.

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

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J.: L. Erlbaum Associates.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). New York: Academic Press.

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

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

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager T. (2000). The unity and diversity of executive functions and their contributions to complex

"frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*, 49–100. doi: 10.1006/cogp.1999.0734

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

Muthén, B. (1982). Some categorical response models with continuous latent variables. In K. G. Joreskog, & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction (the conference volume)*. Amsterdam: North Holland.

Muthén, L. K., & Muthén, B. O. (1998-2011). *M-PLUS Users Guide*. (6th ed.). Los Angeles, California: Muthén and Muthén.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*, 691-692. doi:10.1093/biomet/78.3.691

Orlando-Edelen, M., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Applications to the Mini-Mental State Examination. *Medical Care, 44*(11 Suppl 3), S134-S142. doi: 10.1097/01.mlr.0000245251.83359.8c

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

Potter, G. G., Plassman, B.L., Burke, J. R., Kabeto, M. U., Langa, K. M., Llewellyn, D. J. … Steffens, D. C. (2009). Cognitive performance and informant reports in the diagnosis of cognitive impairment and dementia in African Americans and whites. *Alzheimer's & Dementia, 5*, 445–453. doi: 10.1016/j.jalz.2009.04.1234

Primi, R., Rocha da Silva, M. C., Rodrigues, P., Muniz, M., & Almeida, L. S. (2013). The use of the bi-factor model to test the uni-dimensionality of a battery of reasoning tests. *Psicothema*, *25*(1), 115-122. doi: 10.7334/psicothema2011.393

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

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

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

Reisberg, B., Shulman, M. B, Torossian, C., Leng, L., & Zhu, W. (2010). Outcome over seven years of healthy adults with and without subjective cognitive impairment. *Alzheimer's & Dementia, 6*(1), 11–24. doi: 10.1016/j.jalz.2009.10.002

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

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16* (Suppl 1), 19-31. doi: 10.1007/s11136-007-9183-7

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. doi: 10.1146/annurev.clinpsy.032408.153553

Reuben, D. B., Seeman, T. E., Keeler, E., Hayes, R. P., Bowman, L., Sewall, A., … Guralnik, J. M. (2004). Refining the categorization of physical functional status: The added value of combining self-reported and performance-based measures. *The Journal of Gerontology Series A Biological Sciences and Medical Sciences*, *59*, M1056-M1061. doi:10.1093/gerona/59.10.M1056

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

Saffer, B. Y., Lanting, S. C., Koehle, M. S., Klonsky, E. D., & Iverson, G. L. (2015). Assessing cognitive impairment using PROMIS(®) applied cognition-abilities scales in a medical outpatient sample. *Psychiatry Research, 226*, 169-72. doi: 10.1016/j.psychres. 2014.12.043

Salthouse, T. A. (1996). The processing speed theory of adult age differences in cognition. *Psychological Review, 103*(3), 403–428. doi:10.1037/0033-295X.103.3.403

Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General, 132,* 566–594. doi: http://dx.doi.org/10.1037/0096-3445.132.4.566

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

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

Seybert, J., & Stark, S. (2012). Iterative linking with the Differential Functioning of Items and Tests (DFIT) method: Comparison of testwide and item parameter replication (IPR) critical values. *Applied Psychological Measurement*, *36*, 494-515. doi: 10.1177/0146 621612445182

Shallice, T., Burgess, P., & Robertson, I. (1996). The domain of supervisory processes and temporal organisation of behaviour. *Philosophical Transactions of the Royal Society B*, *351*(1346), 1405–1412. doi: 10.1098/rstb.1996.0124

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

Shilling, V., Jenkins, V., Morris, R., Deutsch, G., & Bloomfield, D. (2005). The effects of adjuvant chemotherapy on cognition in women with breast cancer--preliminary results of an observational longitudinal study. *Breast*, *14*(2), 142–150. doi: 10.1016/j.breast.2004. 10.004.

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

Smith, A. (1982). *Symbol digit modalities test: Manual.* Los Angeles: Western Psychological Services.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M…Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia, 7*(3), 280-292. doi: 10.1016/j.jalz.2011.03.003

Teresi, J. A. & Jones, R. N. (2016). Methodological issues in examining measurement equivalence in Patient Reported Outcomes Measures: Methods overview to the two-part series, "Measurement Equivalence of the Patient Reported Outcomes Measurement

Information System (PROMIS) Short Form Measures." *Psychological Test and Assessment Modeling*, *58*(1), 37-78.

Teresi, J. A., Kleinman, M., Ocepek-Welikson, K., Ramirez, M., Gurland, B., Lantigua, R., & Holmes, D. (2000). Applications of item response theory to the examination of the psychometric properties and differential item functioning of the CARE Dementia Diagnostic Scale among samples of Latino, African-American and White non-Latino elderly. *Research on Aging, 22*, 738-773. doi:10.1177/0164027500226007

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

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

Warnecke, R. B., Johnson, T. P., Chavez, N., Sudman, S., O'Rourke, D. P., Lacey, L., & Horm, J. (1997). Improving question wording in surveys of culturally diverse populations. *Annals of Epidemiology*, *7*, 334–42. doi: 10.1016/S1047-2797(97)00030-6

Wolf, H., Grunwald, M., Ecke, G. M., Zedlick, D., Bettin, S., Dannenberg, C., … Gertz, H. J. (1998). The prognosis of mild cognitive impairment in the elderly. *Journal of Neural Transmission Supplements*, *54*, 31–50. doi: 10.1007/978-3-7091-7508-8_4

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

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). Effects of local item dependence on the validity of IRT item, test, and ability statistics [MCAT Monograph no. 5].

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficient alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods, 6*, 21-29. Available at: http://digitalcommons.wayne.edu/jmasm/vol6/iss1/4

Appendix

Table A1:
PROMIS applied cognition - general concerns item set. Model fit statistics: Comparative fit index (CFI) from the confirmatory factor and bi-factor models and the graded response model fit from IRTPRO for the total sample and the demographic subgroups

Sample	CFA CFI (MPLUS)	IRT Model RMSEA (IRTPRO)
Total Sample (CFA)	0.995	0.05
Random First Half Sample (CFA)	0.996	N/A
Random Second Half Sample (Bi-factor CFA)	0.999	N/A
Female	0.995	0.06
Male	0.996	0.05
Age 21 to 49 Years	0.996	0.06
Age 50 to 64 Years	0.996	0.06
Age 65 to 84 Years	0.994	0.05
Non-Hispanic Whites	0.997	0.05
Non-Hispanic Blacks	0.995	0.06
Hispanics	0.995	0.07
Non-Hispanic Asians/Pacific Islanders	0.994	0.10
Less Than High School	0.994	0.08
High School Graduate	0.995	0.06
Some College	0.997	0.05
College Graduate	0.996	0.07
Graduate Degree	0.996	0.06
Hispanics Interviewed in English	0.996	0.09
Hispanics Interviewed in Spanish	0.994	0.09

Table A2:
PROMIS applied cognition - general concerns item set: Discrimination '*a*' parameter estimates for IRTPRO individual subgroup runs

Item description	Total Sample	Race/Ethnicity				Education					Age Groups			Gender		Hispanics Only Interview Language	
		NH White	NH Black	Hisp.	NH API	No HS	HS	Some Coll.	Coll.	Grad.	21-49	50-64	65-84	Fe-male	Male	Engl.	Span.
I have had trouble forming thoughts	4.35	4.52	4.42	4.32	4.00	4.13	4.31	4.56	4.04	4.51	4.45	4.37	4.14	4.22	4.64	4.72	3.45
My thinking has been slow	4.96	5.30	4.84	4.79	4.64	4.69	4.58	5.36	5.07	5.05	5.27	4.93	4.65	4.87	5.13	5.24	3.91
My thinking has been foggy	5.50	5.86	5.15	5.24	6.09	5.19	4.88	5.80	6.06	5.56	5.80	5.29	5.35	5.29	5.94	5.75	4.36
I have had trouble concentrating	5.65	5.71	5.69	5.57	5.74	5.34	5.46	5.87	5.30	6.10	5.53	5.91	5.27	5.36	6.25	5.86	4.83
I have had to work really hard to pay attention or I would make a mistake	5.63	5.99	5.75	5.14	5.38	5.15	5.89	5.83	5.43	5.45	5.92	5.70	5.22	5.62	5.72	5.14	4.90
It has seemed like my brain was not working as well as usual	6.01	6.63	6.00	5.74	5.81	5.37	6.47	6.61	6.11	6.65	6.29	6.01	5.58	6.04	6.00	6.16	5.01
I have had to work harder than usual to keep track of what I was doing	6.26	7.35	6.32	5.54	5.32	6.12	6.77	6.10	5.94	6.87	6.42	7.02	5.34	6.11	6.54	5.40	5.79
I have had trouble shifting back and forth between different activities that require thinking	5.90	6.12	6.31	6.04	5.04	6.23	5.89	6.03	5.18	5.82	5.48	6.53	5.49	5.71	6.35	5.71	7.28

Table A3:
PROMIS applied cognition - general concerns item set: Final IRT item parameters and DIF statistics for the race/ethnicity groups, non-Hispanic Whites are the reference group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had trouble forming thoughts	Non-Hispanic White	5.08 (0.14)	0.04 (0.02)	0.61 (0.01)	1.13 (0.02)	1.55 (0.03)	NS, Anchor item	
	Non-Hispanic Black							
	Hispanic							
	Non-Hispanic Asian/Pacific Islander							
My thinking has been slow	Non-Hispanic White	5.79 (0.16)	-0.07 (0.02)	0.48 (0.01)	1.05 (0.02)	1.45 (0.03)	NS, Anchor item	
	Non-Hispanic Black							
	Hispanic							
	Non-Hispanic Asian/Pacific Islander							
My thinking has been foggy	Non-Hispanic White	6.45 (0.26)	0.04 (0.02)	0.59 (0.02)	1.15 (0.03)	1.60 (0.04)		
	Non-Hispanic Black	5.85 (0.30)	0.00 (0.02)	0.56 (0.02)	1.09 (0.03)	1.57 (0.05)	2.2 (0.136)	6.3 (0.177)
	Hispanic	6.43 (0.35)	0.12 (0.02)	0.60 (0.02)	1.11 (0.03)	1.51 (0.05)	<0.1 (0.923)	13.7 (0.008)
	Non-Hispanic Asian/Pacific Islander	7.98 (0.54)	0.08 (0.02)	0.62 (0.02)	1.19 (0.04)	1.53 (0.05)	6.3 (0.012)	4.4 (0.351)
I have had trouble concentrating	Non-Hispanic White	6.25 (0.24)	-0.07 (0.02)	0.50 (0.02)	1.07 (0.03)	1.49 (0.04)		
	Non-Hispanic Black	6.43 (0.34)	-0.08 (0.02)	0.50 (0.02)	1.05 (0.03)	1.42 (0.04)	0.2 (0.665)	1.7 (0.793)
	Hispanic	6.86 (0.38)	-0.08 (0.02)	0.46 (0.02)	1.03 (0.03)	1.44 (0.04)	1.9 (0.171)	2.0 (0.740)
	Non-Hispanic Asian/Pacific Islander	7.51 (0.48)	-0.03 (0.02)	0.54 (0.02)	1.15 (0.04)	1.51 (0.05)	5.5 (0.019)	8.0 (0.090)

continued

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had to work really hard to pay attention or I would make a mistake	Non-Hispanic White	6.54 (0.19)	0.02 (0.01)	0.58 (0.01)	1.09 (0.02)	1.53 (0.03)	NS, Anchor item	
	Non-Hispanic Black							
	Hispanic							
	Non-Hispanic Asian/Pacific Islander							
It has seemed like my brain was not working as well as usual	Non-Hispanic White	7.24 (0.30)	-0.07 (0.02)	0.44 (0.02)	0.94 (0.02)	1.38 (0.03)		
	Non-Hispanic Black	6.83 (0.38)	0.02 (0.02)	0.53 (0.02)	1.02 (0.03)	1.45 (0.04)	0.7 (0.392)	***20.9 (<0.001)***
	Hispanic	7.37 (0.20)	0.05 (0.02)	0.50 (0.02)	1.04 (0.03)	1.43 (0.05)	<0.1 (0.883)	***31.3 (<0.001)***
	Non-Hispanic Asian/Pacific Islander	7.37 (0.20)	-0.02 (0.02)	0.51 (0.03)	1.01 (0.04)	1.39 (0.05)	0.6 (0.442)	10.0 (0.040)
I have had to work harder than usual to keep track of what I was doing	Non-Hispanic White	7.35 (0.22)	0.02 (0.01)	0.52 (0.01)	1.00 (0.02)	1.38 (0.02)	NS, Anchor item	
	Non-Hispanic Black							
	Hispanic							
	Non-Hispanic Asian/Pacific Islander							
I have had trouble shifting back and forth between different activities that require thinking	Non-Hispanic White	6.64 (0.27)	0.07 (0.02)	0.61 (0.02)	1.06 (0.02)	1.48 (0.03)		
	Non-Hispanic Black	7.11 (0.41)	0.06 (0.02)	0.56 (0.02)	1.08 (0.03)	1.38 (0.04)	0.9 (0.335)	8.5 (0.074)
	Hispanic	7.14 (0.20)	0.03 (0.02)	0.55 (0.02)	1.01 (0.03)	1.43 (0.05)	3.5 (0.062)	6.1 (0.192)
	Non-Hispanic Asian/Pacific Islander	7.14 (0.20)	-0.00 (0.02)	0.56 (0.03)	1.05 (0.04)	1.36 (0.05)	<0.1 (0.972)	13.1 (0.011)

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters.

* Bolded entries indicate items that evidence DIF after correction for multiple comparisons; "NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A4:
PROMIS applied cognition - general concerns item set: Final IRT item parameters and DIF statistics for the education groups, respondents with graduate degrees are the reference group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had trouble forming thoughts	Less than HS	5.07 (0.26)	-0.19 (0.02)	0.33 (0.02)	0.86 (0.03)	1.28 (0.05)	0.8 (0.364)	6.1 (0.189)
	High School	5.52 (0.28)	-0.22 (0.02)	0.34 (0.02)	0.81 (0.03)	1.15 (0.04)	<.01 (0.875)	11.1 (0.026)
	Some College	5.48 (0.22)	-0.20 (0.02)	0.36 (0.02)	0.84 (0.03)	1.25 (0.04)	<.01 (0.919)	5.9 (0.211)
	College Degree	5.07 (0.28)	-0.14 (0.02)	0.38 (0.03)	0.92 (0.04)	1.40 (0.07)	0.6 (0.434)	2.3 (0.687)
	Graduate Degree	5.46 (0.38)	-0.16 (0.03)	0.40 (0.03)	0.94 (0.05)	1.33 (0.08)		
My thinking has been slow	Less than HS	6.11 (0.16)	-0.29 (0.01)	0.24 (0.01)	0.78 (0.02)	1.17 (0.02)	DIF not significant	
	High School							
	Some College							
	College Degree							
	Graduate Degree							
My thinking has been foggy	Less than HS	6.74 (0.18)	-0.17 (0.01)	0.34 (0.01)	0.86 (0.02)	1.26 (0.02)	NS, Anchor item	
	High School							
	Some College							
	College Degree							
	Graduate Degree							
I have had trouble concentrating	Less than HS	6.88 (0.19)	-0.29 (0.01)	0.25 (0.01)	0.80 (0.02)	1.17 (0.02)	NS, Anchor item	
	High School							
	Some College							
	College Degree							
	Graduate Degree							

continued

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had to work really hard to pay attention or I would make a mistake	Less than HS	6.17 (0.33)	-0.20 (0.02)	0.30 (0.02)	0.85 (0.03)	1.26 (0.04)	0.4 (0.545)	6.5 (0.168)
	High School	7.30 (0.40)	-0.21 (0.02)	0.34 (0.02)	0.79 (0.03)	1.16 (0.04)	1.7 (0.189)	9.6 (0.047)
	Some College	7.10 (0.31)	-0.21 (0.02)	0.32 (0.02)	0.83 (0.02)	1.23 (0.03)	0.8 (0.364)	8.1 (0.088)
	College Degree	7.05 (0.42)	-0.22 (0.02)	0.34 (0.02)	0.76 (0.03)	1.23 (0.05)	0.5 (0.460)	9.3 (0.054)
	Graduate Degree	6.56 (0.48)	-0.15 (0.03)	0.37 (0.03)	0.87 (0.04)	1.35 (0.08)		
It has seemed like my brain was not working as well as usual	Less than HS	6.42 (0.34)	-0.15 (0.02)	0.30 (0.02)	0.79 (0.03)	1.21 (0.04)	5.4 (0.020)	***17.9 (0.001)***
	High School	8.03 (0.45)	-0.24 (0.02)	0.27 (0.02)	0.70 (0.03)	1.13 (0.04)	<.01 (0.978)	8.3 (0.082)
	Some College	8.09 (0.37)	-0.26 (0.02)	0.23 (0.02)	0.72 (0.02)	1.08 (0.03)	<.01 (0.931)	1.7 (0.794)
	College Degree	7.76 (0.47)	-0.30 (0.02)	0.18 (0.02)	0.67 (0.03)	1.11 (0.04)	0.2 (0.658)	1.9 (0.761)
	Graduate Degree	8.13 (0.63)	-0.27 (0.02)	0.19 (0.03)	0.70 (0.03)	1.06 (0.05)		
I have had to work harder than usual to keep track of what I was doing	Less than HS	7.66 (0.21)	-0.21 (0.01)	0.27 (0.01)	0.73 (0.02)	1.10 (0.02)	NS, Anchor item	
	High School							
	Some College							
	College Degree							
	Graduate Degree							
I have had trouble shifting back and forth between different activities that require thinking	Less than HS	7.21 (0.19)	-0.18 (0.01)	0.33 (0.01)	0.78 (0.02)	1.14 (0.02)	NS, Anchor item	
	High School							
	Some College							
	College Degree							
	Graduate Degree							

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters.

* Bolded entries indicate items that evidence DIF after correction for multiple comparisons; "NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A5:
PROMIS applied cognition - general concerns item set: Final IRT item parameters and DIF statistics for the age groups, with the age 21 to 49 group as the reference

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had trouble forming thoughts	Age 21-49	4.98 (0.15)	-0.23 (0.02)	0.34 (0.01)	0. 88 (0.02)	1.31 (0.03)	NS, Anchor item	
	Age 50-64							
	Age 65-84							
My thinking has been slow	Age 21-49	5.66 (0.18)	-0.35 (0.02)	0.22 (0.01)	0.79 (0.02)	1.21 (0.03)	NS, Anchor item	
	Age 50-64							
	Age 65-84							
My thinking has been foggy	Age 21-49	6.29 (0.20)	-0.22 0.02)	0.33 (0.01)	0.89 (0.02)	1.31 (0.03)	NS, Anchor item	
	Age 50-64							
	Age 65-84							
I have had trouble concentrating	Age 21-49	6.46 (0.21)	-0.34 (0.02)	0.23 (0.01)	0.82 (0.02)	1.22 (0.03)	DIF not significant	
	Age 50-64							
	Age 65-84							
I have had to work really hard to pay attention or I would make a mistake	Age 21-49	6.70 (0.35)	-0.18 (0.02)	0.34 (0.02)	0.82 (0.03)	1.35 (0.04)		
	Age 50-64	6.57 (0.30)	-0.24 0.02)	0.30 (0.02)	0.82 (0.02)	1.24 (0.03)	< 0.1 (0.893)	10.9 (0.028)
	Age 65-84	6.24 (0.28)	-0.30 (0.02)	0.32 (0.02)	0.88 (0.03)	1.26 (0.04)	0.5 (0.471)	***19.5 (0.001)***
It has seemed like my brain was not working as well as usual	Age 21-49	7.12 (3.9)	-0.26 (0.03)	0.19 (0.02)	0.75 (0.02)	1.20 (0.04)		
	Age 50-64	6.98 (0.32)	-0.29 (0.02)	0.23 (0.02)	0.72 (0.02)	1.12 (0.03)	<0.1 (0.971)	12.7 (0.013)
	Age 65-84	6.67 (0.30)	-0.32 (0.02)	0.23 (0.02)	0.75 (0.02)	1.15 (0.04)	0.04 (0.522)	12.8 (0.012)

continued

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had to work harder than usual to keep track of what I was doing	Age 21-49	7.34 (0.40)	-0.20 (0.02)	0.26 (0.02)	0.72 (0.02)	1.17 (0.03)		
	Age 50-64	8.15 (0.42)	-0.25 (0.02)	0.26 (0.02)	0.74 (0.02)	1.08 (0.03)	3.1 (0.077)	11.9 (0.018)
	Age 65-84	6.38 (0.28)	-0.30 (0.02)	0.26 (0.02)	0.77 (0.03)	1.17 (0.04)	2.7 (0.099)	8.2 (0.084)
I have had trouble shifting back and forth between different activities that require thinking	Age 21-49	6.23 (0.32)	-0.15 (0.02)	0.35 (0.02)	0.83 (0.03)	1.25 (0.04)		
	Age 50-64	7.60 (0.37)	-0.22 (0.02)	0.31 (0.02)	0.77 (0.02)	1.12 (0.03)	***10.4 (0.001)***	***16.4 (0.003)***
	Age 65-84	6.56 (0.30)	-0.28 (0.02)	0.30 (0.02)	0.81 (0.03)	1.18 (0.04)	1.2 (0.269)	***18.7 (0.001)***

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters."NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A6:
PROMIS applied cognition - general concerns item set: Final IRT item parameters and DIF statistics for gender, with females as the reference group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had trouble forming thoughts	Female	4.25	-0.04	0.63	1.26	1.77	DIF not significant	
	Male	(0.13)	(0.04)	(0.04)	(0.05)	(0.06)		
My thinking has been slow	Female	4.82	-0.17	0.49	1.16	1.66	NS, Anchor item	
	Male	(0.15)	(0.04)	(0.04)	(0.05)	(0.06)		
My thinking has been foggy	Female	5.36	-0.02	0.62	1.27	1.77	NS, Anchor item	
	Male	(0.17)	(0.04)	(0.04)	(0.05)	(0.06)		
I have had trouble concentrating	Female	5.50	-0.17	0.50	1.19	1.66	NS, Anchor item	
	Male	(0.18)	(0.04)	(0.04)	(0.05)	(0.06)		
I have had to work really hard to pay attention or I would make a mistake	Female	5.48	-0.06	0.60	1.12	1.73	DIF not significant	
	Male	(0.18)	(0.04)	(0.04)	(0.05)	(0.06)		
It has seemed like my brain was not working as well as usual	Female	5.85	-0.11	0.49	1.10	1.59	NS, Anchor item	
	Male	(0.19)	(0.04)	(0.04)	(0.05)	(0.06)		
I have had to work harder than usual to keep track of what I was doing	Female	6.08	-0.07	0.54	1.10	1.57	NS, Anchor item	
	Male	(0.20)	(0.04)	(0.04)	(0.05)	(0.06)		
I have had trouble shifting back and forth between different activities that require thinking	Female	5.73	-0.04	0.60	1.17	1.61	DIF not significant	
	Male	(0.19)	(0.04)	(0.04)	(0.05)	(0.06)		

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters."NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Table A7:
PROMIS applied cognition - general concerns item set: Final IRT item parameters and DIF statistics for the language groups for Hispanics only (English & Spanish), with English interview as the reference group

Item name	Group	*a*	*b1*	*b2*	*b3*	*b4*	*a*DIF*	*b*DIF*
I have had trouble forming thoughts	English	4.51	-0.05	0.61	1.20	1.66	NS, Anchor item	
	Spanish	(0.28)	(0.05)	(0.04)	(0.05)	(0.07)		
My thinking has been slow	English	5.02	-0.18	0.49	1.14	1.60	NS, Anchor item	
	Spanish	(0.32)	(0.05)	(0.04)	(0.05)	(0.07)		
My thinking has been foggy	English	5.72	-0.02	0.57	1.19	1.63	0.3 (0.577)	10.8 (0.029)
		(0.42)	(0.05)	(0.04)	(0.06)	(0.07)		
	Spanish	5.32	0.15	0.68	1.25	1.85		
		(0.53)	(0.05)	(0.05)	(0.08)	(0.13)		
I have had trouble concentrating	English	5.84	-0.20	0.43	1.12	1.60	NS, Anchor item	
	Spanish	(0.38)	(0.05)	(0.04)	(0.05)	(0.07)		
I have had to work really hard to pay attention or I would make a mistake	English	5.37	-0.08	0.54	1.18	1.73	NS, Anchor item	
	Spanish	(0.34)	(0.05)	(0.04)	(0.05)	(0.08)		
It has seemed like my brain was not working as well as usual	English	6.02	-0.04	0.50	1.13	1.59	DIF not significant	
	Spanish	(0.39)	(0.05)	(0.04)	(0.05)	(0.07)		
I have had to work harder than usual to keep track of what I was doing	English	5.80	-0.09	0.53	1.05	1.54	NS, Anchor item	
	Spanish	(0.38)	(0.05)	(0.04)	(0.05)	(0.07)		
I have had trouble shifting back and forth between different activities that require thinking	English	5.81	-0.03	0.52	1.09	1.61	5.8 (0.016)	9.0 (0.062)
		(0.42)	(0.05)	(0.04)	(0.05)	(0.07)		
	Spanish	8.31	-0.12	0.59	1.06	1.48		
		(0.97)	(0.05)	(0.05)	(0.06)	(0.09)		

* Statistical test for differences in parameters is Wald test using 1 *df* for the test of differences in the *a* parameters for the comparison groups and 2 *df* for the test of differences in the *b* parameters."NS, Anchor item" refers to a non-significant DIF finding for the item during the initial iterative anchor item selection process. The "non-significant" designation refers to the second stage DIF detection procedure using the anchor items and testing the remaining items. The "non-significant" designation indicates that the item was not found to have DIF in the second stage of DIF detection.

Figure A1:
PROMIS applied cognition - general concerns item set: Scree plot from exploratory factor analysis of the total sample (n = 5477)

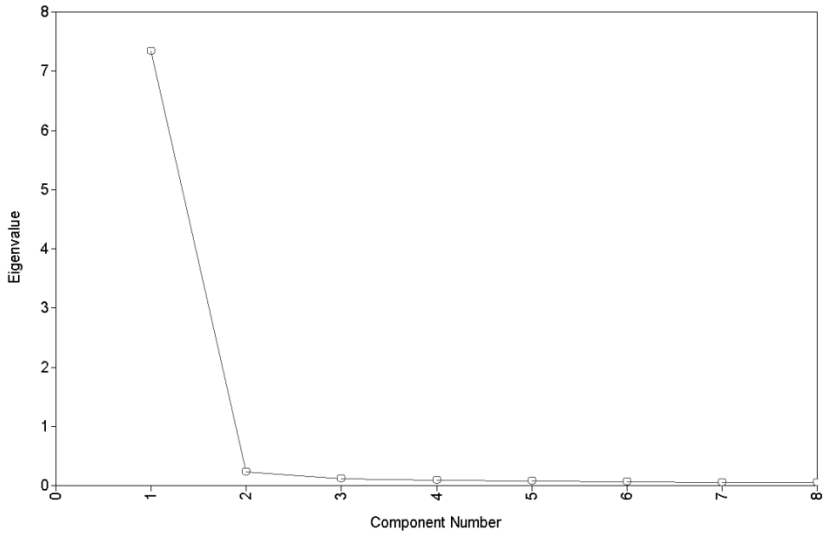


Figure A2:
PROMIS applied cognition - general concerns item set: Item information functions of the total sample

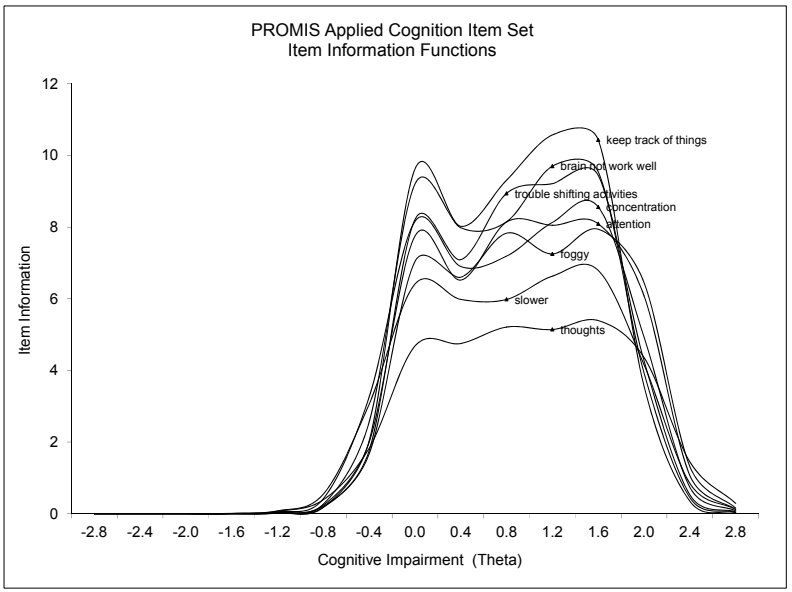
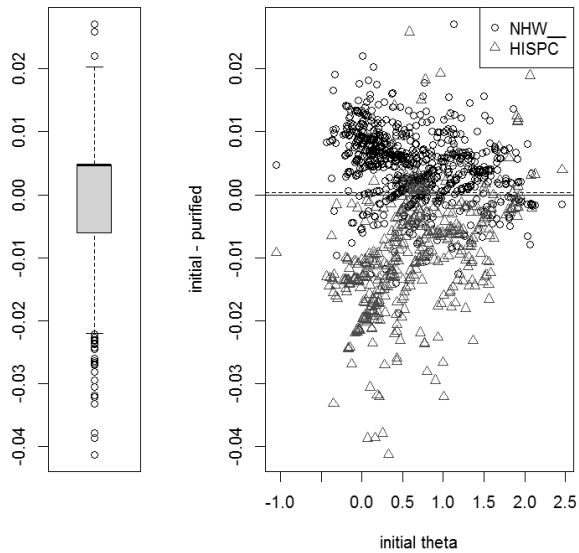
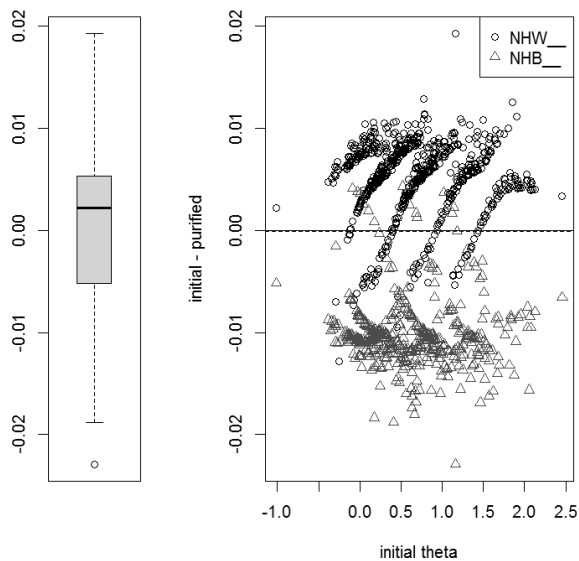


Figure A3:
Individual impact analyses graphs depicting individual-level differential item functioning impact (from lordif)

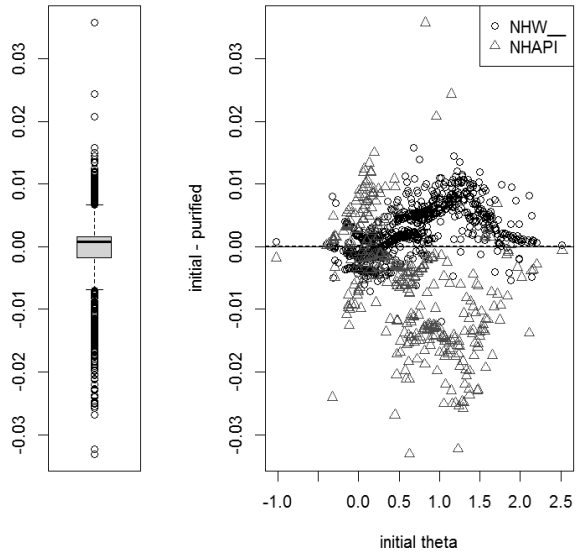
Non-Hispanic White (reference group) vs. Hispanic



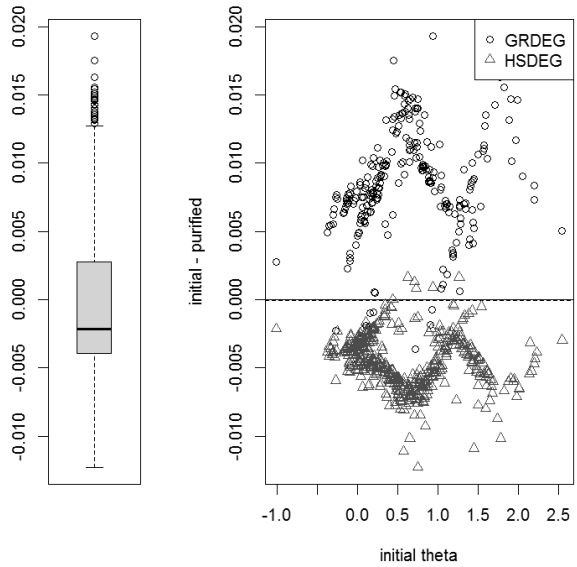
Non-Hispanic White (reference group) vs. non-Hispanic Black



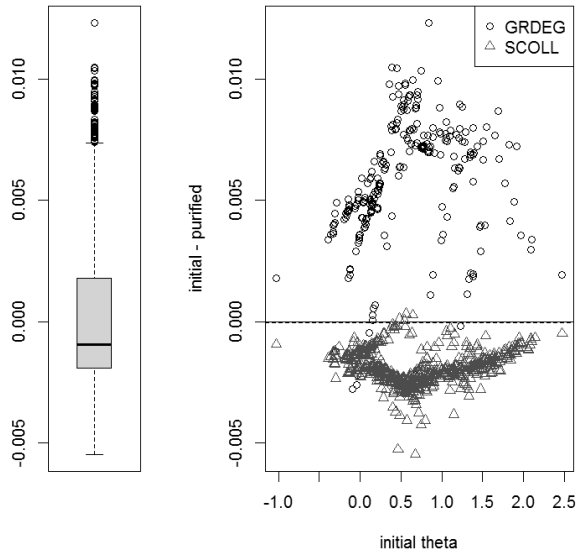
Non-Hispanic White (reference group) vs. NHAPI



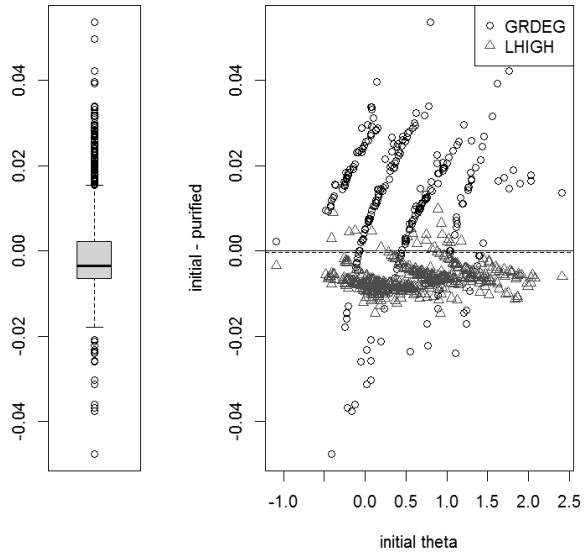
Graduate school (reference group) vs. high school degree



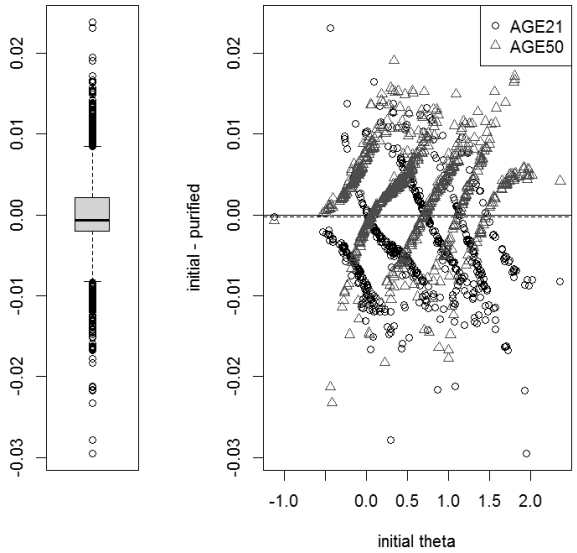
Graduate degree (reference group) vs. some college



Graduate degree (reference group) vs. less than high school



Age 21 to 49 years (reference group) vs. age 50 to 64 years



Age 21 to 49 years (reference group) vs. age 65 to 84 years