Psychological Test and Assessment Modeling, Volume 53, 2011 (4), 440-460

A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos

Frances M. Yang[1,2,3], Kevin C. Heslin[4], Kala M. Mehta[5], Cheng-Wu Yang[6], Katja Ocepek-Welikson[7], Marjorie Kleinman[1], Leo S. Morales[4], Ron D. Hays[4], Anita L. Stewart[5], Dan Mungas[8], Richard N. Jones[1,2,3], Jeanne A. Teresi[1,7,9]

Abstract

Object naming tests are commonly included in neuropsychological test batteries. *Differential item functioning* (DIF) in these tests due to cultural and language differences may compromise the validity of cognitive measures in diverse populations. We evaluated 26 object naming items for DIF due to Spanish and English language translations among Latinos (n=1,159), mean age of 70.5 years old (Standard Deviation (SD)±7.2), using the following four *item response theory*-based approaches: *Mplus/Multiple Indicator, Multiple Causes* (Mplus/MIMIC; Muthén & Muthén, 1998-2011), *Item Response Theory Likelihood Ratio Differential Item Functioning* (IRTLRDIF/ MULTILOG; Thissen, 1991, 2001), difwithpar/Parscale (Crane, Gibbons, Jolley, & van Belle, 2006; Muraki & Bock, 2003), and Differential Functioning of Items and Tests/MULTILOG (DFIT/MULTILOG; Flowers, Oshima, & Raju, 1999; Thissen, 1991). Overall, there was moderate to near perfect agreement across methods. Fourteen items were found to exhibit DIF and 5 items observed consistently across all methods, which were more likely to be answered correctly by individuals tested in Spanish after controlling for overall ability.

Key words: Item response theory, differential item functioning, object naming test, Hispanic/ Latinos, Spanish

---

Correspondence concerning this article should be addressed to: Frances M. Yang, PhD, Institute for Aging Research, Hebrew SeniorLife, Harvard Medical School, Department of Medicine Beth Israel Deaconess Medical Center, Division of Gerontology, 1200 Centre St., Boston, MA 02131, USA; email: francesyang@hsl.harvard.edu

[1]Institute for Aging Research, Hebrew SeniorLife; [2]Harvard Medical School, Beth Israel Deaconess Medical Center, Department of Medicine; [3]Columbia (CALME) RCMAR; [4]University of California, Los Angeles (UCLA) RCMAR; [5]University of California, San Francisco (UCSF) RCMAR; [6]Medical University of South Carolina (MUSC) RCMAR; [7]Hebrew Home at Riverdale; [8]University of California, Davis; [9]Columbia University Stroud Center, Faculty of Medicine; New York State Psychiatric Institute and Research Division

Introduction

Cognitive test items are increasingly being included in large-scale, government-funded item banks. Given the diversity of potential evaluees, it is important that such banks contain items that are conceptually and psychometrically equivalent among groups differing in characteristics such as education, ethnicity and language. Neuropsychological tests assess several domains of cognitive function, including memory, attention, conceptual thinking, verbal abilities, spatial abilities, and executive functioning. Tests of object naming, which measure ability to retrieve verbal information from semantic memory, are commonly included in neuropsychological test batteries. Several object naming tests are currently available (Druks, Masterson, Kopelman, Clare, Rose, & Rai, 2006; Króliczak, Westwood, & Goodale, 2006; Zec, Markwell, Burkett, & Larsen, Zec, Markwell, Burkett, & Larsen, 2005). Typically, the individual being assessed is shown a set of pictures and asked to name the objects represented. Used in conjunction with other cognitive tests, poor performance on object naming tasks may be an indicator of cognitive changes that could support a diagnosis of Alzheimer's disease or other dementia.

Measurement non-invariance due to language of test administration can occur when individuals of different language groups (e.g., Spanish and English speakers) at similar levels of cognitive functioning respond to cognitive test items differently. These differences can result in a type of bias called *differential item functioning* (DIF). This type of item bias may result in differences in test validity across groups. Two types of DIF, uniform and non-uniform can be detected. Uniform DIF occurs when the probability of response is in the same direction across the cognitive function continuum. Non-uniform DIF is evident when DIF is in different directions at different parts of the cognitive function distribution. Analytic methods for detecting DIF are potentially important and useful in evaluating the validity of cognitive functioning and other health outcome measures in diverse populations (see Teresi, Ramirez, Lai, & Silver, 2008).

A fundamentally important issue in assessing DIF is choosing among available methods and software programs. Five different methods were used recently to identify different items with DIF (Crane, Gibbons, Jolley, & van Belle 2006; Dorans & Kulick, 2006; Hays, Morales, & Reise, 2000; Jones, 2006; Orlando-Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006) on the Mini Mental State Examination (MMSE) (Folstein, Folstein, & McHugh, 1975). Findings were not identical across methods. The reasons for the differences in findings were posited to be due to the use of different criteria for identifying and flagging DIF, for example, measures of magnitude versus statistical significance (Borsboom, 2006; Hambleton, 2006; Millsap, 2006).

Comparing findings from different methods can provide insights into whether differences are due to the different assumptions and criteria embedded within the methods. Moreover, convergent findings across methods are more likely to prompt content experts to modify or remove items with consistent DIF of high magnitude. Using real data from the Spanish and English Neurological Assessment Scales (SENAS), this study evaluated four different techniques for detecting DIF. We hypothesized that these four Item-Response Theory (IRT)-based techniques would show substantial agreement in the detection of

DIF among the same set of object naming items, but vary in the number of items flagged with DIF due to different assumptions and criteria used.

Methods

Analytic sample

This study used data from the development of the Spanish and English Neuropsychological Assessment Scales (SENAS) (Mungas, Reed, Crane, Haan, & Gonzalez, 2004; Mungas, Reed, Haan, & Gonzalez, 2005; Mungas, Reed, Marshall, & Gonzalez, 2000). These data are uniquely suited to address questions regarding DIF in cognitive testing instruments because the SENAS scales were developed using IRT, a methodology which allows examination of how an individual's performance on a test is based upon their latent ability or trait (Hambleton, Swaminathan, & Rogers, 1991).

The sample was of community volunteers, recruited via community outreach methods targeting health care systems and community organizations. Details of the sampling plan have been presented elsewhere (Mungas, et al., 2004; Mungas, et al., 2005; Mungas, et al., 2000). A total of 1,779 participants, mean age of 70.5 years old (Standard Deviation (*SD*)±7.2) completed the SENAS object naming test, we used the 59 years, rather than 65 years old, because recruitment began as young as 59 years old in the SENAS. We excluded persons in the following race/ethnic groups: Native Americans (n=1), Asians (n=8), Blacks and African Americans (n=174), Filipinos (n=6), Whites (n=421), other (n=6), and those with missing data (n=2). Also excluded were persons who had missing responses for all items (n=2). The analytic sample consisted of 1,159 English- and Spanish-speaking Latinos (65% of total participants).

Measures

The SENAS battery consists of scales measuring core cognitive abilities, which are described in detail elsewhere (Mungas, et al., 2004; Mungas, et al., 2005; Mungas, et al., 2000). The current study focused on the object naming ability. The SENAS object naming scale is a verbal measure of semantic memory consisting of 44 items scored as zero if incorrect and one if correct. The names of the included items were selected to have similar frequency of usage in the Spanish and English languages, based upon frequency norms of Eaton (2003).

To collect data with the object naming scale, trained examiners placed a stimulus page in front of the participant and asked the following in either English or Spanish, based upon the participant's language preference: "I am going to ask you to name some objects" or "Voy a pedirle que nombre algunos objetos." Then the examiner pointed to the object corresponding to the item and said: "Tell me the name of this" or "Dígame el nombre de éste." The examiner then coded 1=correct and 0=incorrect. Allowable correct responses were standardized and listed on the protocol sheet. Subjects who gave responses that

were technically correct but at a different level of detail or abstraction were given an additional prompt about the level of detail requested.

In this analysis, low variability items (items with extremely low or extremely high proportion correct) were eliminated to avoid imprecise parameter estimation. Specifically, we did not include items that were answered correctly by greater than 95% or less than 5% of participants in one or both of the two language sub-groups. In summary, twenty-six items were included in the analysis.

Statistical procedures

Dimensionality. An underlying assumption of many IRT models is that the items within a scale are unidimensional, i.e., that a single underlying trait exclusively determines the probability of item responses (Embretson & Reise, 2000). While there are a number of different assumptions, methods, and software available to assess for dimensionality, such as assessing the fit of the data within *Rasch* models (Glas & Verhelst, 1995; Rasch, 1960; Rizopoulos, 2006). For this study, we used *exploratory factor analysis* (EFA) in *Mplus* version 4.1 (Muthén & Muthén, 1998-2009), permuted parallel analysis in *Stata* v. 9 (Buja & Eyuboglu, 1992; Yang & Jones, 2008), *unidimtest* in the *R* program (Drasgow & Lissak, 1983; Rizopoulos, 2006), and *PolyBIF* (Gibbons et al., 2007). We performed these analyses combining the Spanish and English groups, as well as separately to establish dimensional factorial invariance.

Item Response Theory. The following four methods used the two parameter logistic (2-PL) or two parameter normal ogive item response model (Hambleton, Swaminathan, and Rogers, 1991; Lord, 1980; Lord & Novick, 1968): 1) *Item Response Theory Likelihood Ratio Differential Item Functioning* (IRTLRDIF) (Thissen, 2001) and *MULTILOG* (Thissen, 1991); 2) *Parscale and difwithpar* (P. Crane et al., 2007; Muraki & Bock, 2003) used sequentially; 3) *Multiple indicators, multiple causes (MIMIC) model* (Jöreskog & Goldberger, 1975) in *Mplus* (Muthén & Muthén, 1998-2011); and 4) *Differential Functioning of Items and Tests* (DFIT), *Equate* (Baker, 1995), and *MULTILOG* (Thissen, 2001) used in combination. These methods are described below.

IRTLRDIF/MULTILOG (Thissen, 1991; 2001). The freeware program IRTLRDIF procedure has been described in detail elsewhere (Orlando-Edelen, et al., 2006; Teresi, Kleinman, & Ocepek-Welikson, 2000; Thissen, Steinberg, & Wainer, 1993). Briefly, IRTLRDIF involves nested model comparisons of log-likelihoods to detect DIF due to a single two level grouping variable. An iterative process is performed to identify "anchor items" (free of DIF) and "candidate items" (with DIF). Bonferroni or other adjustment methods (e.g., Benjamini-Hochberg) are recommended (Thissen, Steinberg, & Kuang, 2002). The Bonferroni correction applied to this study was based on the procedures recommended by Teresi, Kleinman, and Ocepek-Welikson (2000), using the following chi-square cut-off value and degrees of freedom (df) for each parameter evaluated separately, and adjusted for 26 items: 9.64 (df=1). Items with DIF statistics above this cut-off were considered to have DIF. The MULTILOG software program (Thissen, 1991) is then used to estimate final IRT item parameters and their standard errors. DIF magnitude

measures are not formally a part of this model; however, expected item scores can be used as an accompanying method for quantifying item level magnitude (see Orlando-Edelen, Thissen, Teresi, Kleinman, and Ocepek-Welikson, 2006).

Difwithpar/Parscale. This procedure uses the ordinal logistic regression (OLR) approach to DIF detection (Camilli & Shepard, 1994; Zumbo, 2007). A hybrid IRTOLR approach incorporates ability estimates from IRT software, Parscale (Muraki & Bock, 2003), in the OLR model (Crane, Jolley, & van Belle, 2003). Difwithpar is a Stata routine that calls the Parscale program for latent trait estimation, which is then used in Stata OLR as the conditioning variable. Note there is a freeware program similar to difwithpar, DIFdetect (Crane, Jolley, & van Belle, 2003), developed to detect DIF using IRT ability estimates from other external programs.

DIF detection is based on nested model comparisons. Non-uniform DIF is present if the interaction between the trait estimate and group in predicting item response is statistically significant. Two criteria are available for detection of uniform DIF; a significance test and the recommended magnitude indices. The criteria that was used to determine the presence of uniform DIF is if the regression coefficient of the item response on the trait estimate changes by 10% with and without control for group membership (Crane et al., 2006). Assessment of magnitude of uniform DIF is built into the model using the flagging criteria described above.

Mplus/MIMIC model (Jöreskog & Goldberger, 1975). The MIMIC model is a special case of a confirmatory factor analytic model with covariates (Jones & Gallo, 2002). Graphical depictions of the MIMIC model can be found elsewhere (see Jones, 2003, 2006; Yang & Jones, 2008; Yang, Tommet, & Jones, 2009). When the response data are categorical the model is analogous to the *item response theory* model. The analytic procedure used here was implemented with *Mplus* software, v. 4.1 (Muthén & Muthén, 1998-2009), using the WLSMV estimator and delta parameterization under the multivariate probit modeling framework (Muthén & Muthén, 1998-2009). Three steps are required. The first step estimates a MIMIC model without direct effects, which is essentially a confirmatory factor analysis model with covariates that influence a single underlying latent trait. The second step involves a forward stepwise model building procedure to identify significant direct effects, which reflect the presence of uniform DIF (non-uniform DIF is not assessed). Indirect effects are the relationships between the covariates and observed dependent variables (latent factor indicators) mediated by latent factors (Muthén, 1989). The matrix of fit derivatives (scaled as chi-square and referred to as modification indices) for the regressions of the 26 object naming items on language of test administration was obtained. The third step is to evaluate the significance of model modifications implied by the modification indices through robust chi-square model difference testing was performed using the *Mplus* DIFFTEST function (Yang & Jones, 2007). Magnitude measures are not formally assessed; however, the absolute values of the standardized parameter estimates for the direct effects (items with DIF) can be compared (Yang, et al., 2009).

DFIT/MULTILOG. The DFIT framework is used to detect differential functioning in binary and polytomous items (Flowers, Oshima, & Raju, 1995; Flowers, et al., 1999), using area-based statistics (Hays, et al., 2000; Velicer, Martin, & Collins, 1996). *Non-*

compensatory differential item functioning (NCDIF) is estimated under the assumption that all other items in the scale are free from DIF. Raju and colleagues (1995) use the NCDIF index in DFIT to determine DIF. NCDIF is the average difference squared between the true or expected scores for an individual as a member of the group tested in English and as a member of the group tested in Spanish (see Morales, Flowers, Gutierrez, Kleinman & Teresi, 2006). The *compensatory differential item functioning* (CDIF), and *differential test functioning* (DTF) indexes are also a product of DFIT. Unlike NCDIF, CDIF is not based on the assumption that all other items in the measure are unbiased, but rather takes into account the covariance of the differences in the expected items scores for the given item, and the differences in the total expected sale scores. DTF is the sum of the individual CDIFs.

First, MULTILOG software is used to fit separate IRT models for the reference (Spanish) and focal (English) groups to produce theta estimates for the latent cognitive ability trait, the location (difficulty) of each item on the ability continuum, and the slope of each item. The ability continuum ranges from negative (lower or worse semantic memory ability) to positive (higher or better semantic memory ability). Second, the Spanish and English item parameter estimates are placed on a common metric by computing a set of linking parameters using Baker's Equate software program (Baker, 1995).

The item parameters produced from MULTILOG are subsequently analyzed using the DFIT software. Using the linked parameter estimates from the two preliminary analyses, the DFIT software is used to compute NCDIF, CDIF and DTF. If any items have significant DIF based on the NCDIF index values, subsequent iterations are required. In the next iteration, the linking parameters are re-estimated with Baker's Equate program, restricting the analysis to the set of items without DIF. Subsequent iterations of DFIT include all items, however, with purified linking constants. Iterations are continued until no additional items with DIF are identified. The final set of items used to compute the linking parameters is referred to as the set of "anchor items." The "anchor items" are the set of object naming items free of DIF.

Based on previous simulation studies (Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006), binary items with an NCDIF index value greater than 0.006 are identified as having "significant" DIF. For this analysis, differential functioning at the scale level is identified by summing the CDIF index values for all items in the scale and comparing the sum with a cutoff value of 0.156 (26 items × 0.006 per item). Higher values indicate the presence of differential functioning at the scale level. Because CDIF values can have positive or negative values, it is possible for items with differential functioning to offset each other (i.e., compensatory DIF) resulting in no differential functioning at the scale level, also known as DIF cancellation. Assessment of magnitude of item-level DIF is part of the DIF detection process using NCDIF.

To show the level of agreement between methods concerning the number of items identified with DIF, Kappa coefficients were calculated. According to Landis and Koch (1977), the general rules used to interpret the findings for this study are as follows: values greater than 0.80 represent almost perfect agreement, values between 0.61 to 0.80

represent substantial agreement, values between 0.41 and 0.60 represent moderate agreement, and values less than zero to 0.40 represent poor to fair agreement.

Results

Table 1 shows the characteristics of the 1,159 SENAS participants included in these analyses, of which 383 were tested in English and 776 were tested in Spanish. The age of the participants ranged from 59 to 91 years for those tested in English, with a mean of 69.8 years and a *SD*±7.0. For older Latinos tested in Spanish, the mean age was 70.8 years (*SD*±7.4). For older Latinos tested in English, the mean level of formal education was 11 years (*SD*±4.1) with a range of 0 to 21, while those tested in Spanish had a mean level of formal education of approximately 5 years (*SD*±4.3) with a range from 0 to 19 years. Twenty-six items were examined for DIF due to language. Among the 26 items used in this analysis, there was a higher proportion among those tested in English who answered the items correctly.

Evidence for sufficient unidimensionality was suggested by a permuted parallel analysis that resulted in a plot of the observed eigenvalues against the random sample eigenvalues. Based on both the EFA and *unidimtest*, even though the two-factor model fit well with the data, there were no items with higher loadings on the second factor compared to the first factor. Next, we tested the general factor and specific factor in the bifactor model (using Gibbons's POLYBIF program). All the items loaded strongly on the first factor, except for item 9 (pick), which loaded higher on the specific factor. All items loaded highly on the first factor and none loaded highly on the second factor based on a lambda less than 0.4. In total, the weight of the evidence seems to suggest that these data conform generally to a unidimensional model.

IRTLRDIF/MULTILOG. The IRT item parameters estimated using the four DIF detection methods are presented in Table 2, including those from MULTILOG (after Bonferroni correction of IRTLRDIF results), with discrimination (*a*) parameters rescaled by dividing

Table 1:
Sample characteristics of the older latinos in the SENAS (*n*=1159)

	Tested in English (*n*=383)	**Tested in Spanish (*n*=776)**	**Total (*n*=1159)**	***p*-value**
Gender				
Male	161 (42%)	294 (38%)	455 (39%)	0.174
Female	222 (58%)	482 (62%)	704 (61%)	
Age (*Mean*, *SD*)	69.8, 7.0	70.8, 7.4	70.5, 7.2	0.023
Education (*Mean*, *SD*)	11.0, 4.1	4.6, 4.3	6.7, 5.2	<0.001

Note: *p*-values are for statistical tests of differences between the two groups: tested in English and tested in Spanish.

Table 2:

Comparative results of item parameters with Standard Errors (SE): Presence of DIF related to language of assessment in the Spanish and

ITEM	Mplus/MIMIC			IRTLRDIF/MULTILOG*				difwithpar/Parscale				DFIT/MULTILOG				Unsigned Area				Signed Area			
	Discrimination (a)	Difficulty (b)		Discrimination (a/1.7)		Difficulty (b)		Discrimination (a)		Difficulty (b)		Discrimination (a/1.7)		Difficulty (b)									
	E & S	E	S	E	S	E	S	E	S	E	S	E	S	E	S	M/M	I/M	P/D	M/D	M/M	I/M	P/D	M/D
9 pick-pico	0,7	-1,9	-2,7	0,6	0,6	-1,2	-2,1	0,8	0,6	-1,2	-2,4	0,8	0,5	-0,8	-2,2	0,9	0,9	1,2	1,4	-0,9	-0,9	-1,2	-1,4
10 bird-ave	0,9	-1,6	-1,6	0,8	0,8	-1,5	-1,5	0,9	0,9	-1,7	-1,7	0,8	0,8	-1,4	-1,5	0,0	0,0	0,0	0,1	0,0	0,0	0,0	-0,1
11 coin-moneda	0,9	-0,4	-2,1	0,8	0,8	-0,9	-1,5	1,1	0,8	-1,0	-1,8	1,1	0,7	-0,5	-1,6	1,7	0,7	0,9	1,0	-1,7	-0,7	-0,9	-1,0
12 avocado-aguacate	1,0	-1,3	-1,3	0,9	0,9	-1,2	-1,2	0,9	0,9	-1,4	-1,4	1,0	0,8	-0,9	-1,2	0,0	0,0	0,0	0,3	0,0	0,0	0,0	-0,3
13 gate-puerta	1,0	-0,8	-0,8	0,9	0,9	-0,7	-0,7	1,6	0,8	-0,6	-1,0	1,5	0,8	-0,2	-0,6	0,0	0,0	0,5	0,6	0,0	0,0	-0,4	-0,4
14 cemetary/cementerio	1,4	-0,6	-0,6	1,3	1,3	-0,3	-0,3	1,4	1,4	-0,6	-0,6	1,5	1,2	-0,4	-0,2	0,0	0,0	0,0	0,3	0,0	0,0	0,0	0,3
15 lantern-linterna	1,0	-0,5	0,0	0,8	0,8	-0,7	-0,1	1,0	1,0	-0,5	-0,5	1,1	0,8	-0,4	0,0	0,5	0,6	0,0	0,5	0,5	0,6	0,0	0,4
16 knot-nudo	1,1	-0,6	-0,6	0,9	0,9	-0,4	-0,4	1,0	1,0	-0,7	-0,7	0,9	0,9	-0,3	-0,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
17 teepee-tipi	1,1	-0,3	0,2	1,0	1,0	-0,4	0,2	1,1	1,1	-0,2	-0,2	1,1	0,9	-0,2	0,3	0,5	0,6	0,0	0,5	0,5	0,6	0,0	0,5
18 spear-lanza	0,9	-0,1	-0,1	0,8	0,8	0,2	0,2	0,8	0,8	-0,2	-0,2	1,0	0,7	0,4	0,2	0,0	0,0	0,0	0,3	0,0	0,0	0,0	-0,1
19 artichoke-alcachofa	1,0	-0,1	-0,1	0,9	0,9	0,2	0,2	1,0	1,0	-0,1	-0,1	0,8	0,9	0,2	0,3	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,1
20 llama-llama	1,5	0,0	0,0	1,4	1,4	0,3	0,3	1,2	1,8	0,0	0,0	1,2	1,7	0,4	0,4	0,0	0,0	0,2	0,2	0,0	0,0	0,0	0,0
21 castle-castillo	1,3	0,0	0,0	1,3	1,3	0,3	0,3	1,4	1,4	0,0	0,0	1,1	1,3	0,3	0,5	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,2
22 porcupine-puercoespin	1,2	0,0	-0,3	1,1	1,1	0,3	0,3	1,2	1,2	0,0	0,0	1,3	1,1	0,6	0,3	0,3	0,0	0,0	0,2	-0,3	0,0	0,0	-0,2
23 olive-oliva	1,0	0,5	0,9	1,1	1,1	0,8	0,8	1,1	1,1	0,5	0,5	0,9	1,0	0,7	1,1	0,4	0,0	0,0	0,5	0,4	0,0	0,0	0,5
24 shrimp-camarón	0,7	0,3	-1,5	0,7	0,7	1,8	0,0	0,7	0,7	1,4	-0,3	0,6	0,7	1,9	0,1	1,7	1,8	1,7	1,8	-1,7	-1,8	-1,7	-1,8
25 plum-ciruela	0,6	0,7	0,7	0,6	0,6	1,1	1,1	0,6	0,6	0,7	0,7	0,6	0,6	1,2	1,2	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,0
26 lobster-langosta	0,9	0,8	0,8	0,9	0,9	1,1	1,1	0,9	0,9	0,8	0,8	0,9	0,9	1,2	1,2	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0
27 dragonfly-dragón volador	0,8	0,8	0,3	0,8	0,8	1,5	0,9	1,1	0,7	1,1	0,7	1,0	0,7	1,5	1,2	0,5	0,6	0,5	0,5	-0,5	-0,6	-0,3	-0,4
28 mule-mula	0,7	0,9	0,9	0,6	0,6	1,3	1,3	0,6	0,6	0,9	0,9	0,6	0,7	1,4	1,2	0,0	0,0	0,0	0,3	0,0	0,0	0,0	-0,2
29 date-dátil	1,3	0,9	0,9	1,3	1,3	1,2	1,2	1,4	1,4	0,9	0,9	1,3	1,3	1,3	1,3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
30 pheasant-faisán	1,2	1,1	1,1	1,2	1,2	1,5	1,5	1,3	1,3	1,1	1,1	1,0	1,3	1,5	1,6	0,0	0,0	0,0	0,2	0,0	0,0	0,0	0,1
31 jewel-alhaja	1,1	1,3	1,3	1,1	1,1	1,6	1,6	1,2	1,2	1,2	1,2	1,0	1,0	1,6	1,9	0,0	0,0	0,0	0,3	0,0	0,0	0,0	0,3
32 stone-piedra	0,9	1,1	0,8	0,7	0,9	1,5	1,5	0,6	1,3	1,4	0,8	0,6	1,2	1,9	1,3	0,3	0,3	0,8	0,9	-0,3	0,0	-0,6	-0,7
33 fog-niebla	1,1	1,5	1,8	1,2	1,2	1,8	1,8	1,3	1,3	1,4	1,4	0,9	1,3	1,8	1,9	0,3	0,0	0,0	0,3	0,3	0,0	0,0	0,1
34 dove-paloma	0,9	1,5	1,1	0,9	0,9	2,1	1,6	1,0	0,9	1,7	1,3	0,9	0,9	2,2	1,8	0,4	0,5	0,4	0,5	-0,4	-0,5	-0,4	-0,5

*IRTLRDIF results are after Bonferonni correction

by the constant 1.7 (Lord, 1980). A total of three iterations were required for IRTLRDIF that resulted in nine anchor items (10, 12, 16, 19, 21, 25, 26, 29, and 30, see Table 3 for description of the items corresponding to these items). After the anchor was established, the remainder of the items were tested against that set in the fourth iteration; we identified an additional six items free of DIF (14, 18, 20, 28, 31, and 33) and 11 items with DIF (9, 11, 13, 15, 17, 22, 23, 24, 27, 32, and 34) due to language of test administration. The item with the largest magnitude of DIF was shrimp (camarón, item 24). After Bonferroni correction, seven items (9, 11, 15, 17, 24, 27, and 34) showed uniform DIF (Table 3). After Bonferroni correction, conditional on ability, those tested in English were more

Table 3:
Summary of comparative results of items with DIF detected by Mplus/MIMIC, IRTLRDIF, difwithpar/Parscale, and DFIT/MULTILOG

Items with DIF in any method	**Mplus/ MIMIC**	**IRTLRDIF/ MULTILOG**		**difwithpar/ Parscale**		**DFIT/ MULTI LOG**
	U-DIF	U-DIF	NU-DIF	U-DIF	NU-DIF	DIF
Item 9 : Pick-Pico	**Yes†**	**Yes†**	**No**	**Yes†**	**No**	**Yes†**
Item 11 : Coin-Moneda	**Yes†**	**Yes†**	**No**	**Yes†**	**No**	**Yes†**
Item 13 : Gate-Puerta	No	No	Yes*	No	Yes	Yes
Item 14 : Cemetary/Cementerio	No	No	No	No	No	Yes
Item 15 : Lantern-Linterna	Yes	Yes	No	No	No	Yes
Item 17 : TeePee-Tipi	Yes	Yes	No	No	No	Yes
Item 20 : Llama-Llama	No	No	No	No	Yes	No
Item 22 : Porcupine-Puercoespin	Yes†	Yes†*	No	No	No	Yes†
Item 23 : Olive-Oliva	Yes	Yes*	No	No	No	Yes
Item 24 : Shrimp-Camaron	**Yes†**	**Yes†**	**No**	**Yes†**	**No**	**Yes†**
Item 27 : Dragonfly-Dragon Volador	**Yes†**	**Yes†**	**Yes***	**Yes†**	**Yes**	**Yes†**
Item 32 : Stone-Piedra	**Yes†**	**Yes***	**Yes**	**No**	**Yes**	**Yes†**
Item 33 : Fog-Niebla	Yes	No	No	No	No	No
Item 34 : Dove-Paloma	Yes†	Yes†	No	Yes†	No	No
Total of Items with DIF in each method	11	7‡		8		11

Items with DIF in all of the methods is **bolded**; U-DIF, Uniform Differential Item Functioning, NU-DIF, Non-Uniform DIF

YES* Items found to be DIF-free after Bonferroni adjustment; †Items favor those tested in Spanish

‡Total number of items with DIF after Bonferroni adjustment

likely to answer the following two items with uniform DIF correctly: items 15 (lantern-linterna) and 17 (teepee-tipi), while those tested in Spanish were more likely to answer the following five items with uniform DIF correctly: items 9 (pick-pico), 11 (coin-moneda), 24 (shrimp-camarón), 27 (dragonfly-dragón volado), and 34 (dove-paloma). After Bonferonni correction, only item 32 (stone-piedra) evidenced non-uniform DIF. At average and above average abilities, those tested in Spanish were more likely to answer the item correctly; while at lower ability levels, those tested in English were more likely to answer the item correctly. This information is summarized with signed and unsigned areas in Table 2.

Mplus/MIMIC. Of the 11 items found with DIF in IRTLRDIF before Bonferroni correction, the following 10 items were also found with DIF using the *Mplus*/MIMIC model: 9, 11, 15, 17, 22, 23, 24, 27, 32, and 34. In comparison with the IRTLRDIF procedure with Bonferroni correction, the MIMIC approach identified seven items (9, 11, 15, 17, 24, 27, and 34) in common that evidenced uniform DIF. In addition, the MIMIC model identified DIF for item 33 (fog-niebla). Again, the item with the largest magnitude of DIF was item 24 (shrimp-camarón).

Difwithpar/Parscale. Five items identified with uniform DIF (9, 11, 24, 27, and 34) using Difwithpar/Parscale were also found to have DIF using both the *Mplus*/MIMIC and IRTLRDIF/MULTILOG approaches. Difwithpar/Parscale also identified one additional item (13) that was also found to have DIF using IRTLRDIF/MULTILOG, before the Bonferroni correction, method. Item 20 (llama) was the only item found to have DIF attributable to language that was not found to have DIF using any of the other procedures. Table 3 provides a summary of a total of five items with uniform DIF (9,11,24,27, and 34) and four items with non-uniform DIF (13, 20, 27, and 32) found through the difwithpar/Parscale method.

DFIT/MULTILOG. The same five items as the three other methods mentioned above (9, 11, 24, 27, and 32) were found to have DIF attributable to language of test administration in DFIT/MULTILOG. In addition, item 14 (cemetery-cementario) demonstrated DIF only for the DFIT method. As with two other DIF methods used in the study, before Bonferroni adjustment the DFIT methods detected DIF in items 13, 15, 17, 22, and 23. Items 20, 33, and 34 did not show DIF under DFIT, but did evidence DIF by one or more of the other methods. In the final run, the differential test functioning (DTF) index (sum of the CDIF indices) was 0.31. This exceeded the cutoff value of 0.156, which suggests DTF between the two language administration groups. While the overall DTF index is reduced to 0.14 when item 11 (coin-moneda) is excluded, the item with the largest NCDIF value (indicative of a higher magnitude of DIF) is item 24 (shrimp-camarón) with a value of 0.12. This is consistent with the results from IRTLRDIF/MULTILOG.

For this study, we used the 2-PL model to estimate both discrimination (*a*) and difficulty (*b*) parameters using each of the methods described above. As an aside, there is also a third parameter (*c*), also known as the guessing or pseudo-chance-level (Hambleton, et al., 1991) parameter that is incorporated into the three-parameter logistic model (Birnbaum, 1968). We note that there are other approaches for detecting DIF, such as using the Rasch model. Within Rasch modeling, there are also different types of software

for estimating item parameters and DIF detection, one of which is *Andersen's Likelihood Ratio Test* (LRTest/eRm; Andersen, 1973) in R (Mair & Hatzinger, 2007). We conducted the LRTest to determine if the Rasch model would be appropriate for comparison with the other methods. An LR-value of 356, χ^2 *df*=25, $p<.001$ was observed, indicating that there was a significant difference in the difficulty parameters between the English and Spanish language groups. Therefore, the use of the 2-PL model for this analysis was confirmed.

SUMMARY. Descriptions of all the items used in this analysis corresponding to their numbers are given in Table 3. There were five items (9, 11, 24, 27, and 32) that showed DIF with respect to language of test administration across all four methods. Those tested in Spanish were more likely to get these five items correct when compared to those tested in English. An additional six items (13, 15, 17, 22, 23, and 34) evidenced DIF by at least three of the four methods, before Bonferroni correction. All of these items except items 22 and 34 were more likely answered correctly by those tested in English. Consistently, each method showed that item 24 (shrimp-camarón) showed the greatest magnitude of DIF; with those tested in English having more difficulty than those tested in Spanish. Finally, there was agreement in findings of no DIF across all methods for items: 19, 12, 18, 19, 21, 25, 26, 28, 29-31, yielding a total agreement rate of 62%. The mean difference in overall mean (± Standard Deviation (*SD*)) test performance difference between English and Spanish groups for all items were 15.9 (±5.3) and 10.3 (±5.4), respectively. After omitting the five items with DIF common across the four methods, the mean difference in overall mean test performance was 13.1 (±4.4) and 7.7 (±4.6), in respective order for English and Spanish.

Table 3 shows all items with uniform DIF (U-DIF) and non-uniform DIF (NU-DIF) indicated by a "YES." Items found to be DIF-free after Bonferroni adjustment are noted by "YES*." For example, for item 32, both uniform and non-uniform DIF was found. But after Bonferroni adjustment, only non-uniform DIF was present in item 32.

Using the item parameters estimated with each of the DIF methods, the test response function (Figure 1) showed that the expected score for respondents tested in Spanish was slightly higher than that of their English-language counterparts between ability (*theta*) levels of –3 and 3. At average object naming ability (*theta*=0), the probability of a correct response is slightly higher for Spanish speakers than English speakers using the following methods: IRTLRDIF/MULTILOG, *Mplus*/MIMIC, and difwithpar/Parscale. However, the overall DIF cancelled at the scale level because some items with DIF were more difficult for English speakers, while others were more difficult for Spanish speakers.

Kappa coefficients were calculated to examine the agreement across methods (Table 4). The results indicated that *Mplus*/MIMIC is in almost perfect agreement with IRTLRDIF/MULTILOG with and without Bonferroni correction (κ=0.75 and κ=0.84, respectively). *Mplus*/MIMIC and DFIT/MULTILOG were in substantial agreement (κ=0.68) with each other. Difwithpar/Parscale demonstrated moderate agreement with IRTLRDIF/MULTILOG (κ=0.59) and substantial agreement with Bonferroni corrected IRTLRDIF/MULTILOG (κ=0.64). DFIT/MULTILOG also showed moderate agreement

Figure 1: Test characteristic curve for object naming items by Spanish and English Language of test administration using IRTLRDIF/MULTILOG, MIMIC, *Mplus*, DFIT/MULTILOG, and difwithpar/Parscale.

Table 4:
Kappa coefficients for IRTLRDIF/MULTILOG, Mplus/MIMIC, DFIT/MULTILOG, and difwithpar/Parscale.

		(1)	(2)	(3)	(4)	(5)
(1)	IRTLRDIF/MULTILOG	1.00	0.84	0.84	0.75	0.59
(2)	Mplus/MIMIC		1.00	0.68	0.75	0.43
(3)	DFIT/MULTILOG			1.00	0.59	0.43
(4)	IRTLRDIF with Bonferroni correction/MULTILOG				1.00	0.64
(5)	difwithpar/Parscale					1.00

with Bonferroni corrected IRTLRDIF/MULTILOG (κ=0.59), but almost perfect agreement with IRTLRDIF/MULTILOG (κ=0.84). Relative to comparisons between the other three methods, the lowest kappa coefficients were found between difwithpar/Parscale and both *Mplus*/MIMIC (κ=0.43) and DFIT/MULTILOG (κ=0.43) at moderate agreement.

Discussion

Overall, there is evidence for at least moderate agreement across the four DIF detection methods. There is almost perfect agreement between a commonly-used method for detecting DIF, IRTLRDIF/MULTILOG, with two other methods: DFIT/MULTILOG and *Mplus*/MIMIC. A recently proposed method for detecting DIF is the difwithpar/Parscale approach of Crane and colleagues (2003), which is only in moderate agreement with the other methods.

It is important in the development phase of assessments to perform qualitative analyses to ensure conceptually equivalent measures. Poor translation and lack of conceptual and cultural equivalence can impact quantitative results, such as those presented here. Many existing cognitive status measures were not developed with culturally diverse groups in mind, with the exception of the SENAS. Adequate Spanish-language cognitive measures are essential for accurate clinical assessment of Spanish-speaking patients, as well as for making meaningful comparisons between groups in studies of race/ethnic differences in cognitive status. Although a number of IRT-based approaches exist for assessing the comparability of self-reported measures across culturally diverse groups, previous work has not systematically compared estimates of DIF generated by these different approaches. Though language-related DIF is an important topic unto itself, application of four approaches helps compare and contrast the different methods. The number of SENAS object naming items identified with DIF due to language of test administration ranged from 7 to 11 across the four methods, with agreement on only five items, after Bonferroni adjustment. Excluding difwithpar/Parscale, there is agreement on the pres-

ence of DIF with respect to 7 items (9, 11, 15, 17, 24, 27, and 32; see Table 3 for description of items).

The five items identified with DIF across all four methods are item 9 (pick-pico), item 11 (coin-moneda), item 24 (shrimp-camarón), item 27 (dragonfly-dragón volado), and item 32 (stone-piedra). Each of these items were easier for respondents tested in Spanish than for those tested in English, except for item 32, which varied between the language groups based on ability levels. The reason that the four items were easier to answer is perhaps due to cultural and regional differences in familiarity with pictures from the semantic memory test such as exposure to shrimp. As the item with the largest DIF magnitude across all methods, "shrimp" is more difficult for respondents tested in English than "camarón" is for those tested in Spanish after controlling for overall object naming ability. This suggests that the stimulus (a shrimp with head, tail, and legs) is more familiar to Spanish speakers or that the word "camarón" may be easier to retrieve in response to that picture than is "shrimp" in English. Supplementary data on types of incorrect responses would be valuable to understand better the reasons that this item showed DIF.

The variability in the number of items with DIF across the four analytic approaches is likely due to differences in operational definitions and tests of DIF. While all methods presented here use an IRT-based method, they vary in the criteria used to flag DIF. IRTLRDIF uses a likelihood ratio test statistic, and accompanying chi-square test of DIF. Also used in the final step was the Bonferroni correction, which is a simple and conservative criterion to determine DIF-free items by lowering the alpha value to avoid false positives when simultaneously comparing the presence of DIF across all items. Both uniform and non-uniform DIF is assessed. The MIMIC approach is similar and estimates DIF using direct effects from a measurement structural equation model as indicated by model misfit indices; only uniform DIF is detected. The difwithpar approach is different from the other methods, as it incorporates both significance tests for non-uniform DIF and changes in the difficulty parameters for uniform DIF, a hybrid use of both significance tests and measures of magnitude. The DFIT program defines DIF as the difference between probabilities of a positive item response for individuals from different groups at the same level of the latent trait, and detects, but does not distinguish between uniform and non-uniform DIF.

Both the IRTLRDIF and the MIMIC methods identify items with DIF on the basis of statistical significance, which is determined in part by sample size, although variants of the IRTLRDIF approach includes the incorporation of magnitude measures for the final selection of items with salient DIF (see Teresi, et al., 2000). In contrast, the DFIT procedure uses cut-off values of DIF magnitude that were determined based on simulations, rather than statistical significance for detecting items with DIF. The IRTOLR approach is based on statistical tests to identify non-uniform DIF, and the incorporation of magnitude measures to flag uniform DIF. Specifically for this analysis; the uniform DIF test for the IRTOLR approach is based on a 10% change in beta.

The presence of item-level DIF does not necessarily imply that group comparisons of the latent variable are biased at the scale level. Other items with DIF that favor English speakers may cancel out the DIF exhibited by some items that favor Spanish speakers.

An advantage of DFIT over the other three procedures examined is that it indicates whether such DIF cancellation occurs at the scale level. IRTLRDIF accompanied by expected item and scale scores also permits such evaluation. Similarly, the IRTOLR approach can be accompanied by tests of impact through comparisons of means. However, like IRTLR, the impact analyses are external to the procedure. A relative advantage of the MIMIC procedure is that it allows one to adjust the items with DIF due to other covariates in the model. A possible disadvantage of the single group MIMIC model is that it does not detect non-uniform DIF; while DFIT detects the presence of non-uniform DIF, but does not specifically identify these items.

In terms of the ease of use or user-friendliness, the DFIT/MULTILOG may be the most challenging for researchers beginning to run DIF analyses because the procedure requires the use of the following three different programs: MULTILOG, Baker's EQUATE, and DFIT. The procedure that is relatively easier, requiring two programs, is using IRTLRDIF and MULTILOG. However, the procedure includes several iterations within IRTLRDIF: choosing anchor items, purifying anchor items, and calculation of the Bonferroni correction to determine the final items with DIF. The difwithpar/Parscale is implemented via a freely available Stata macro, and we have developed Stata macros for governing the *Mplus*/MIMIC and IRTLRDIF procedures that are freely available upon request.

Limitations. There are several limitations of this study. First, because of low variability (prevalence of errors less than 5% in this sample), 18 items from the original 44-item object naming test are excluded from the analyses (eight items with low variability in English alone, five in Spanish alone and five with low variability in both). Low variability can result in estimation problems for the methods used in this study. These items may well be important for measuring object naming ability in clinical settings, and indeed, an explicit goal in SENAS construction was for items to span a very broad range of difficulty. The higher ability items will likely be more relevant for English speaking and more highly educated individuals, while the low ability items will be more applicable to Spanish speaking, low education, and more impaired individuals. This study addresses items that fall within the primary range of overlap for the English-speaking and Spanish-speaking distributions, and this is the range where measurement bias is most important. That is, items outside this range make a limited contribution to assessing object naming ability for one of the two groups.

Second, extension of the findings to all English and Spanish speakers generally is not warranted, as the study only included older Mexican Americans, living in Northern California, who chose to be tested in either English or Spanish. More complete evaluation of the different methods for DIF detection should be completed with a simulation study. Although simulation studies are regarded as closer to the gold standard, comparing DIF analyses using real data across different methods is still informative as this is the first study of its kind. Examining the operating characteristics of a measure assumes a gold standard outcome, so future research will require both comparisons to simulation data and a clinical outcome (Holland & Wainer, 1993; Teresi, Stewart, Morales, & Stahl, 2006).

Finally, we acknowledge that the addition of other covariates in the model – such as acculturation level, wealth, income, and education – are important to determine DIF due to language differences, but half of the methods used in this study can only determine DIF due to one covariate. Therefore, in order to fulfill the purpose of this study, we were limited to one covariate to determine DIF across four different methods. However, the future direction of this study is to include other covariates in the model using *Mplus*/MIMIC.

Conclusion

There was substantial convergence of results across methods, but there also were differences. DIF is important to the extent that it biases scale level measurement, that is, results in differential validity across groups. Thus, the similarities and differences in the results from these four methods must be understood in the context of resulting effects on individual ability levels and average differences across groups. If the cumulative effect of DIF at the scale level does not change estimated ability for individual examinees, then DIF does not present a measurement problem even if many individual items have DIF. However, even if only a few items have DIF but there is a systematic bias of ability estimates, this would present an important measurement problem. Accounting for effects of DIF on ability estimates is an important component of the process of evaluating measurement bias.

A question arises regarding the practical utility of analyses of DIF and of the best methods for DIF detection. As computerized adaptive testing (CAT) gains popularity in the fields of health and neuropsychology (Reeve, 2006), it is increasingly important to ensure that the item banks feeding the CAT are adequate to the task. For example, the Division of Neurosciences at the United States National Institutes of Health has undertaken a major effort to assemble assessment tools that clinicians and researchers might use to measure outcomes; a key area of inquiry is cognitive function. Efforts such as Toolbox (www.nihtoolbox.org) that focus on assessment of neurological and behavioral function, and the Patient Reported Outcome Measurement Information System (PROMIS) (www.nihpromis.org) that focuses on the construction of items banks rely on the inclusion of items that have been examined for measurement equivalence. The results of efforts such as those presented here have been integrated into the products of these international initiatives. When several methods consistently identify items with a high magnitude of DIF, content experts and investigators are more likely to consider modification. For example, high magnitude DIF found in two of the PROMIS depression item bank items resulted in their deletion from the item bank (Teresi et al., 2009). The shrimp item identified in these analyses will be removed from the object naming items of the SENAS, thus demonstrating the importance and practical consequences of such analyses.

As the populations of many countries become more linguistically, racially, and culturally diverse, issues of measurement bias in both research and clinical settings will become more important. This line of inquiry has received surprisingly little attention in the neu-

ropsychology literature. The methods used in this study can make an important contribution to improving measurement in diverse populations.

Acknowledgement

Support for this collaborative project was provided by the National Institute on Aging through the six Resource Centers for Minority Aging Research (RCMAR: P30AG015294, P30AG021677, P30AG015292, P30AG021684, P30AG015272, and P30AG015281), and the RCMAR Coordinating Center (P30AG021684). We would also like to acknowledge the UC Davis Alzheimer's Disease Center (UCD ADC, P30AG010129) for providing the SENAS data funded by AG10220 (PI Mungas) and AG12975 (PI Haan). Additional funding provided through the Harvard Older Americans Independence Center (P60AG00812 PI Lipsitz), the National Institute on Aging (5-T32 AG023480 PI Lipsitz; AG025308 PI Jones; AG025444 PI Mehta). We would like to thank (in alphabetical order) the advice, mentorship, and support from: Drs. Laura Gibbons, Jack Goldberg, Rafael A. Lantigua, Mildred Ramirez, Thomas N. Templin, and Barbara Tilley.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

Baker, F. B. (1995). *EQUATE Computer program, Version 2.1*. Madison, Wisconsin: Laboratory of Experimental Design. Department of Educational Psychology. University of Wisconsin.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Borsboom, D. (2006). When does measurement invariance matter? *Med Care, 44*(11 Suppl 3), S176-181.

Buja, A., & Eyuboglu, N. (1992). Remarks on Parallel Analysis *Multivariate Behavioral Research, 27*(4), 509-540.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. Newbury Park, California: Sage Publishers.

Crane, P., Gibbons, L., Jolley, L., & van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Med Care, 44*(11 Suppl 3), S115-123.

Crane, P., Gibbons, L., Ocepek-Welikson, K., Cook, K., Cella, D., Narasimhalu, K., et al. (2007). A comparison of three sets of criteria for determining the presence of differential item functioning using ordinal logistic regression. *Qual Life Res, 16 Suppl 1*, 69-84.

Crane, P., Jolley, L., & van Belle, G. (2003). *DIFdetect [computer program and documentation, available at http://www.alz.washington.edu/DIFDETECT/welcome.html]*. Seattle, WA: University of Washington, National Alzheimer's Coordinating Center.

Crane, P. K., Gibbons, L. E., Jolley, L., van Belle, G., Selleri, R., Dalmonte, E., et al. (2006). Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. *International Psychogeriatrics*, 1-11.

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

Drasgow, F., & Lissak, R. I. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*, 363-373.

Druks, J., Masterson, J., Kopelman, M., Clare, L., Rose, A., & Rai, G. (2006). Is action naming better preserved (than object naming) in Alzheimer's disease and why should we ask? *Brain Lang, 98*(3), 332-340.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Flowers, C., Oshima, T., & Raju, N. S. (1995). *A Monte Carlo assessment of the DFIT with dichotomously-scored unidimensional tests*, Georgia State University, Atlanta, GA.

Flowers, C., Oshima, T., & Raju, N. S. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res, 12*(3), 189-198.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2007). Full-Information Item Bifactor Analysis of Graded Response Data *Applied Psychological Measurement, 31*(1), 4-19.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models. Foundations, Recent Developments, and Applications* (pp. 69-95). New York: Springer.

Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park: SAGE Publications.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Identification of Potentially Biased Test Items. *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publications, Inc.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 28*(9SII), II-28-II-42.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. New York: Lawrence Erlbaum Associates.

Jones, R. N. (2003). Racial bias in the assessment of cognitive functioning of older adults. *Aging & Mental Health, 7*(2), 83-102.

Jones, R. N. (2006). Identification of measurement differences between English and Spanish language versions of the Mini-Mental State Examination. Detecting differential item functioning using MIMIC modeling. *Med Care, 44*(11 Suppl 3), S124-133.

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the mini-mental state examination: effects of differential item functioning. *J Gerontol B Psychol Sci Soc Sci, 57*(6), P548-558.

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

Kessler, R. C., Merikangas, K. R., Berglund, P., Eaton, W. W., Koretz, D. S., & Walters, E. E. (2003). Mild disorders should not be eliminated from the DSM-V. *Arch Gen Psychiatry, 60*(11), 1117-1122.

Króliczak, G., Westwood, D. A., & Goodale, M. A. (2006). Differential effects of advance semantic cues on grasping, naming, and manual estimation. *Experimental Brain Research*.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

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

Lord, F., & Novick, M. (1968). Latent traits and item characteristic functions (Chapter 16) *Statistical Theories of Mental Test Scores* (pp. 358-393). Reading, MA: Addison-Wesley.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20.

Millsap, R. E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Med Care, 44*(11 Suppl 3), S171-175.

Morales, L. S., Flowers, C., Gutierrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the Differential Item and Test Functioning (DFIT) Framework. *Med Care, 44*(11 Suppl 3), S143-151.

Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & Gonzalez, H. (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): further development and psychometric characteristics. *Psychol Assess, 16*(4), 347-359.

Mungas, D., Reed, B. R., Haan, M. N., & Gonzalez, H. (2005). Spanish and English neuropsychological assessment scales: relationship to demographics, language, cognition, and independent function. *Neuropsychology, 19*(4), 466-475.

Mungas, D., Reed, B. R., Marshall, S. C., & Gonzalez, H. M. (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. *Neuropsychology, 14*(2), 209-223.

Muraki, E., & Bock, D. (2003). PARSCALE for windows (Version 4.1). *Chicago: Scientific Software International*.

Muthén, B. O. (1989). Dichotomous factor analysis of symptom data. *Soc Meth and Res, 18*(1), 19-65.

Muthén, L., & Muthén, B. (1998-2011). Mplus Version 6.11. Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998-2009). Mplus Version 5.2. Los Angeles: Muthén & Muthén.

Orlando-Edelen, M., Thissen, D., Teresi, J., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach. Application to the Mini-Mental State Examination. *Med Care, 44*(11 Suppl 3), S134-142.

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.

Rasch, G. (1960). Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.

Reeve, B. B. (2006). Special issues for building computerized-adaptive tests for measuring patient-reported outcomes: the National Institute of Health's investment in new technology. *Med Care, 44*(11 Suppl 3), S198-204.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25.

Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Stat Med, 19*(11-12), 1651-1683.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Crane, P. K., Jones, R. N., et al. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly, 51*(2), 148-180.

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psychology Science Quarterly, 50*(4), 538-612.

Teresi, J. A., Stewart, A. L., Morales, L. S., & Stahl, S. M. (2006). Measurement in a multi-ethnic society. Overview to the special issue. *Med Care, 44*(11 Suppl 3), S3-4.

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

Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning* (2.0b ed.). University of North Carolina at Chapel Hill: L.L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77-83.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Velicer, W. F., Martin, R. A., & Collins, L. M. (1996). Latent transition analysis for longitudinal data. *Addiction, 91 Suppl*, S197-209.

Yang, F. M., & Jones, R. N. (2007). Center for Epidemiologic Studies-Depression Scale (CES-D) item response bias found with Mantel-Haenszel method was successfully replicated using latent variable modeling. *J Clin Epidemiol, 60*(11), 1195-1200.

Yang, F. M., & Jones, R. N. (2008). Measurement differences in depression: chronic health-related and sociodemographic effects in older Americans. *Psychosomatic Medicine, 70*(9), 993-1004.

Yang, F. M., Tommet, D., & Jones, R. N. (2009). Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: an extension of the bi-factor item response theory model for use in differential item functioning. *J Psychiatr Res, 43*(12), 1025-1035.

Zec, R. F., Markwell, S. J., Burkett, N. R., & Larsen, D. L. (2005). A longitudinal study of confrontation naming in the "normal" elderly. *J Int Neuropsychol Soc. , 11*(6), 716-726.

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Asessment Quarterly, 4*(2), 223-233.