Published in final edited form as:
J Int Neuropsychol Soc. 2009 September ; 15(5): 758–768. doi:10.1017/S1355617709990361.

Differential Item Functioning of the Boston Naming Test in Cognitively Normal African American and Caucasian Older Adults

Otto Pedraza[1], **Neill R. Graff-Radford**[2], **Glenn E. Smith**[3], **Robert J. Ivnik**[3], **Floyd B. Willis**[4], **Ronald C. Petersen**[5], and **John A. Lucas**[1]

[1]Department of Psychiatry and Psychology, Mayo Clinic, Jacksonville, Florida

[2]Department of Neurology, Mayo Clinic, Jacksonville, Florida

[3]Department of Psychiatry and Psychology, Mayo Clinic, Rochester, Minnesota

[4]Department of Family Medicine, Mayo Clinic, Jacksonville, Florida

[5]Department of Neurology, Mayo Clinic, Rochester, Minnesota

Abstract

Scores on the Boston Naming Test (BNT) are frequently lower for African American when compared to Caucasian adults. Although demographically-based norms can mitigate the impact of this discrepancy on the likelihood of erroneous diagnostic impressions, a growing consensus suggests that group norms do not sufficiently address or advance our understanding of the underlying psychometric and sociocultural factors that lead to between-group score discrepancies. Using item response theory and methods to detect differential item functioning (DIF), the current investigation moves beyond comparisons of the summed total score to examine whether the conditional probability of responding correctly to individual BNT items differs between African American and Caucasian adults. Participants included 670 adults age 52 and older who took part in Mayo's Older Americans and Older African Americans Normative Studies. Under a 2-parameter logistic IRT framework and after correction for the false discovery rate, 12 items where shown to demonstrate DIF. Six of these 12 items ("dominoes," "escalator," "muzzle," "latch," "tripod," and "palette") were also identified in additional analyses using hierarchical logistic regression models and represent the strongest evidence for race/ethnicity-based DIF. These findings afford a finer characterization of the psychometric properties of the BNT and expand our understanding of between-group performance.

Keywords

Boston Naming Test; Item response theory; Differential item functioning; Ethnicity; Race; Bias

Introduction

The Boston Naming Test (BNT; Kaplan et al., 1983) is one of the most widely used neuropsychological measures in the clinical assessment and investigational study of visual naming ability. Numerous investigators, however, have urged caution in using the BNT in African Americans and other ethnic minorities because of differential performances as compared to Caucasians (Boone et al., 2007; Fillenbaum et al., 1997, 1998; Inouye et al., 1993; Lichtenberg et al., 1994; Manly et al., 1998; Wagner et al., 2007; Welsh et al., 1995;

Correspondence concerning this article should be addressed to: Otto Pedraza, Department of Psychiatry and Psychology, Mayo Clinic, 4500 San Pablo Road, Jacksonville, FL, USA 32224, Tel: 904-953-7286; Fax: 904-953-0461, otto.pedraza@mayo.edu.

Whitfield et al., 2000). Potential explanations for the observed discrepancy between African American and Caucasian BNT performance include differences in educational attainment and literacy (Manly et al., 2002, 2004), health-related risk factors (Wagner et al., 2007; Whitfield et al., 2000), and cultural appropriateness of test items (Manly et al., 1998; Miles, 2002; Whitfield et al., 2000).

From a clinical standpoint, demographically-corrected normative data for African Americans can mitigate the impact of these discrepancies on the likelihood of erroneous diagnostic impressions (Boone et al., 2007; Heaton et al., 2004; Lucas et al., 2005a; Smith et al., 2008). However, group norms do not sufficiently address or ultimately advance our understanding of the underlying psychometric and sociocultural factors that lead to between-group score discrepancies (Brandt, 2007; Manly, 2005).

Differential item functioning (DIF) represents a modern psychometric approach to the investigation of between-group score discrepancies. Under equivalent testing conditions, it is expected that individuals from different groups but comparable ability level will have a similar probability of responding correctly to a given test item. An item displays DIF when the conditional probability of obtaining a correct response differs between individuals who have been matched on the underlying ability construct (Hambleton et al., 1991; American Educational Research Association, 1999). Using the framework of the current study, it would be expected that a randomly selected Caucasian adult and a randomly selected African American adult with comparable naming ability will have a similar probability of responding correctly to any given BNT item. If the conditional probability of a correct response differs between the two ability-matched groups, then the item demonstrates DIF. It should be noted that for an item to be free of DIF, the conditional probabilities between the groups do not need to be equal, but sufficiently similar. As described below, this will be operationalized as a statistically non-significant difference between item parameter estimates.

An item can demonstrate DIF in one of two forms. Uniform DIF is present when the probability of a correct response is greater for one group than another across all levels of ability. Conversely, nonuniform DIF is present when the probability of a correct response varies across the ability spectrum. For example, when comparing two groups of individuals on a test such as the BNT, nonuniform DIF would be present if the probability of a correct response for a particular item is higher for one group at the lower levels of naming ability, but higher for the other group at higher naming ability levels.

Item response theory (IRT) provides an attractive framework for the investigation of uniform and nonuniform DIF (Teresi et al., 2000). Conceptually, IRT models are based on the notion that a person's performance on a particular test depends on the parametric properties of each test item and the person's trait or latent ability level (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985). In other words, an IRT model expresses the probabilistic association between a person's observable item responses and their unobservable but estimated ability level. Under a 2-parameter logistic IRT model, the latent ability level (θ) is estimated based upon the person's pattern of passed versus failed items when taking into account each item's discrimination and difficulty parameters. Item discrimination (α) represents the degree to which the item can distinguish individuals with higher ability from those with lower ability and is closely related in classical test theory to the biserial correlation between each item response and the total test score. Item difficulty (β) represents the ability level at which a person has a 50% chance of responding correctly to an item. A 2-parameter model that estimates item discrimination and difficulty seems particularly well suited for the investigation of cognitive ability. Models incorporating a third, ‘guessing’ parameter, are most helpful in multiple-choice test formats more frequently encountered in educational than mental status assessments (Teresi

et al., 2000). A 1-parameter (Rasch) model has not been recommended for DIF-detection in health data (Teresi, 2006).

Item discrimination and difficulty are best visualized and more easily understood using item characteristic curves (ICC), which relate the probability of an item response to the underlying ability construct. In an ICC, item difficulty is represented by the location along the x-axis at which point the probability of a correct response for a binary item is 50%, and item discrimination is represented by the slope of the trace line at that location parameter. A steeper slope (corresponding to a higher α value) reflects a higher degree of discrimination. If a test item is DIF-free, the ICCs on that item are comparable and highly overlapping. An item with uniform DIF is characterized by different β parameters between the two groups, resulting in nonoverlapping but relatively parallel curves. In contrast, an item with nonuniform DIF has different α parameters between the groups, resulting in nonparallel curves (see Figure 1 for an illustration).

An advantage of IRT over classical test theory is that reliability is not constrained to a single coefficient, but instead can be measured continuously over the entire ability spectrum. In IRT models, reliability is equivalent to the concept of information, which is inversely related to standard error of measurement and represents the degree of precision at each ability level (Embretson & Reise, 2000). Item information is therefore maximized by higher discrimination parameters and an adequate match between item difficulty and a person's ability level. Moreover, item information is additive and yields a test information curve that represents the overall degree of precision at each ability level.

Modern psychometric methods hold substantial promise in propelling new perspectives on the relationship between racial/ethnic background and cognitive functions (Pedraza & Mungas, 2008). Knowledge of item-level psychometric properties takes us beyond demographically-based group norms, thus advancing our understanding of the underlying factors contributing to test score discrepancies. Within this framework, the current study investigates the presence of DIF on the BNT between African American and Caucasian older adults, with the goal of highlighting items that may be problematic when comparing BNT scores across these groups.

Method

Participants

Participants included 670 adults who took part in Mayo's Older Americans and Older African Americans Normative Studies (MOANS and MOAANS, respectively) and for whom item-level data from the Boston Naming Test were available. Study criteria and recruitment protocol for the MOANS and MOAANS projects have been described previously (Ivnik et al., 1990; Lucas et al., 2005a). In brief, cognitively normal older adults were defined as community-dwelling, independently functioning individuals examined by their primary care physician within one year of study entry and who met the following criteria: (1) normal cognition based on self, informant, and physician reports; (2) capacity to independently perform activities of daily living based on informant report; (3) no active or uncontrolled CNS, systemic, or psychiatric condition that would adversely affect cognition, based on physician report; and (4) no use of psychoactive medications in amounts that would be expected to compromise cognition or for reasons indicating a primary neurologic or psychiatric illness. All data were obtained in full compliance with study protocols approved by the Mayo Clinic Institutional Review Board.

Table 1 presents the demographic characteristics of the study participants. There was no significant difference between Caucasian and African American participants in the proportion of males to females ($\chi^2(1) = 2.04$, $p = .15$). However, Caucasian participants were significantly

older ($t(668) = 7.72$, $p < .001$, Cohen's $d = .60$) and had more years of formal education ($t(665) = 7.26$, $p < .001$, Cohen's $d = .56$).

Measure

The Boston Naming Test was administered using standardized published instructions (Kaplan et al., 1983), beginning with item 30 and proceeding until a basal level of 8 consecutive correct responses was established. Consequently, items 1-29 were administered to less than one-third of the total sample, resulting in a restricted sample size for parametric DIF analyses. Thus, only items 30-60 were considered for the current study. Test administration proceeded until each participant reached a ceiling performance of 6 consecutive failures using a lenient discontinuation rule (Ferman et al., 1998). Under this lenient rule, correct responses provided after phonemic cues were not counted toward the 6 consecutive failures. Within the African American sample, regional synonyms were accepted as correct for certain items (e.g., “mouth harp” for harmonica, “tom walkers” for stilts) (Lucas et al., 2005b). Thirty-eight participants did not reach the final test item using this discontinuation criterion. For half of these participants ($n = 19$), the modal number of non-administered items beyond the discontinuation criteria was 1. To explore the potential effect of inclusion versus exclusion of the remaining 19 participants who had a greater proportion of non-administered items, all IRT DIF analyses were performed using two separate datasets. In the first dataset (Restricted), the 19 participants with non-administered items were excluded from all analyses, resulting in a working sample size of 651 participants. In the second dataset (Full), all 670 participants were included and items not administered beyond the discontinuation rule were coded as incorrect, consistent with standard test administration instructions. The two sets of results were nearly identical, with only one item (“scroll”) demonstrating DIF when analyzed under the Full dataset, but no DIF when analyzed under the Restricted dataset. Given the similarity in findings between the two datasets and the use of discontinuation rules in standard clinical and research practice, only the results from the Full dataset will be presented and discussed.

Statistical Analyses

Power and sample size for IRT—Formal power analytic tools are not currently available to make sample size determinations for IRT studies. However, simulation data using 2-parameter logistic IRT models (Holman et al., 2003) suggest that our sample size was more than sufficient to detect moderate effects.

IRT assumptions—Accurate parameter and ability estimation under the current IRT model requires that the assumptions of unidimensionality and local independence be satisfied. Unidimensionality implies that a single latent trait is sufficient to account for the observed pattern of item responses. As such, the probability of responding correctly to a test item will be a function of that single underlying trait. Local independence implies that only the latent trait accounts for the relationship among test items at any given level of ability; thus, after partialing out the underlying trait there should be minimal to no residual correlation between item pairs. Unidimensionality and local independence are interrelated assumptions, such that violations of conditional independence between any two item responses may result in spurious, unintended dimensions that lead to inaccurate ability and parameter estimates.

Unidimensionality was assessed separately on the African American and Caucasian data using several methods. First, exploratory principal components analyses were conducted on the tetrachoric correlations using PRELIS 2.0, followed by confirmatory factor analyses using an asymptotic distribution-free estimator (i.e., diagonally weighted least squares) using LISREL 8.80 (Jöreskog & Sörbom, 2006). Acceptable model fit was evaluated with the comparative fit index (CFI, values >0.90 indicate better fit) and the root-mean-square error of approximation (RMSEA, values <0.10 indicate better fit). A known limitation of ADF estimators, however,

is the requirement for substantially large sample sizes to generate proper solutions (Boomsma & Hoogland, 2001). For instance, Jöreskog and Sörbom (1997) recommend a minimum sample size of $1.5p(p+1)$, where p represents the number of observed indicators. Alternatively, nonparametric approaches are also available to evaluate essential unidimensionality and local independence. DIMTEST 2.0, a nonparametric conditional covariance-based test formulated by Stout (1987) and Nandakumar and Stout (1993), with refinement by Stout, Froelich, and Gao (2001), was used to further evaluate these assumptions. DIMTEST yields a T-statistic that tests the null hypothesis of essential unidimensionality.

Analytic procedure—The steps for conducting IRT-based DIF analyses have been described in great detail elsewhere (e.g., Orlando-Edelen et al, 2006; Teresi et al., 2007). The first step was to identify a group of DIF-free anchor items that could be used to link the two participant groups in terms of their ability (θ). Anchor items were generated in an iterative manner using IRTLRDIF, a software program for IRT modeling based on a nested comparison approach (Thissen, 2001). IRTLRDIF uses maximum likelihood to test a compact model, in which all parameters are constrained to be equal for a studied item, to an augmented model in which one or more parameters are freely estimated. The difference between the log-likelihood statistics of the two models is distributed as a chi-square statistic (G^2), with degrees of freedom equal to the difference in parameter estimates between the two models. Under this iterative procedure, all items are initially evaluated for DIF by sequentially testing the two models. A significant G^2 statistic at the nominal $p < .05$ level indicates that at least one of the parameters differs between the groups and is assumed to demonstrate DIF. After all test items were tested once using this procedure and those with potential DIF removed, the procedure was repeated as many times as necessary until "purification" of the anchor set was achieved. This represented the DIF-free anchor item set for all subsequent DIF analyses.

Next, the remaining set of test items was sequentially evaluated for DIF against the set of purified anchor items. Nonuniform DIF was examined by testing the discrimination parameter (α) between the compact and augmented models for each item, followed by examination of uniform DIF by testing the difficulty parameter (β). To adjust for multiple comparisons, the false discovery rate was controlled using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) as implemented by Thissen et al. (2002) because this method has been shown to demonstrate greater power than the Bonferroni approach. Final item parameter estimates, standard errors, and summary statistics were then estimated using MULTILOG (Thissen, 2003). Finally, the ICCs for each item demonstrating DIF were plotted for visual inspection, and item and test information were calculated to gauge reliability across the spectrum of naming ability.

Within a parametric approach, additional computational models are also available to investigate DIF. In the current investigation, we opted to supplement the IRT-based DIF analyses with a logistic regression approach (Jodoin & Gierl, 2001; Swaminathan & Rogers, 1990; Zumbo, 1999) with the goal of bolstering the detection of DIF by minimizing the likelihood of method-dependent spurious findings. The use of multiple computational procedures in the detection of DIF has been advocated (Hambleton, 2006). Hierarchical logistic regression models were applied to each item using the binary responses as the dependent variable and the BNT total score, grouping variable (African American versus Caucasian), and group-by-total score interaction as the independent variables. The total score was entered into the model in the first step, followed by the grouping variable and finally by the interaction term. This model yields for each item a test for DIF by subtracting the chi-square value in the third step from the chi-square value in the first step. This results in a 2-degree of freedom chi-square test of uniform and nonuniform DIF (Swaminathan & Rogers, 1990). The Benjamini-Hochberg procedure was again used to control the false discovery rate. The effect size criteria proposed by Jodoin and Gierl (2001) was adopted to gauge the magnitude of DIF in the

regression models, with Nagelkerke R^2 values less than .035 representing negligible DIF, .035 to .070 moderate DIF, and greater than .070 large DIF.

Results

Consistent with prior normative studies across racial/ethnic groups, our sample of Caucasian adults obtained a higher mean BNT score than our sample of African American adults ($t(668) = 15.7$, $p < .001$, Cohen's $d = 1.21$). Except for item 51 (“latch”), in which 55% of African American and 54% of Caucasian adults provided a correct response, African Americans had a lower percentage of correct responses across the remaining test items (Figure 2).

Assumptions

The principal components analysis from the African American group showed the first eigenvalue was 14.65 and the second eigenvalue 1.63, a ratio of nearly 9:1. Results from a single-factor CFA showed a CFI of .91 and RMSEA of .09. The DIMTEST statistic was $T = 1.26$ ($p = .10$). These findings were interpreted to provide strong support for the assumption of unidimensionality in this sample. In the Caucasian group, the principal components analysis showed the first eigenvalue was 8.60 and the second 3.50, a ratio of 2.5:1. CFA produced multiple Heywood cases, resulting in ‘not positive definite’ errors and non-admissible solutions despite modifications, likely due to sample size limitations. Results from DIMTEST showed a T statistic of 1.38 ($p = .08$). Despite the greater heterogeneity in the Caucasian data, these findings were interpreted to reflect borderline unidimensionality in this sample and considered sufficient to proceed with IRT modeling.

IRT-based DIF

To identify the anchor item set, an initial run using IRTLRDIF found 15 DIF-free items (“harmonica,” “acorn,” “igloo,” “stilts,” “cactus,” “hammock,” “knocker,” “stethoscope,” “pyramid,” “accordion,” “asparagus,” “compass,” “sphinx,” “yoke,” and “trellis”). During a second iteration, one of these items demonstrated DIF (“pyramid”) and was removed. No additional items were found to demonstrate DIF during a third iteration. Thus, these 14 DIF-free items represented the purified anchor set used to link the two groups. The remaining 17 items (including “pyramid”) were considered the candidate items for DIF analyses.

Using the DIF-free anchor set, 15 of the 17 candidate items initially demonstrated DIF. After controlling for multiple comparisons, 12 of the 17 candidate items continued to demonstrate DIF. “Dominoes” and “escalator” showed DIF in both the α and β parameters and “rhinoceros” showed DIF in the α parameter only (nonuniform DIF). “Muzzle,” “unicorn,” “noose,” “latch,” “tripod,” “scroll,” “tongs,” “palette,” and “protractor” showed DIF in the β parameter only (uniform DIF). Final IRT parameter estimates and standard errors for items demonstrating DIF are presented in Table 2.

Among the 9 items demonstrating uniform DIF, 7 of them showed a difference of at least 0.5 standard deviations in β parameters between the two participant groups. “Tripod” and “palette” showed higher β parameters in African American adults, whereas “muzzle,” “unicorn,” “noose,” “latch,” “scroll,” “tongs,” and “protractor” showed higher β parameters in Caucasian adults. Item characteristic curves for these items are shown in Figure 3.

The 31 studied items covered a wide range of difficulty. For Caucasian adults, difficulty parameters ranged from the least difficult item at $\beta = -5.03$ (“dominoes”) to the most difficult item at $\beta = 1.82$ (“protractor”). In fact, the probability of responding correctly to the item “dominoes” was at minimum 70% across the entire ability spectrum. For African American adults, the least difficult item was “escalator” ($\beta = -2.08$), while “protractor” represented the

most difficult item at β = 1.28. It is notable that "protractor" was by far the most difficult item for both participant samples while the next most difficult item, "compass," had a β parameter that was substantially lower (β = .70).

Range of discriminability parameters was comparable between the two participant samples. The least discriminating item among Caucasian adults was "dominoes" (α = .41), which as noted earlier also represented the easiest item. Among African American adults, the least discriminating item was "latch" (α = .62), which was also the second-least discriminating item among Caucasian adults. The most highly discriminating item in both participant samples was "igloo" (α = 3.46).

Figure 4 displays the test information function for each participant sample. In both groups, the BNT provided the most information at approximately -1.0 standard deviation of naming ability. In other words, precision of measurement was greatest at a mild level of naming difficulty, with slightly greater precision at that ability range for African American than Caucasian adults.

LR-based DIF

Sixteen items had difference chi-square tests that were significant at the nominal $p < .05$ level. After controlling for multiple comparisons, 14 out of these 16 items remained statistically significant (Table 3). DIF effect size estimates showed 8 of these items to demonstrate negligible DIF and 6 items to demonstrate moderate DIF (i.e., "dominoes," "escalator," "muzzle," "latch," "tripod," "palette"). No item demonstrated a large magnitude of DIF in the regression models.

Discussion

The present study sought to investigate differential item functioning (DIF) on the Boston Naming Test between African American and Caucasian adults age 52 and older. Using IRT-based methodology, 12 of the studied items demonstrated DIF, suggesting that the conditional probability of responding correctly to these items differed significantly between the two groups after matching for the latent naming ability. The items "dominoes" and "escalator" showed uniform and nonuniform DIF, which reflects nonequivalence in the difficulty and discriminability parameters between the two groups. The item "rhinoceros" showed DIF in the discriminability parameter only, whereas items "muzzle," "unicorn," "noose," "latch," "tripod," "scroll," "tongs," "palette," and "protractor" showed DIF only in the difficulty parameter.

There has been considerable debate within the DIF literature about the extent to which item parameter estimates, and hence DIF detection, are dependent on the methods used to calculate those parameters. An emerging opinion promotes the use of multiple computational procedures for DIF detection. To minimize the likelihood that the current findings were specific to the use of IRT, we reanalyzed the data using hierarchical logistic regression models with item response as the binary dependent outcome in each model. Results showed 14 items to demonstrate DIF, with 6 of these items considered to have at least "moderate" DIF based upon suggested criteria. These 6 items ("dominoes," "escalator," "muzzle," "latch," "tripod," "palette") were similarly identified as demonstrating DIF through the IRT analyses and represent the strongest evidence for race/ethnicity-based DIF in the Boston Naming Test.

The presence of DIF on the Boston Naming Test is problematic from two broad perspectives. First, it raises some concerns about the construct validity of the test when a construct-irrelevant aspect, namely race or ethnic group membership, is associated with nonequivalence in the conditional probability of obtaining a correct item response. In other words, after matching individuals on naming ability level, at minimum 6 of the 31 items studied using IRT and logistic

regression methods (or 12 of 31 studied items using IRT methods only) demonstrate a significantly different probability that a person will respond correctly, solely as a result of their racial/ethnic group membership. Second, it bolsters the notion that the use of ethnicity-based norms to evaluate the clinical impact of the summed total score may mask psychometric problems that are present at the item level.

The current results present additional psychometric information about the Boston Naming Test that, to our knowledge, has not been previously reported. Specifically, examination of the IRT difficulty and discriminability parameters shows lack of a monotonic relationship among ordered items. Despite administration rules conforming to an incremental administration, each successive item does not necessarily represent a psychometric increase in difficulty when compared to its previous item. In fact, the observed pattern most closely resembles an oscillating profile of increasing and decreasing item difficulty as administration progresses from items 30 through 60. This pattern is evident in both the African American and Caucasian participant samples. Additionally, one would expect uniformity in the degree to which each item differentiates those with better or worse naming ability. However, the discrimination parameters show considerable variability from item to item throughout both groups of adults. Graves et al. (2004) previously applied a 1-parameter Rasch model to a mixed sample of 206 adults (n = 62 considered cognitively normal) in order to develop a short version of the BNT; but unfortunately, item difficulty parameters were not reported. Additional IRT-based investigations certainly appear warranted to better understand the finer psychometric properties of this instrument, with a particular emphasis on item parameters across the full range of naming performance in normal and cognitively impaired populations.

From a practical standpoint, these results could be utilized in a future refinement of the BNT to mitigate item bias or differential functioning. Specifically, the items shown here to be free of DIF could be retained in a future revision of the test, with reordering of those items based upon estimated rather than hypothesized difficulty parameters. Alternatively, a scoring algorithm can be devised and implemented to weight responses from DIF-free items more heavily than responses from DIF-loaded items. Given the widespread use of the BNT within and beyond neuropsychology, and the existing large normative datasets across the developmental span, a new scoring algorithm may seem most practical.

Potential limitations to the current investigation include restricting the range of the BNT to items 30 through 60, restricting the participant sample to cognitively normal older adults, the characterization of the two groups, and possible multidimensionality of the data from Caucasian adults. First, item range restriction was necessary for psychometric reasons. The test was administered using standardized rules that instruct examiners to begin with item 30 and proceed until a basal level of 8 consecutive correct responses is reached. Fewer than one-third of our cognitively normal participant sample failed to reach this basal level. As a result, the majority of items 1-29, and particularly items 1-25, were administered to less than 200 subjects per group. IRT analyses of these items would have resulted in unstable and perhaps misleading parameter estimates.

Our sample was restricted to cognitively normal adults for two reasons. First, the principal goal of the study was to investigate the psychometric properties of the BNT at the item level, with a particular focus on differential functioning between two racial/ethnic groups. The relationship between item parameters and clinical dysfunction is not essential to understanding DIF, as the analytic method presupposes matching on ability level. Furthermore, restricting the sample to cognitively normal adults minimized the likelihood that an unequal distribution between the two groups of clinical factors unrelated to naming (e.g., reduced visual acuity or perception, slowed processing speed) could have contributed to differential performance across items. The

relationship between item parameters and naming dysfunction remains an important topic of investigation and will be pursued in future studies.

The two participant groups consisted of self-identified African American and Caucasian adults that have taken part in the Mayo normative studies. Numerous publications over the past 19 years have described the methodology used to select the participant sample. Test selection, administration, and scoring are comparable between the two Mayo sites in Rochester, Minnesota, and Jacksonville, Florida. Because all of the African American adults were recruited in Jacksonville, it could be reasonably argued that the current results represent DIF based upon geographic (i.e., north vs. south) rather than ethnicity-based factors. This certainly presents a caveat to the findings discussed above as well as a testable hypothesis. Unfortunately, we do not have a sufficiently large sample of African American and Caucasian adults within both sites to examine DIF based on geographic distribution.

Lastly, it is possible that the BNT data obtained from Caucasian adults were not sufficiently unidimensional to meet IRT assumptions, although it seems unclear why ‘naming’ should be a more strongly unitary construct in one group versus the other. To our knowledge the invariance properties of the BNT have yet to be uniformly established across these two groups, and this may represent another topic worthy of further study.

In sum, the current investigation highlights the benefits of modern psychometric methods in the investigation of between-group discrepancies. Our findings suggest that the unexamined use of race/ethnicity-based norms, although necessary in clinical decision-making, potentially masks underlying psychometric problems that may contribute to between-group discrepancies at the item level. The degree to which these findings extend beyond the BNT to other established and commonly used neuropsychological instruments remains largely unexplored.

Acknowledgments

We extend our gratitude to Katja Ocepek-Welikson at the Hebrew Home for the Aged in Riverdale, New York, for technical assistance with the IRTLRDIF program, and to Francine C. Parfitt, Tracy L. Kendall, and Sylvia B. Stewart at the Mayo Memory Disorders Clinic for data preparation. This study was supported in part by a grant from the Robert and Clarice Smith Fellowship Program and a Supplement grant from the National Institute on Aging (AG016574-07S1). The study authors do not have any sources, financial or otherwise, that could result in a conflict of interest pertaining to this manuscript.

References

American Education Research Association, American Psychological Association, and the National Council on Measurement in Education. Standards for educational and psychological testing. Washington, D.C.: American Education Research Association; 1999.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B 1995;(57):289–300.

Boomsma, A.; Hoogland, JJ. The robustness of LISREL modeling revisited. In: Cudeck, R.; du Toit, S.; Sörbom, D., editors. Structural equation models: Present and future. Lincolnwood, IL: Scientific Software International; 2001. p. 139-168.

Boone KB, Victor TL, Wen J, Razani J, Pontón M. The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. Archives of Clinical Neuropsychology 2007;22(3):355–365. [PubMed: 17320344]

Brandt J. 2005 INS Presidential Address: Neuropsychological crimes and misdemeanors. The Clinical Neuropsychologist 2007;21(4):553–568. [PubMed: 17613978]

Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahway, NJ: Lawrence Erlbaum Associates; 2000.

Ferman TJ, Ivnik RJ, Lucas JA. Boston Naming Test discontinuation rule: Rigorous versus lenient interpretations. Assessment 1998;5(1):13–18. [PubMed: 9458337]

Fillenbaum GG, Huber M, Taussig IM. Performance of elderly White and African American community residents on the abbreviated CERAD Boston Naming Test. Journal of Clinical & Experimental Neuropsychology 1997;19(2):204–210. [PubMed: 9240480]

Fillenbaum GG, Peterson B, Welsh-Bohmer KA, Kukull WA, Heyman A. Progression of Alzheimer's disease in black and white patients: the CERAD experience, part XVI. Consortium to Establish a Registry for Alzheimer's Disease. Neurology 1998;51(1):154–158. [PubMed: 9674795]

Graves RE, Bezeau SC, Fogarty J, Blair R. Boston naming test short forms: A comparison of previous forms with new item response theory based forms. Journal of Clinical and Experimental Neuropsychology 2004;26(7):891–902. [PubMed: 15742540]

Hambleton RK. Good practices for identifying differential item functioning. Medical Care 2006;44(Suppl 3):S182–S188. [PubMed: 17060826]

Hambleton, RK.; Swaminathan, H. Item response theory Principles and applications. Boston: Kluwer-Nijhoff Publishing; 1985.

Hambleton, RK.; Swaminathan, H.; Rogers, HJ. Fundamentals of item response theory. Newbury Park, CA: Sage Publications; 1991.

Heaton, RK.; Miller, SW.; Taylor, MJ.; Grant, I. Revised comprehensive norms for an expanded Halstead-Reitan battery: Demographically-adjusted neuropsychological norms for African American and Caucasian adults. Lutz, FL: Psychological Assessment Resources, Inc.; 2004.

Holman R, Glas CAW, de Haan RJ. Power analysis in randomized clinical trials based on item response theory. Controlled Clinical Trials 2003;24:390–410. [PubMed: 12865034]

Inouye SK, Albert MS, Mohs R, Sun K, Berkman LF. Cognitive performance in a high functioning community dwelling elderly population. Journal of Gerontology: Medical Sciences 1993;48:M146–M151.

Ivnik RJ, Malec JF, Tangalos EG, Petersen RC, Kokmen E, Kurland LT. The Auditory Verbal Learning Test (AVLT): Norms of ages 55 and older. Psychological Assessment 1990;2:304–312.

Jodoin MG, Gierl MJ. Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. Applied Measurement in Education 2001;14:329–349.

Jöreskog, KG.; Sörbom, D. LISREL 8.80. Chicago, IL: Scientific Software International; 2006.

Jöreskog, KG.; Sörbom, D. LISREL 8: User's reference guide. 2nd. Chicago, IL: Scientific Software International; 1997.

Kaplan, E.; Goodglass, H.; Weintraub, S. The Boston Naming Test. Philadelphia: Lea & Febiger; 1983.

Lichtenberg PA, Ross T, Christensen B. Preliminary normative data on the Boston Naming Test for an older urban population. Clinical Neuropsychologist 1994;8:109–111.

Lucas JA, Ivnik RJ, Smith GE, Ferman TJ, Willis FB, Petersen RC, Graff-Radford NR. Mayo's Older African American Normative Studies: Normative data for commonly used clinical neuropsychological measures. The Clinical Neuropsychologist 2005a;19:162–183. [PubMed: 16019702]

Lucas JA, Ivnik RJ, Smith GE, Ferman TJ, Willis FB, Petersen RC, Graff-Radford NR. Mayo's Older African American Normative Studies: Norms for Boston Naming Test, Controlled Oral Word Association, Category Fluency, Animal Naming, Token Test, WRAT-3 Reading, Trail Making Test, Stroop Test, and Judgment of Line Orientation. The Clinical Neuropsychologist 2005b;19:243–269. [PubMed: 16019707]

Manly JJ. Advantages and disadvantages of separate norms for African Americans. Clinical Neuropsychologist 2005;19(2):270–275. [PubMed: 16019708]

Manly JJ, Byrd DA, Touradji P, Stern Y. Acculturation, reading level, and neuropsychological test performance among African American elders. Applied Neuropsychology 2004;11(1):37–46. [PubMed: 15471745]

Manly JJ, Jacobs DM, Sano M, Bell K, Merchant CA, Small SA, Stern Y. Cognitive test performance among nondemented elderly African Americans and whites. Neurology 1998;50(5):1238–1245. [PubMed: 9595969]

Manly JJ, Jacobs DM, Touradji P, Small SA, Stern Y. Reading level attenuates differences in neuropsychological test performance between African American and White elders. Journal of the International Neuropsychological Society 2002;8:341–348. [PubMed: 11939693]

Miles, GT. Neuropsychological assessment of African Americans. In: Ferraro, FR., editor. Studies on neuropsychology, development, and cognition. Lisse, Netherlands: Swets & Zeitlinger Publishers; 2002. p. 63-77.

Muthén B, Kaplan D. A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. British Journal of Mathematical and Statistical Psychology 1992;45:19–30.

Nandakumar R, Stout W. Refinements of Stout's procedure for assessing latent trait unidimensionality. Journal of Educational Statistics 1993;18:41–68.

Orlando-Edelen MO, Thissen D, Teresi JA, Kleinman M, Ocepek-Welikson K. Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental State Examination. Medical Care 2006;44(Suppl 3):S134–S142. [PubMed: 17060820]

Pedraza O, Mungas D. Measurement in cross-cultural neuropsychology. Neuropsychology Review 2008;18(3):184–193. [PubMed: 18814034]

Rilling LM, Lucas JA, Ivnik RJ, Smith GE, Willis FB, Ferman TJ, Petersen RC, Graff-Radford NR. Mayo's Older African American Normative Studies: Norms for the Mattis Dementia Rating Scale. The Clinical Neuropsychologist 2005;19:229–242. [PubMed: 16019706]

Smith, GE.; Ivnik, RJ.; Lucas, JA. Assessment techniques: Tests, test batteries, norms, and methodological approaches. In: Morgan, J.; Ricker, J., editors. Textbook of Clinical Neuropsychology. New York: Taylor & Francis; Group: 2008. p. 38-57.

Stout W. A nonparametric approach for assessing latent trait unidimensionality. Psychometrika 1987;52 (4):589–617.

Stout, W.; Froelich, A.; Gao, F. Using resampling methods to produce an improved DIMTEST procedure. In: Boomsma, A.; van Duijn, MAJ.; Snijders, TAB., editors. Essays on item response theory. New York: Springer-Verlag; 2001. p. 357-376.

Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement 1990;26:361–270.

Teresi JA. Different approaches to differential item functioning in health applications: Advantages, disadvantages and some neglected topics. Medical Care 2006;44(11 Suppl 3):S152–S170. [PubMed: 17060822]

Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. Statistics in Medicine 2000;19:1651–1683. [PubMed: 10844726]

Teresi JA, Kleinman M, Ocepek-Welikson K, Ramirez M, Gurland B, Lantigua R, Holmes D. Applications of item response theory to the examination of the psychometric properties and differential item functioning of the Comprehensive Assessment and Referral Evaluation Dementia Diagnostic Scale among samples of Latino, African American, and White non-Latino elderly. Research on Aging 2000;22(6):738–773.

Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, Morales LS, Orlando-Edelen M, Cella D. Evaluating measurement equivalence using the item response theory log-likelihood ration (IRTLR) method to assess differential item functioning (DIF): Applications (with illustrations) to measures of physical functioning ability and general distress. Quality of Life Research 2007;16(Suppl 1):43–68. [PubMed: 17484039]

Thissen, D. IRTLRDIF v2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. L.L. Thurstone Psychometric Laboratory, University of North Carolina at Chapel Hill; 2001.

Thissen, D. MULTILOG 7.0: Multiple, categorical item analysis and test scoring using item response theory. Chicago: Scientific Software International; 2003.

Thissen D, Steinberg L, Kuang D. Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. Journal of Educational and Behavioral Statistics 2002;27(1):77–83.

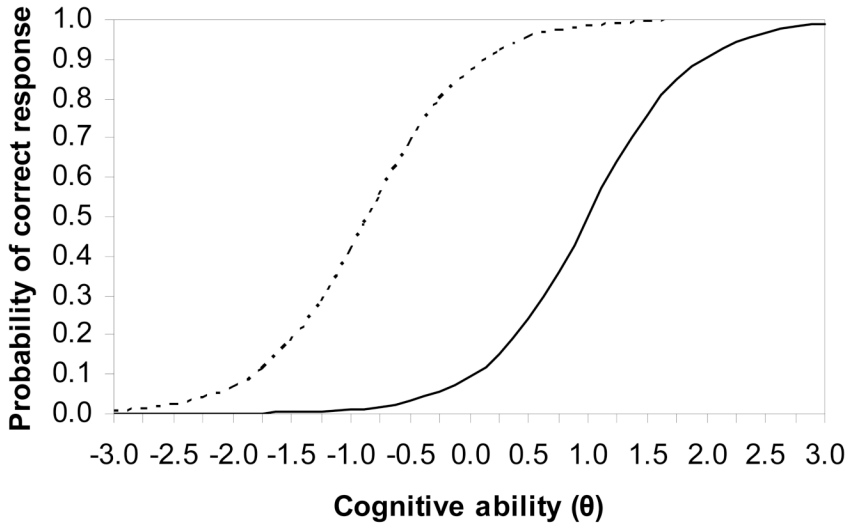
Wagner MT, Wymer JH, Carlozzi NE, Bachman D, Walker A, Mintzer J. Alzheimer Study Group. Preliminary examination of progression of Alzheimer's disease in a rural Southern African American cohort. Archives of Clinical Neuropsychology 2007;22(3):405–414. [PubMed: 17296283]

Welsh K, Fillenbaum G, Wilkinson W, Heyman A, Mohs RC, Stern Y, Harrel Z, Edland S, Beekly D. Neuropsychological test performance in African American and White patients with Alzheimer's disease. Neurology 1995;45:2207–2211. [PubMed: 8848195]

Whitfield KE, Fillenbaum GG, Pieper C, Albert MS, Berkman LF, Blazer DG, Rowe JW, Seeman T. The effect of race and health-related factors on naming and memory. The MacArthur Studies of Successful Aging. Journal of Aging & Health 2000;12(1):69–89. [PubMed: 10848126]

Zumbo, BD. A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Directorate of Human Resources Research and Evaluation, Department of National Defense; Ottawa, Canada: 1999. Retrieved on 6/8/2005 from http://educ.ubc.ca/faculty/zumbo/DIF/index.html

a.



b.

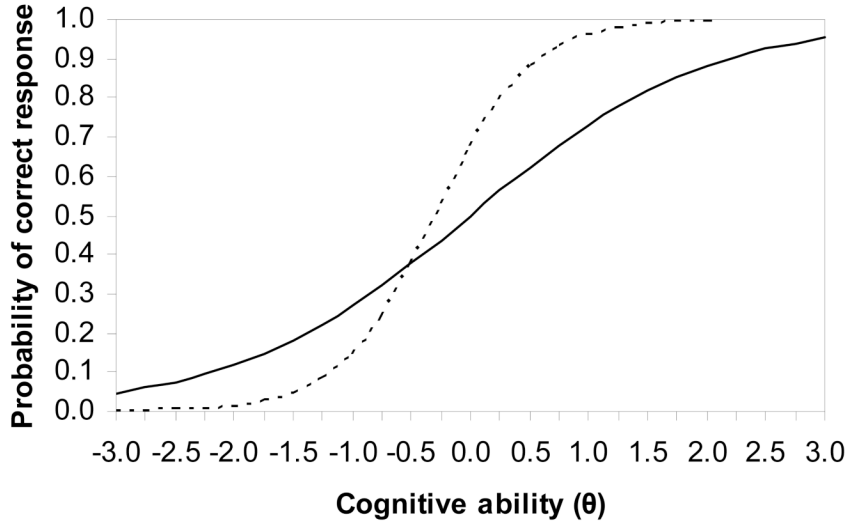


Figure 1. Sample item demonstrating uniform DIF (a) and nonuniform DIF (b)

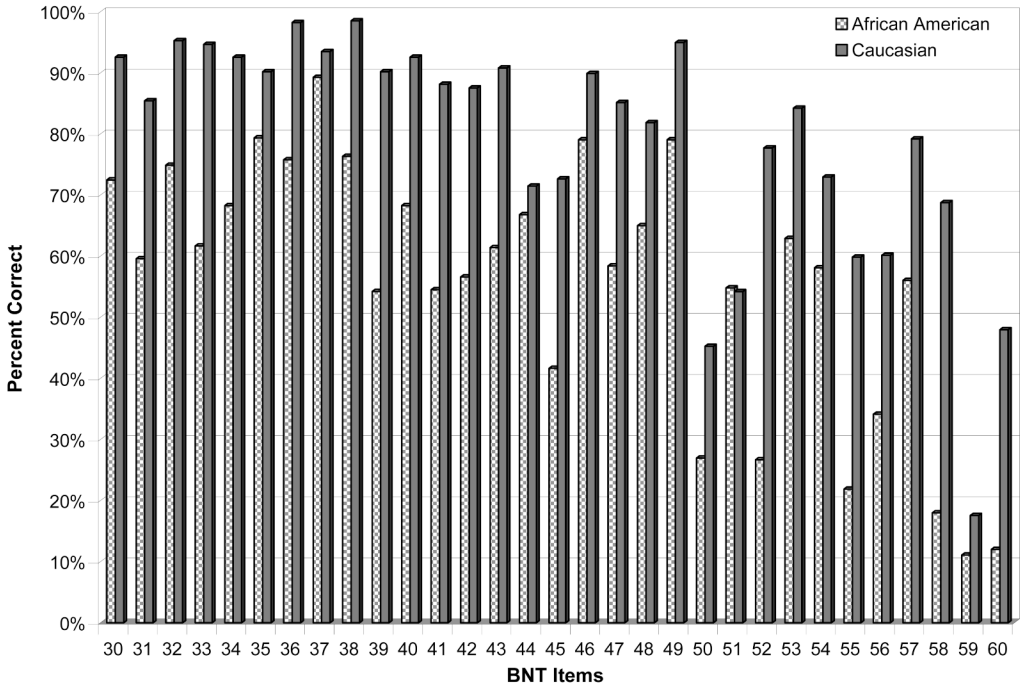


Figure 2. Mean percent of correct item responses by participant group

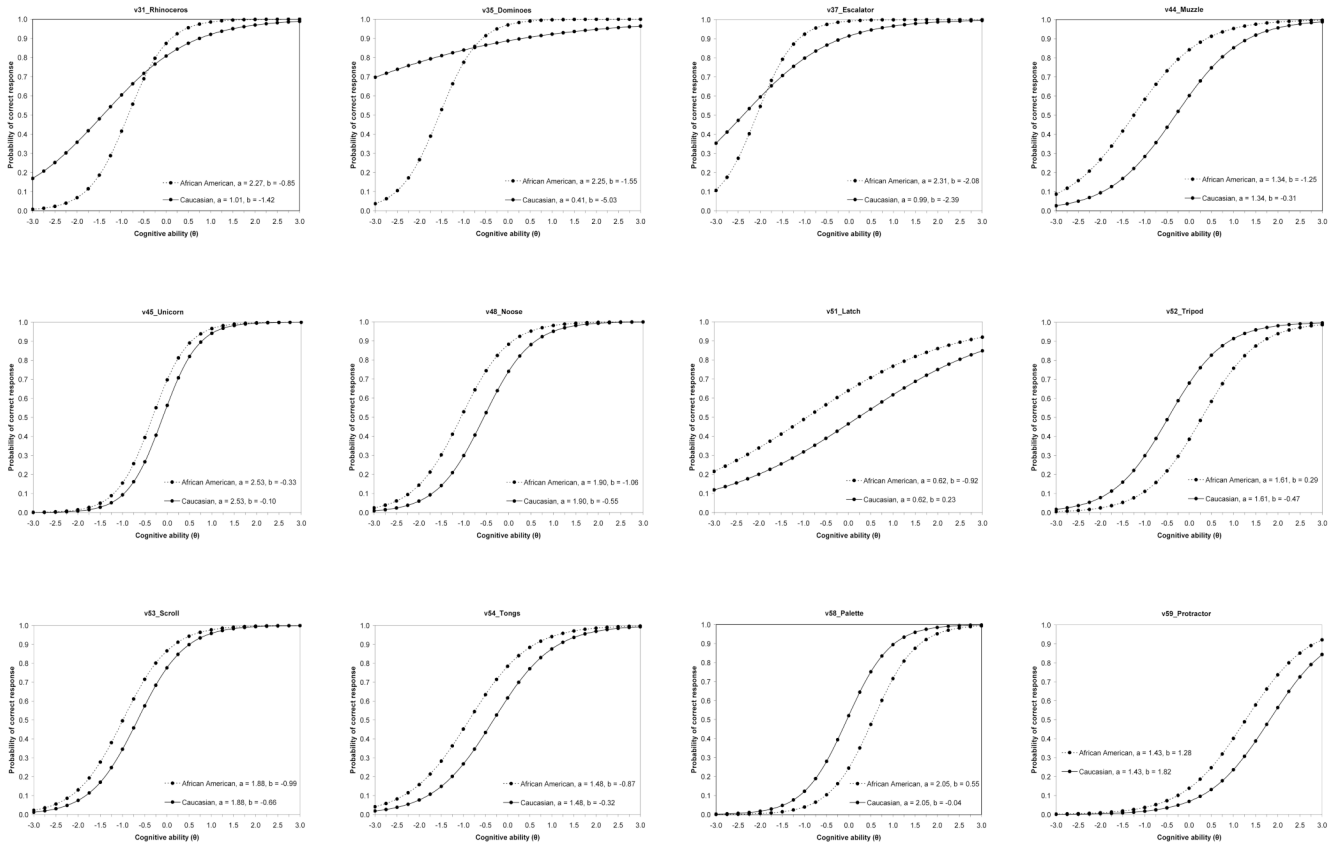


Figure 3. Item characteristic curves for 12 items demonstrating DIF using IRT-based methods

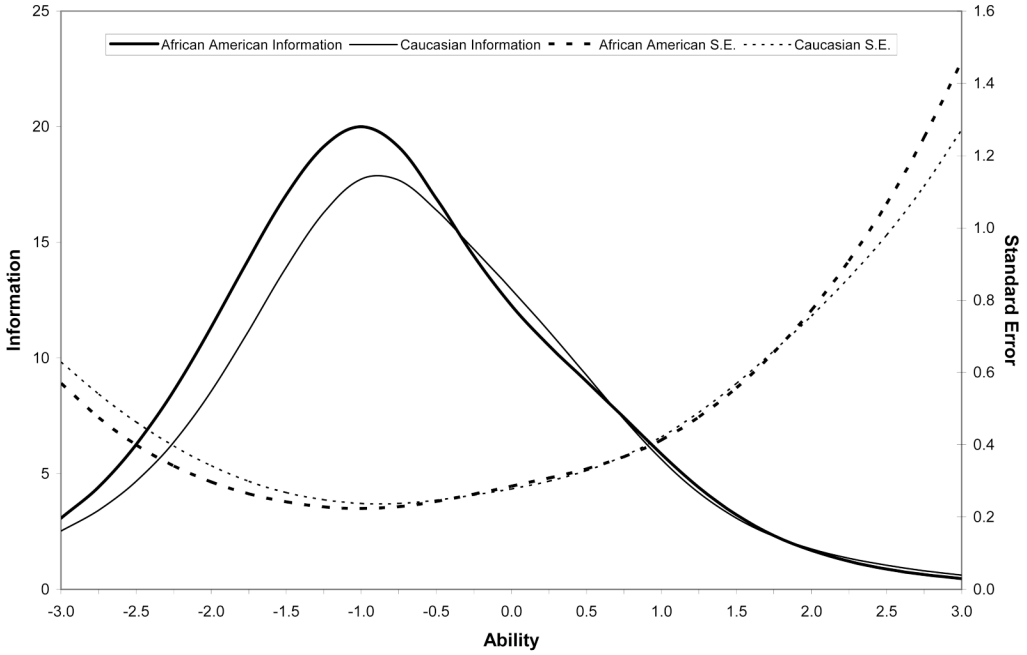


Figure 4. Test information curve by participant group across range of naming ability

Table 1

Demographic characteristics and Boston Naming Test (BNT) mean scores

		Sex	Age			Education			BNT		
	n	(% female)	M	SD	Range	M	SD	Range	M	SD	Range
Caucasian	336	70.2	75.3	9.2	52-98	14.0	2.7	4-20	52.9	4.9	31-60
African American	334	75.1	70.3	7.5	52-95	12.3	3.5	0-20	43.3	10.1	16-60

Table 2

Item parameters, standard errors, and type of IRT-based DIF

		Caucasian		African American				
Item #	**Name**	**α**	**β**	**α**	**β**	**α-DIF Test: G^2 statistic**	**β-DIF Test: G^2 statistic**	**Type of DIF**
31	Rhinoceros	1.01 (0.35)	-1.42 (0.61)	2.27 (0.37)	-0.85 (0.09)	10.2 **	3.2	NU
35	Dominoes	0.41 (0.39)	-5.03 (5.23)	2.25 (0.44)	-1.55 (0.12)	16.6 ***	10.9 **	NU
37	Escalator	0.99 (0.54)	-2.39 (1.38)	2.31 (0.58)	-2.08 (0.22)	7.1 **	19.0 ***	NU
44	Muzzle	1.34 (0.16)	-0.31 (0.14)	1.34 (0.16)	-1.25 (0.14)	0.0	32.1 ***	U
45	Unicorn	2.53 (0.27)	-0.10 (0.08)	2.53 (0.27)	-0.33 (0.08)	0.2	10.5 **	U
48	Noose	1.90 (0.19)	-0.55 (0.11)	1.90 (0.19)	-1.06 (0.11)	0.2	16.6 ***	U
51	Latch	0.62 (0.13)	0.23 (0.25)	0.62 (0.13)	-0.92 (0.26)	0.0	14.5 ***	U
52	Tripod	1.61 (0.17)	-0.47 (0.12)	1.61 (0.17)	0.29 (0.13)	4.2 *, 1	17.7 ***	U
53	Scroll	1.88 (0.19)	-0.66 (0.12)	1.88 (0.19)	-0.99 (0.11)	2.3	8.5 **	U
54	Tongs	1.48 (0.18)	-0.32 (0.12)	1.48 (0.18)	-0.87 (0.12)	2.4	13.8 ***	U
58	Palette	2.05 (0.22)	-0.04 (0.09)	2.05 (0.22)	0.55 (0.12)	0.8	14.0 ***	U
59	Protractor	1.43 (0.13)	1.82 (0.15)	1.43 (0.13)	1.28 (0.18)	0.0	9.2 **	U

* $p < .05$,

** $p < .01$,

*** $p < .001$.

α = discriminability parameter, β = difficulty parameter; G^2 = chi-square with 1 d.f., U = uniform DIF, NU = nonuniform DIF.

1 DIF test not statistically significant after Benjamini-Hochberg procedure.

Table 3

DIF results using hierarchical logistic regression

		χ^2			R^2		
Item #	**Name**	**Step 1**	**Step 3**	**Difference**	**Step 1**	**Step 3**	**Difference**
35	Dominoes	138.59	156.67	18.08	0.326	0.363	0.037
37	Escalator	83.13	103.85	20.72	0.262	0.323	0.061
38	Harp	219.89	228.91	9.02	0.528	0.546	0.018
39	Hammock	246.20	260.31	14.11	0.444	0.464	0.020
41	Pelican	196.86	212.92	16.06	0.365	0.390	0.025
44	Muzzle	110.99	150.76	39.77	0.215	0.284	0.069
45	Unicorn	351.79	362.30	10.51	0.548	0.561	0.013
48	Noose	201.23	215.02	13.79	0.378	0.400	0.022
51	Latch	45.16	68.45	23.29	0.087	0.130	0.043
52	Tripod	293.81	337.52	43.71	0.474	0.528	0.054
54	Tongs	163.15	177.76	14.61	0.298	0.322	0.024
58	Palette	328.29	370.83	42.54	0.520	0.570	0.050
59	Protractor	98.39	107.52	9.13	0.244	0.265	0.021
60	Abacus	285.20	299.31	14.11	0.492	0.510	0.018