Evaluating the Influence of Differential Item Functioning for Race and Gender on STAR Reading Items

Damien C. Cormier

Written for Renaissance Learning Inc.

Renaissance Learning
PO Box 8036
Wisconsin Rapids, WI 54495-8036
Phone: (800) 338-4204
Outside the US: 1.715.424.3636
Fax: (715) 424-4242
Email: answers@renlearn.com
Support Email: support@renlearn.com
Web Site: www.renlearn.com

Copyright Notice

Copyright © 2011-2012 by Renaissance Learning, Inc. All Rights Reserved.

This publication is protected by US and international copyright laws. It is unlawful to duplicate or reproduce any copyrighted material without authorization from the copyright holder. This document may be reproduced only by staff members in schools that have a license for STAR Reading software. For more information, contact Renaissance Learning, Inc., at the address above.

Accelerated Math, Accelerated Reader, AR, ATOS, English in a Flash, MathFacts in a Flash, NEO, Renaissance, Renaissance Learning, the Renaissance logo, Renaissance Place, STAR Early Literacy, STAR Reading, STAR Reading, and the STAR logo are trademarks of Renaissance Learning, Inc., and its subsidiaries, registered, common law, or pending registration in the US and other countries.

CONTENTS

Executive Summary 3
Introduction 10
Star Reading 10
Design 10
Test Administration Procedures 11
Content Development 12
Validity and Differential Item Functioning 13
Uniform and Non-Uniform DIF 15
Purpose 18
Method 18
Sample 19
Variables 20
Item Correct 20
Gender 21
Race 21
STAR Reading Scaled Score 22
Assessment Items 22
Procedure 22
Results 30
Informal DIF 30
Mantel-Haenszel Procedure 31
Logistic Regression 32
Limitations 36
Conclusion 36
References 39
Appendix A: Asian and Native American DIF Analyses 41
Procedure 42
Informal DIF Results 42
Logistic Regression 44
Conclusion 45
Appendix B 46

Appendix C.....56
Appendix D.....61
Appendix E.....71
Logistic Regression DIF Syntax for SPSS v.19.....72
Gender Analysis.....72
Race Analysis.....72

EXECUTIVE SUMMARY

Modern test theory views validity as an ongoing process that is assumed to never be complete (Messick, 1995). The collection of validity evidence in terms of "consequences of test use, measurement error issues, and the context of use" (Zumbo, 1999, p. 11) is part of the process that is part of responsible test development. It has been suggested that these questions need to be answered if we are to appropriately justify the use and interpretation of assessment results (Messick, 1995a). The evaluation of differential item functioning (DIF) is an important component of the various pieces of information that are needed to build a strong validity argument for a given assessment. If an item demonstrates bias toward one subgroup, then the subgroup that is disadvantaged by this item may be more likely to be subjected to inappropriate decisions based on their assessment results. Thus, the evaluation of the influence of DIF is essential to determine item bias.

The purpose of the research reported here was to systematically evaluate items from the STAR Reading assessment to determine whether DIF exists for gender and race subgroups. Initial DIF analyses evaluated DIF for gender, and for white, black, and Hispanic subgroups; a second phase of analysis, reported in Appendix A, evaluated it in smaller samples of Asian

and Native American subgroups, . The evaluation was conducted in three phases, which began by considering all of the assessment items, as recommended by Osterlind and Everson (2009). This comprehensive evaluation of DIF for STAR Reading items consists of (a) an informal DIF procedure; (b) the Mantel-Haenszel procedure; and (c) a logistic regression procedure. In addition, this report also provides adequate information on the procedures used and described herein, so that they can be applied to future analyses and contexts. This is particularly important, given the need for the collection of ongoing validity evidence of evolving assessment procedures.

Methods

[Star Reading uncalibrated items from Batches RS5 and RS6 were used for analysis. The batches were combined and went live around Jan. 4th, 2011. A total of 17,050,812 cases were included in the original file. The number of cases (item responses) that had demographic information on gender was equal to 11,596,983 (68.0% of the total number of cases) and the number of cases that had demographic information on race was equal to 4,845,736 (28.4% of the total number of cases). Summary demographic information by cases and for individual assessment items are presented in Table 1. The total number of items contained in the merged data file was 1,222. As seen in Table 1, every assessment item contained in the data file had at least a number of cases with information on gender and race. Thus, the analyses were able to be conducted on the entire set of items available within these STAR Reading batches.

]

Items. The items of interest here were 1,222 new items written for use in STAR Reading Enterprise, a new version of STAR Reading designed as a standards-referenced reading achievement test, STAR Reading Enterprise items measure 5 major content domains, 10

general skill sets, and more than 300 discrete skills. (Not all of those skills were represented in the data set reported here.) Each of these new items was embedded, along with 2 to 4 others, as unscored items within STAR Reading tests administered nationwide.

Tests. Students took STAR Reading tests in the normal course of their reading studies. Each STAR Reading test consisted of 25 adaptively administered scored test items, with another 3 to 5 randomly selected, unscored items embedded in the STAR Reading test. Student scores on the STAR Reading tests served as the criterion measure for the DIF analyses; the new items were not counted in calculating the STAR Reading scores.

Data. A total of 17,050,812 cases (item response records) were included in the original data file, Gender information was recorded for 11,596,983 (68.0%) of those cases; demographic information on race was recorded for 4,845,736 (28.4%).

Data analyses. The purpose of the analyses was to systematically evaluate the prospective new STAR Reading items to determine whether DIF existed for gender and race subgroups. The evaluation was conducted using an approach recommended by Osterlind and Everson (2009). It consisted of three phases: (a) an informal DIF procedure, to identify items with the largest differences between reference and focal subgroups; (b) the Mantel-Haenszel procedure, to identify items with "uniform DIF); and (c) a logistic regression procedure, to determine whether items showed evidence of non-uniform DIF.

Results of Initial Analyses

#

<u>The initial informal DIF analysis, which was limited to subgroups with large student samples (males vs. females, and white-black-Hispanic students)</u> included all of the 1,222 items from

the STAR Reading dataset, and looked at subgroups differences in proportion correct ("p-value") on each item, Overall, item p-value difference statistics based on race comparisons demonstrated greater differences between groups than those seen for the gender comparison. Actual differences ranged from very small (less than 0.001) to as large large as 0.25 between groups. As planned, the 100 items with the largest p-value difference statistics were used for the remaining analyses.

The Mantel-Haenszel procedure wasinitially completed for three comparisons: White-Black, White-Hispanic, and Male-Female. Later, the analyses were extended to include White-Asian and white-Native American comparisons; these are reported separately in Appendix A. Students were grouped into four quartiles, based on their STAR Reading scale scores, to control for ability. Overall, there were items identified as demonstrating significant DIF across all ability levels for all comparison groups. There is, however, a considerable decrease in the number of items demonstrating DIF associated with an increase in the ability quartiles. This trend was consistent across the comparisons, with appreciably more items identified as demonstrating DIF in the first quartile compared to the fourth ability quartiles for the gender comparison. That is, the M-H procedure tended to identify DIF more often in the lowest ability groups than in the higher quartile groups. These results supported further exploration of these data in order to gather more information on the interaction between ability and performance on the assessment items to determine if the significant differences seen here are true DIF or if they can in fact be attributed to variations in ability occurring within the arbitrary quartiles that were selected for the M-H analysis.

Logistic regression analyses produced a great deal of data, including chi-square values as indicators of significant differences, and Nagelkerke R-squared values as measures of effect size. First, chi-square values and Nagelkerke R-square values were obtained for three models: An ability

model, an ability and uniform DIF model, and an ability, uniform and non-uniform DIF model. Initial analyses of the logistic regression data suggested that a portion of the items for the White-Black and White-Hispanic comparisons did demonstrate statistically significant uniform and non-uniform DIF, and that for the Gender comparisons, all items demonstrated both uniform and non-uniform DIF. However, analyses of effect sizes, using the Nagelkerke R-squared values, showed them to be very small: smaller than .01 for more than 50% of the items, and smaller than .025 for almost all of the items. Furthermore, the contribution of non-uniform DIF was less than .001 for the vast majority of the items.

Because only Nagelkerke values above 0.05 are considered to have practical significance, it can be concluded that once ability is adequately controlled for statistically, DIF between race or gender groups is reduced to a negligible value. In other words, a student's race or gender was not related to the probability of a student being able to correctly answer any of the items analyzed. Thus, this final analysis has demonstrated that the items analyzed do not demonstrate a noteworthy bias toward a particular gender or race.

Results of Small Subgroup Analyses

The results of these analyses were consistent with those seen in the large subgroup comparisons (e.g. White-Black and White-Hispanic). In brief, once ability was controlled for statistically, any DIF that was observed in the informal DIF procedure was non-significant, or effect sizes were negligible. Therefore, despite the analyses being conducted on only a small subset of items that met the sample size threshold, the results of the analyses suggest that no DIF is occurring when comparing the performance of Asian or Native American students to White students. In summary, the results of these analyses further support the use of the

STAR Reading and STAR Reading assessments with students from the culturally and linguistically diverse backgrounds represented in this study.

Interpretation

Based on the logistic regression procedure, and particularly on the Nagelkerke R-squared analyses, it can be concluded that once ability was adequately controlled for statistically, DIF between race or gender groups – including whites, blacks, Hispanics, Asians, and Native Americans -- was reduced to a negligible value. In other words, neither a student's race nor gender was related to the probability of a student being able to correctly answer any of the items analyzed. Thus, this final analysis demonstrated that the items analyzed do not exhibit bias toward a particular gender or race.

Action

Because the DIF analyses reported here resulted in no meaningful DIF when the effect of ability level was properly controlled for by means of logistic regression, no remedial action to any of the 1,222 items studies was needed. However, analyses of future sets of new items may well find some DIF that is meaningful in magnitude. In such cases, items that demonstrate DIF will be withheld from operational use, and subjected to editorial review to try to identify what features or content may have caused differential functioning. In cases where the cause is not apparent, the items will be rejected, and not used in operational STAR tests. Where the cause is suspected or apparent, such items will be either be rejected permanently, or revised, recalibrated, and evaluated anew for DIF. In either event, no items found to have DIF in appreciable magnitude will be employed in the operational item banks of the STAR tests.

INTRODUCTION

STAR READING

STAR Reading is a computer-adaptive test assessment designed to inform instructional practices. The assessment can be considered within a 3-tier system ranging from formative classroom assessments, to interim periodic assessments, and finally summative assessments. STAR Reading is appropriate for students who have acquired at least 100 words of sight vocabulary. STAR reading assessments have three purposes and they are: (a) providing educators with a relatively quick measure of reading comprehension; (b) providing a measure of reading achievement relative to a national normative sample; and (c) providing data that allows for progress monitoring of student performance across time. For more about the advantages of these purposes see the STAR Reading Technical Manual (Renaissance Learning Inc.; RLI, 2012.

DESIGN

STAR Reading is a computer-adaptive test, which uses the information from previous items to select items that are at an appropriate level of ability for each individual student. This key advantage of computer adaptive testing allows for a high degree of psychometric reliability . In addition, the selection of items within a proximal range of the student's ability reduces the likelihood of frustration due to items being too hard or disengagement because items are too easy for a given student. Moreover, only exposing the student to items at his or her ability level results in a decrease in the overall time required to obtain an accurate reading score

because there is no need for the student to respond to items that are not going to provide information about his or her ability level.

STAR Reading assessments take less than 15 minutes to administer and have the ability of providing teachers with meaningful information to guide their instructional practices with individual students. The STAR Reading assessment items are pulled from a large pool of items. Specifically, the Renaissance Place STAR Reading version 4.3 has 1,792 items that have been grouped into 54 difficulty levels. For more about the design, improvements over time in STAR Reading software, and test security procedures, see the STAR Reading Technical Manual (RLI, 2012).

Test Administration Procedures

The STAR Reading assessments are standardized – meaning that the administration of the assessments is conducted in the same way every time a student takes an assessment. The STAR Reading software includes a test monitoring feature, which helps teachers maintain a standardized testing situation every time an assessment is administered to a student.

The user interface was designed to be "simple and effective" (RLI, 2012). Students are permitted to have a practice session, which allows them to become familiar with the software's environment and features. Although there should be no difference between beginner, novice, or expert computer users in their ability to effectively use the STAR Reading software, there is a feature built-in to the practice sessions that ensure that a student struggling to use the software will be able to request support from his or her teacher during the assessment procedure.

During the administration, the STAR Reading software uses adaptive branching; beginning with items of low difficulty relative to the students expected level of ability. The test

administration involves 25 items and most students are expected to complete the assessment in less than 10 minutes (RLI, 2012).

Content Development

The content of the assessment items contained in the STAR Reading program was developed for students from kindergarten through grade 12. This means that the difficulty level of the items ranged from a kindergarten level to beyond high school, so that students who have exceeded the norm are unlikely to reach a ceiling when their reading achievement is assessed. In addition to this difficulty range, the STAR Reading items were also developed to take into consideration the practical implication that students may be tested up to five times per year (and as often as weekly, in some cases) using the software. Thus, in order to reduce the likelihood of students being exposed to items more than once, a large item bank was developed across all grade levels. A vocabulary list graded with the Advantage/TASA Open Standard for Readability (ATOS) software was instrumental in developing the assessment items. For specific information on the content development process, including how ATOS was used to grade vocabulary word lists, see the STAR Reading Technical Manual (RLI, 2012). For more information on the ATOS formula and its development, see Milone (2008). The STAR Reading Technical Manual (RLI, 2012) also provides information on item development, including vocabulary-in-context and authentic text passage item specifications.

The STAR Reading Technical Manual also includes detailed information on item and scale calibration, as well as norming. These are mentioned here because they may be useful to a reader interested in a better understanding of the STAR Reading assessments. However, the item and scale calibration are outside the scope of this report – the analyses described in detail later are conducted on uncalibrated items that were embedded in STAR Reading for the purpose of gathering data for item analysis and calibration. The only influence that these

details have on the analyses in this report is that the scaled scores from the assessment are used as a measure of a student's overall ability in reading, which is described in more detail in the procedures section.

VALIDITY AND DIFFERENTIAL ITEM FUNCTIONING

The importance of gathering validity evidence is often stressed as fundamental to test development to support appropriate use and interpretation of assessment results (Osterlind & Everson, 2009). Messick (1995a) emphasized that validity is not attributable to an assessment, but rather to the interpretation that is made from the assessment's score. This includes the consideration that "scores are a function of not only the items or stimulus conditions, but also of the persons responding as well as the context of the assessment" (p.1).

A number of internal factors have been listed as important considerations in systematically evaluating the content validity of an assessment (Lissitz & Samuelsen, 2007). One of these, reliability, is a criterion for validity and involves consistency in the results of an assessment procedure. Therefore, reliability is sure to have significant implications for the validity of the inferences drawn from the test results, which is why it is often discussed concurrently with validity. Within this framework, the examination of whether there are differences in performance between students from different racial groups or genders is considered to be a source of evidence in the reliability of a test, when ability has been controlled for statistically (Lissitz & Samuelsen, 2007). If there were to be differences in performance between subgroups of students for an item, the item would be considered to be biased due to the significant difference in the probability of students from one group being able to correctly answer the item.

Typically, the examination of individual test items for test bias involves the criterion of probability of success on the item. The probability for success represents the likelihood of a group of students with the same characteristic, such as race or gender, correctly answering the item. A test item is considered to be unbiased when "the probability for success on the item is the same for equally able examinees of the same population regardless of their subgroup group membership" (Osterlind, 1983). It is important to note that although mean subgroup differences in total score do not necessarily indicate that any bias exists within a test. In other words, it is inappropriate to draw the inference the group level differences in mean test score are caused by item bias. There are a multitude of variables that could be influencing the group level differences and without empirical evidence any inference drawn would be purely speculative. Thus, it is important to systematically examine individual items to determine if there is a bias between subgroups of a given population in their probabilities of correctly answering, while controlling for differing levels of ability that may exist within the sample. This type of analysis is referred to as differential item functioning (DIF).

It is not surprising that the definition of DIF is closely related to the aforementioned definition of probability of success –DIF, as defined in the Standards for Educational and Psychological Testing exists "when examinees of equal ability differ on average, according to their group membership, in their responses to a particular item." (APA, NCME, AERA, 1999). This is framed as "fairness as a lack of bias" in the Standards for Educational and Psychological Testing (APA, NCME, AERA, 1999, p.74). Messick (1995b) supports the use of DIF as a method of identifying items that demonstrate construct-irrelevant difficulties between groups that are not related to differences in construct-relevant abilities demonstrated between groups.

Although the evaluation of the influence of DIF is important in gathering validity information for all assessments, Zwick (2000) suggests that it may be particularly important for those such as STAR Reading, which rely on computer-adaptive methods:

> With sometimes more of fewer items, the worth of each item's contribution to the developed score is heightened when compared with tests in which all items are presented to every examinee. Because of this, examining for DIF may be more important in CAT than in nonadaptive tests, and any flawed item may be more consequential to the score. (p.68)

In short, the evaluation of DIF for the STAR Reading items would not only strengthen the validity argument of the assessment, but will also help determine if certain items should be revised or removed from the assessment that is currently in place. Furthermore, the importance of conducting DIF analyses on STAR Reading items appears to have significant implications for this assessment, given its computer-adaptive delivery system.

Uniform and Non-Uniform DIF

There are typically two types of DIF that are evaluated using modern DIF detection procedures. An item that is free of either type of DIF has item characteristic curves that are nearly identical for both groups being compared (see Figure 1). *Uniform DIF* occurs when differential item functioning is occurring in a consistent way across ability levels (see Figure 2). Conversely, n*on-uniform DIF* occurs when the influence of DIF changes depending on the ability level. An extreme example of non-uniform DIF would be for the probability of a correct response on a given item to be greater for the reference group than it is for the focal group at a low ability level, but at a high ability level the probability of a correct response is greater for the focal than it is for the reference group (see Figure 3). For more information on the

differences between uniform and non-uniform DIF and item characteristic curves, see Zumbo (1999) or Osterlind and Everson (2009).

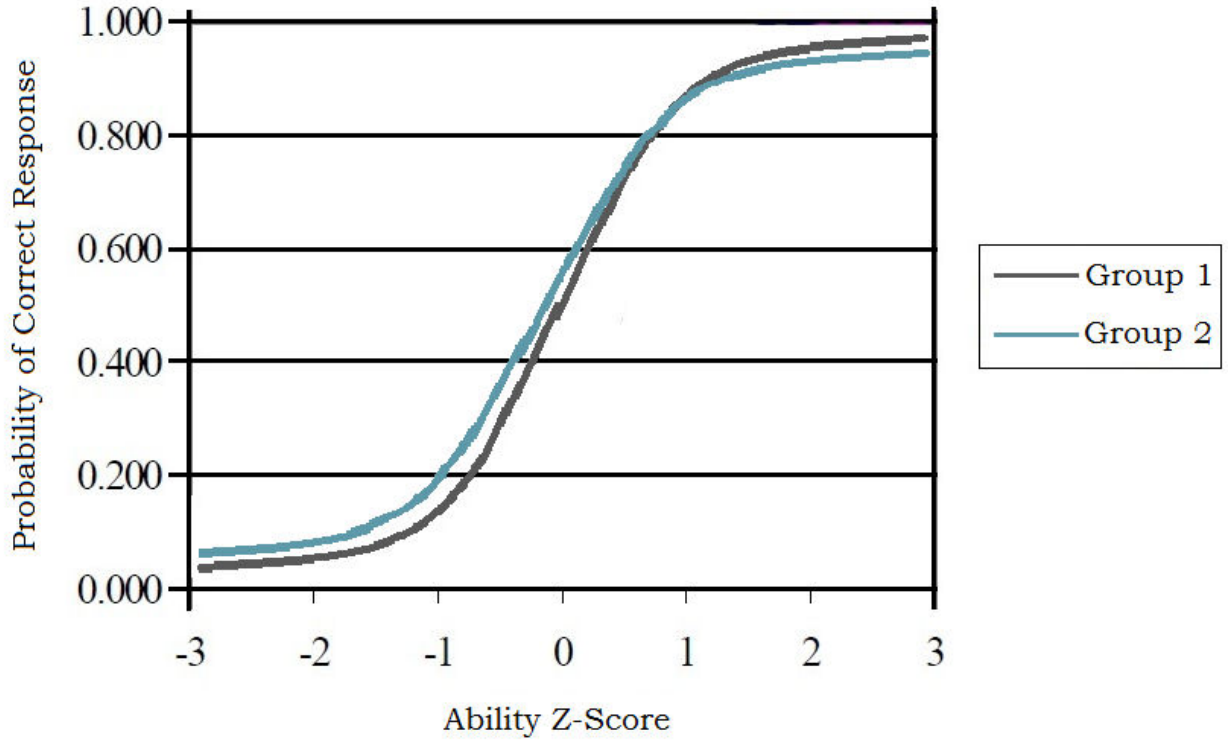


Figure 1. Item characteristic curves for reference and focal groups not demonstrating any type of DIF.

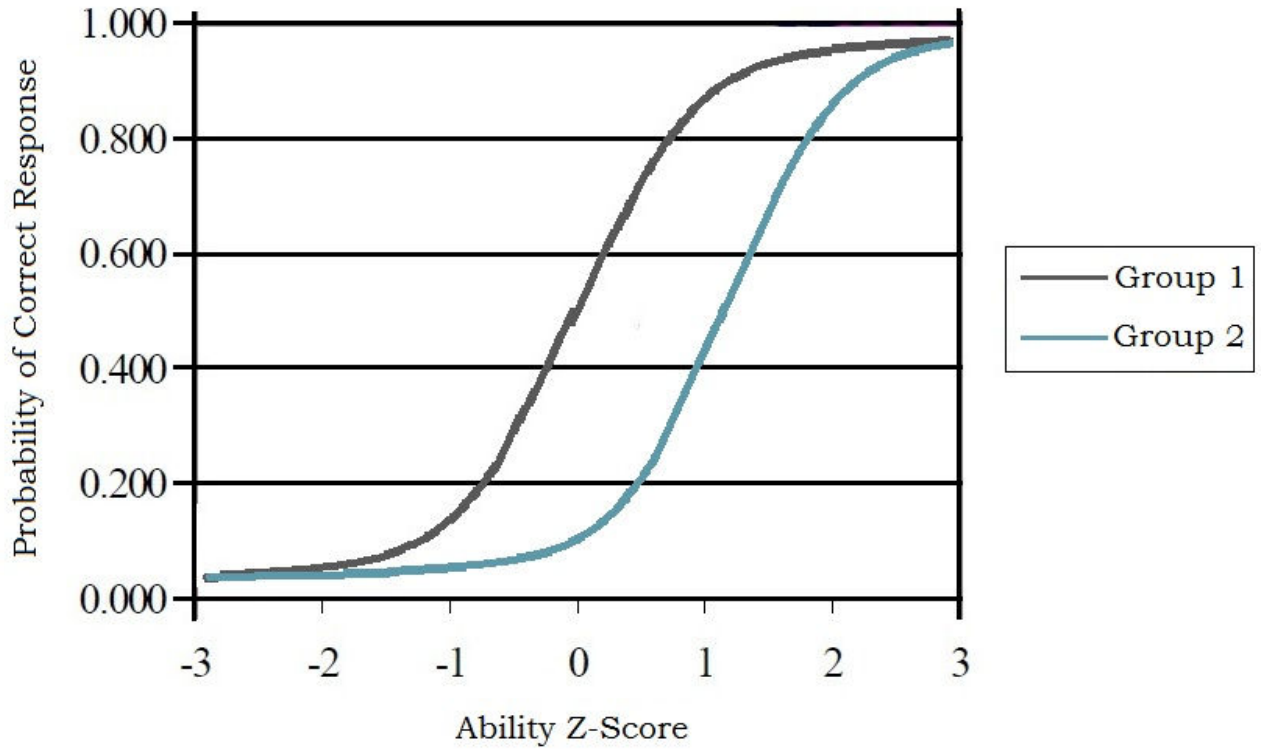


Figure 2. Item characteristic curves for reference and focal groups demonstrating uniform DIF.

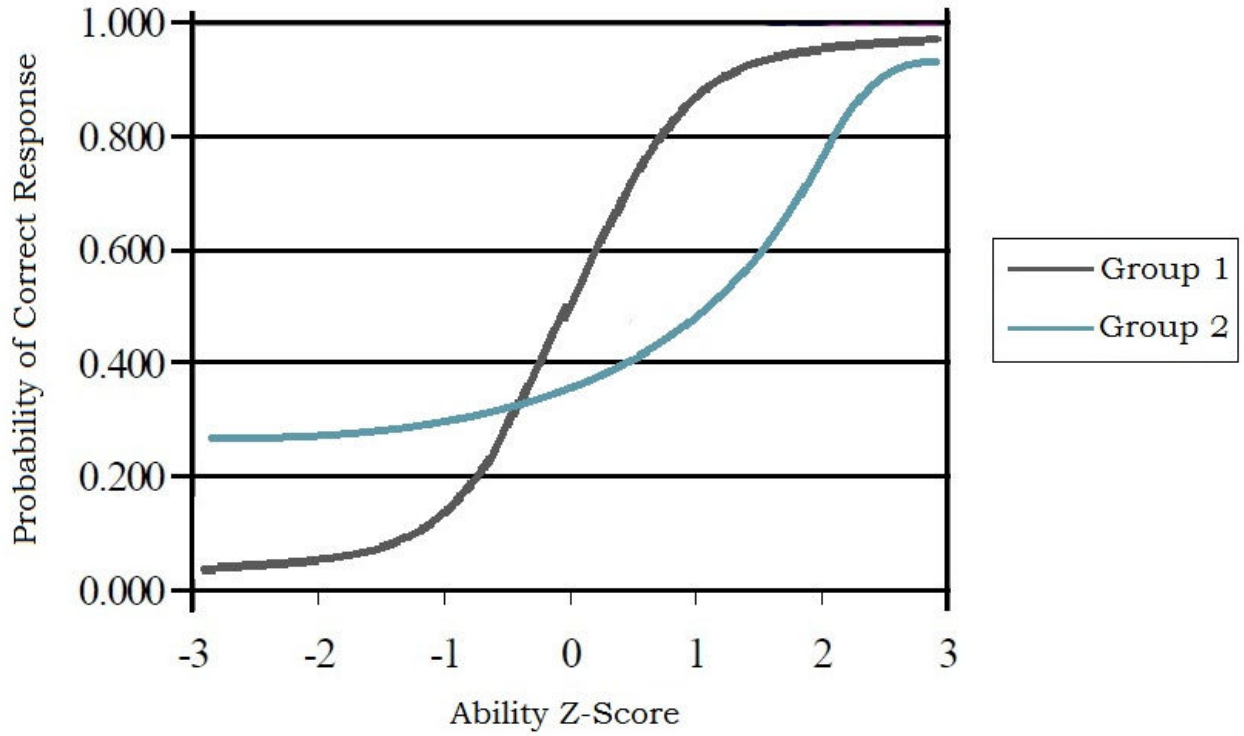


Figure 3. Item characteristic curves for reference and focal groups demonstrating non-uniform DIF.

PURPOSE

The purpose of this report is to systematically evaluate items from the STAR Reading assessment to determine whether DIF exists for gender and race subgroups. The evaluation is conducted in three phases, which begins by considering the full set of uncalibrated items that are being evaluated, as recommended by Osterlind and Everson (2009). This comprehensive evaluation of DIF for Star Reading items consists of (a) an informal DIF procedure; (b) the Mantel-Haenszel procedure; and (c) a logistic regression procedure.

METHOD

SAMPLE

Star Reading uncalibrated items from Batches RS5 and RS6 were used for analysis. The batches were combined and went live around Jan. 4th, 2011. A total of 17,050,812 cases were included in the original file. The number of cases (item responses) that had demographic information on gender was equal to 11,596,983 (68.0% of the total number of cases) and the number of cases that had demographic information on race was equal to 4,845,736 (28.4% of the total number of cases). Summary demographic information by cases and for individual assessment items are presented in Table 1. The total number of items contained in the merged data file was 1,222. As seen in Table 1, every assessment item contained in the data file had at least a number of cases with information on gender and race. Thus, the analyses were able to be conducted on the entire set of items available within these STAR Reading batches.

Although the sample sizes do attain the threshold of minimum sample size for some items, the samples for the Asian and Hispanic groups were not satisfactory for the complete DIF analysis described in detail later due to the need to break samples into ability ranges for the H-M analysis, which would cause them to be one fourth of their reported size in Table 1. This limitation highlights the difficulty in doing thorough DIF analyses for many measures, due to the sample size requirements. There were a number of cases that were filtered out of the data files during their original creation, as part of the data cleaning process. For example incomplete tests, duplicate assessments, duplicate students, and duplicate response records. This allowed for a dataset containing cases considered to be valid for the purposes of this analysis.

Table 1

Sample Demographics

	Dataset Cases Summary		Item-Level Sample Size Summary Statistics			
	Count[1]	Percent	*M*	*SD*	Min.	Max.
Gender						
Male	5,777,090	49.8	4,727	1,778	2,598	14,460
Female	5,819,893	50.2	4,762	1,829	2,489	14,432
Race						
White	2,281,189	46.9	1,867	731	813	5,830
Black	1,064,735	21.9	871	396	393	5,633
Hispanic	1,199,611	24.7	982	345	411	2,915
Asian	188,185	3.9	154	69	33	565
Native American	112,016	2.3	92	43	30	528

[1]The number of items that had race or gender demographic information is 1,222.

VARIABLES

Three files included in the calibration data were merged to obtain all the necessary information for the analyses. The raw data files used were the (a) Student Demographic Table; (b) Assessment Table; and (c) Uncalibrated Item Response Table. They were merged using SPSS software version 19 by matching either on the customer and user ID keyed variables or the customer and assessment ID keyed variables, depending on the structure of the tables. A number of variables in the resulting merged file were unused and only the variables necessary for the analyses are described in detail below.

ITEM CORRECT

The item correct variable (seen in the dataset as *iItemStatus*) was originally found in the Uncalibrated Item Response Table. This variable was coded with "1" representing a correct answer and "4" representing an incorrect answer. The variable was recoded using the *recode into different variable* function in SPSS, with "1" representing a correct response and "0" representing an incorrect response. The variable was labeled *ItemCorrect* to represent the dummy coding. All other values were coded as missing values (*System Missing* in SPSS) using

the *ELSE* function. The dummy coding was used because it will be needed in the informal DIF analysis described in detail later.

Gender

The gender variable (seen in the dataset as *chGender*) was found in the Student Demographics Table prior to recoding. Originally, the variable was coded using a string value of "M" for male and "F" for female. Cells with the value "U" were meant to represent that the gender of the student was "Not Specified" by the user. For the purposes of the analysis, the gender variable was recoded to a numeric value with a dummy coding using the *recode into different variable* function in SPSS, with "1" representing male and "0" representing female. The variable was labeled *GenderMale* to represent the dummy coding.

Race

The race variable (seen in the dataset as *iEthnicityID*) was found in the Student Demographics Table prior to recoding. The variable was originally coded using the numeric codes seen in Table 2. Based on the general demographic information seen in Table 1, small sample sizes were expected for races other than White, Black, and Hispanic. Thus, smaller sub-population demographics were collapsed into broader categories as seen in Table 2. As was done with previously described variables, the *recode into different variable* function in SPSS was used and the variable was labeled "RaceID". All other values were recoded to *System Missing* using the *ELSE* function within *recode into different variable.*

Table 2

Dataset Values of Race/Ethnicity Variables Before and After Recoding

Code	Description	Re-Coded Value	Re-Coded Description

786	American Indian or Alaskan Native	5	Native American
787	Asian or Pacific Islander	4	Asian
788	Black	2	Black
789	Hispanic	3	Hispanic
790	White	1	White
2093	Not Provided	7	System Missing
2094	Subcontinent Asian American	4	Asian
2095	Other Race or Ethnicity	6	Other
2096	Asian Pacific American	4	Asian
2097	Native American	5	Native American
2098	Native Hawaiian	5	Native American
2099	Pacific Islander	4	Asian

STAR Reading Scaled Score

STAR Reading software allows for the creation of an almost limitless number of test forms, based on the interaction between the tests items and the students taking the test. Thus, in order to be able to compare test scores between students, as well as for norm-referencing purposes, a conversion of raw test scores was needed. The common score used in this case was a scaled score, which was created using a two-step procedure. For more information on this procedure, see the STAR Reading Technical Manual (RLI, 2012). Star Reading Scaled Scores range from 0-1400, and these were provided in the Assessment Table as iScaledScore. This variable was not recoded for the analysis and the original labeled was also retained.

Assessment Items

Individual items needed to be identified for the proposed analyses. Thus, the iAssessmentItemID variable from the Uncalibrated Item Response Table was used to identify individual items. Although the original item identification numbers were used in the analyses, the variable was relabeled to ItemID.

Procedure

Individual files were created for each analysis based on the merged file that was created from the original data files used for calibration. The reason for individual files was to reduce the number of cases that would need to be processed either through sorting or case selection procedures, which would have a heavy processing load given the relatively large size of the complete STAR Reading data file. Therefore, one SPSS data file containing only cases with a value for the gender variable was created. In addition, another SPSS data file containing only cases that had a value for the race variable was created. The creation of the files was done using the Select Cases function in the SPSS program and the Delete Unselected Cases option was used to eliminate all unused cases.

REFERENCE AND FOCAL GROUPS

When conducting DIF analyses, the groups being compared are generally referred to as the reference and focal groups. The reference group represents the group that the item or test is thought to favor, which is typically assumed to be the majority group or males (Osterlind & Everson, 2009). However, it should be noted that for the purposes of analysis, it makes no difference which group is selected as the focal or reference group. Thus, this nomenclature is only useful for the purposes of discussion and will be used throughout this report.

STEP 1: INFORMAL DIF PROCEDURE

The informal DIF method was recommended as a quick, yet informative, method of obtaining information about items that may be flagged for potential DIF. The method used here is a variation of the method outlined in Osterlind and Everson (2009), as a preliminary approach to examining a set of assessment items for DIF. Osterlind and Everson (2009) use an ordinal

ranking of items based on their respective p-value[1] differences between groups. Here, a difference in the absolute p-value between comparison groups is used and referred to as a *p-value difference statistic*. This was considered to be an improvement due to the relatively large number of items to be analyzed. The use of the rank ordering system would have been far more difficult to evaluate and the difference in absolute p-value difference statistics captures similar information, as a greater absolute p-value difference statistic is likely to result in a greater discrepancy in the rank ordering of an item between groups.

P-values were calculated using the Aggregate function from the SPSS v.19 software. The aggregate function computes the mean correct value for each item by group. The aforementioned dummy scoring 0-1 (1 being a correct response) allowed the mean to be interpreted as the proportion (in the form of a percentage) of the group of getting that item correct. In order to accomplish this, ItemID was identified as a break variable. The mean function was then selected for Summary of Variables. The sample size was also selected to be a part of the resulting aggregated file. Finally, an aggregated data file was created for each of the comparison groups (i.e. male, female, White, Black, Hispanic, Asian, Native American) by using the Select Cases function from the data menu. The summary statistics provided by this set of procedures were then pasted into a Microsoft Excel file.

The p-values of each item were then calculated in the *Excel* file using a simple equation that provided the absolute value of the p-value difference between the comparison groups. The equation was as follows:

$$=\text{ABS}(\text{CELL X} - \text{CELL Y}) \qquad (1)$$

[1] *P-values* represent the mean proportion of students in a group answering the item correctly.

where CELL X and CELL Y represent the cells containing for mean p-value for the groups being compared. After this was calculated for all items, the spreadsheet was *formatted as a table* and the absolute p-value difference statistics were sorted from largest to smallest for evaluation. Due to the arbitrary nature of this procedure, no cut-score was selected prior to analysis. The biggest drawback of this analysis is that it does not control for differences in ability that may exist between groups. However, this limitation is addressed to some degree in the next step and very well in the third step of the previously described three-step procedure. Nonetheless, the informal DIF results will provide some information about items that may be demonstrating DIF. This is particularly useful in this case due to the relatively large number of items. In this case, the top 100 items, which is roughly 10 percent of the items, demonstrating the greatest p-value difference statistics for each comparison (i.e. Male-Female, White-Black, and White-Hispanic) was selected for further analysis.

STEP 2: MANTEL-HAENSZEL PROCEDURE FOR EVALUATING DIF

The Mantel-Haenszel procedure is known today as one of the most commonly used methods of detecting DIF (Osterlind & Everson, 2009). The Mantel-Haenszel procedure uses a 2x2 (for dichotomous items) cross-tabulation of the two groups by the item response groups (correct or incorrect). The respective counts for each column and row intersection represent the count of the number of cases belonging to each pairing of focal-reference group by item response group. The Mantel-Haenszel procedure uses a variation of a typical chi-square procedure known as the full chi-square. For more information on the calculation of the Mantel-Haenszel procedure statistics, see chapter 7 of Osterlind and Everson (2009).

This chi-square contingency table is run for each item, at different levels of ability. Multiple levels of ability were considered because it was suggested by Osterlind and Everson (2009) to

be "(…) a critically important part of all procedures of DIF investigation is to match examinees by ability or proficiency on the appraised construct. The matching is usually done by total test score" (p. 14-15).

Thus, for the current analysis, the groups were split into quartiles to account for some of the non-uniform DIF that may be occurring based on ability. Quartiles were created based on a re-coding of the percentile rank variable by putting students with a percentile rank ranging from 0 to 25 in the 1st quartile, students with a percentile rank from 25 to 50 in the 2nd quartile and so on, using the recode into different variable function in SPSS v.19.0.

After the recoding is complete, the Split File function is used to organize the output by quartile and item number. The Crosstabs function from the SPSS v19.0 software is used for this analysis, given that it provides the appropriate Mantel-Haenszel statistic in the statistic sub-menu, using a common odds ratio of 1.0.

The output from this analysis will provide a number of statistical tests. The noteworthy one for the purposes of this analysis will be the Mantel-Haenszel test of conditional independence. The cross-tabs for the groups will also be noted to see how discrepant group membership will need to be in order to observe a significant effect based on a test of significance. Due to the relatively large number of statistical tests being run, a conservative alpha (significance cut-off) for the tests of significance was set at $p > .001$ to reduce the likelihood of committing a Type II error. The p-value was selected based on an approximation of a Bonferonni adjustment, which would have set the p-value for significance at $p = .0005$.

STEP 3: LOGISTIC REGRESSION DIF PROCEDURE

The logistic regression approach extends and improves upon the previously described Mantel-Haenszel procedure by controlling for ability as a continuous variable, instead of creating

arbitrary ability sub-groups. It also provides information on the extent to which uniform and non-uniform DIF may be occurring. The logistic regression procedure uses a logistic function, which takes the following general form:

$$f(z) = \frac{1}{1 + e^{-z}} \qquad (2)$$

The current analysis models z using a number of regression coefficients within the function as follows:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \qquad (3)$$

The βs in the equation represent the various regression weights, estimated in the course of the analysis, for each of the independent variables. The variable X_1 represents ability, which is the total STAR Reading score (TotalScore) in the dataset. The variable X_2 represents the group membership such as gender (GenderMale) and race (RaceID). The multiplied variables X_1X_2 represent the interaction term, which provide information on the extent to which non-uniform DIF is occurring, if this interaction term is significant. The benefit of using a regression equation is that the effect of each independent variable is tested while controlling for the others in the equation. For example, when we interpret the interaction term for non-uniform DIF, we are making this interpretation controlling for ability and group membership (uniform DIF).

The logistic regression DIF procedure to be used here is described in detail in Zumbo (1999). Zumbo (1999), provides SPSS v.17 syntax to appropriately assess uniform and non-uniform DIF using logistic regression. This syntax was modified for the purposes of the comparisons made in the current procedure, using the variables from the STAR Reading dataset, and is available in Appendix E.

As was the case for the Mantel-Haenszel procedure, there is a great deal of output that is produced by the SPSS processor based on the provided syntax. However, the noteworthy pieces of output are: (a) the model chi-square and its significance; and (b) the Nagelkerke R-square (Nagelkerke, 1991) for each block or step-wise entry of the variables into the logistic regression model.

It should be noted that the Nagelkerke R-squared used in these analyses is referred to as a "pseudo r-squared" (Peck, Roeser Zarrrett, & Eccles, 2008) because it is not interpreted in the same way as a typical r-squared value due to the nature of the logistic regression models. Basically instead of the r-squared representing variance accounted for by the variables, its value represents how well the variables in the model predict the likelihood of a correct answer on the item being evaluated.

The data from this output will be summarized into an *Excel* spreadsheet. In order to determine the unique contribution of uniform and non-uniform DIF, a series of simple formulas will need to be calculated from the output data. The following formula[2] was used to obtain the chi-square value to test for the uniform DIF effect only:

=(Chi-square value for Step 2) – (Chi-square value for Step 1) (3)

The following formula was used to obtain the chi-square value to test for the simultaneous contributions of uniform and non-uniform DIF:

=(Chi-square value for Step 3) – (Chi-square value for Step 1) (4)

---

[2] Although the general form of the formulas presented in this section represent the calculations that were done to obtain the necessary information on uniform and non-uniform DIF, the actual cell values will have to be substituted if these formulas are to be used in similar analyses. If a user does not understand how to convert the formulas here to a format recognized by Excel, the program's help menu can provide the appropriate form for these functions.

In order to test for the statistical significance of these values, the following formula was used:

=CHIDIST(Chi-square value, 1) (5)

=CHIDIST(Chi-square value, 2) (6)

It should be noted that two degrees of freedom are used in formula 4 when testing only for uniform DIF because of the way this chi-square was calculated. Specifically, the chi-square value at step 2 had two degrees of freedom and this was subtracted the chi-square value at step one, which had 1 degree of freedom. Thus, the resulting degrees of freedom for this chi-square is the difference between these two values (df = 1). The same calculation was done for formula 5, with the resulting degrees of freedom being equal to 2.

The r-square attributed to uniform DIF was obtained through a similar procedure. The following formula was entered into the table to obtain the value for all items:

=(R-squared[3] for Step 2) – (R-squared for Step 3) (7)

Similarly, the r-square attributed to non-uniform DIF was calculated by using the following formula:

=(R-squared for Step 3) – (R-squared for Step 2) (8)

Basically, the subtraction of the output removes the effects of the variables that are included in both equations and provide the effect of the unique variables in each of the equations. For more information on this procedure, see Zumbo (1999).

[3] Nagelkerke r-squared

This final analysis should provide comprehensive information related to the extent to which DIF is occurring for the items analyzed. The use of this three-step procedure is likely to produce convergent evidence about the presence and magnitude of the DIF effect, if it occurring.

RESULTS

INFORMAL DIF

The informal DIF analysis included all of the 1,222 items from the STAR Reading dataset. Summary statistics from the p-value difference statistics can be found in Table 3. Overall, item p-values based on race comparisons demonstrated greater differences between groups than those seen for the gender comparison. However, when evaluating all items, some had p-value differences equal or less than 0.001 across all comparisons. This indicates that the probability of answering these items correctly was basically the same for these items, for all subgroups. Despite some of the items demonstrating very low p-value differences, others demonstrated a relatively high p-value differences difference, with some having over a p-value difference of 0.25 between groups. As planned, the top 100 items demonstrating the largest difference in p-value between comparison groups were used for the remaining analyses. Summary statistics for this group can be found in Table 3 and more detailed item-level statistics can be found in Appendix B.

Table 3

Summary of Informal DIF P-Value Difference Statistics

Comparison Groups (Reference-Focal)	P-Value Difference Statistics

		Mean	SD	Min	Max
Complete Sample					
	White-Black	0.091	0.045	0.001	0.270
	White-Hispanic	0.084	0.042	0.000	0.308
	Male-Female	0.030	0.021	0.000	0.115
Top 100 Items with Largest P-Value Difference					
	White-Black	0.181	0.024	0.157	0.270
	White-Hispanic	0.177	0.033	0.144	0.308
	Male-Female	0.075	0.011	0.000	0.115

Mantel-Haenszel Procedure

The Mantel-Haenszel procedure to assess DIF for STAR Reading items was completed for three comparisons, which were White-Black, White-Hispanic, and Male-Female. The four ability quartiles allow for statistical control of ability, to some extent. The results of this analysis are summarized in Figure 1 and shown in more detail in Appendix C.

Overall, there were items identified as demonstrating significant M-H DIF across all ability levels for all comparison groups. There is, however, a considerable decrease in the number of items demonstrating DIF associated with an increase in the ability quartiles; that is, as grop abilityincreased, the number of items with significant M-H DIF declined. This trend is consistent across the comparisons, with a particularly large difference in the number of items identified as demonstrating DIF between the first and fourth ability quartiles for the gender comparison (see Figure 1).

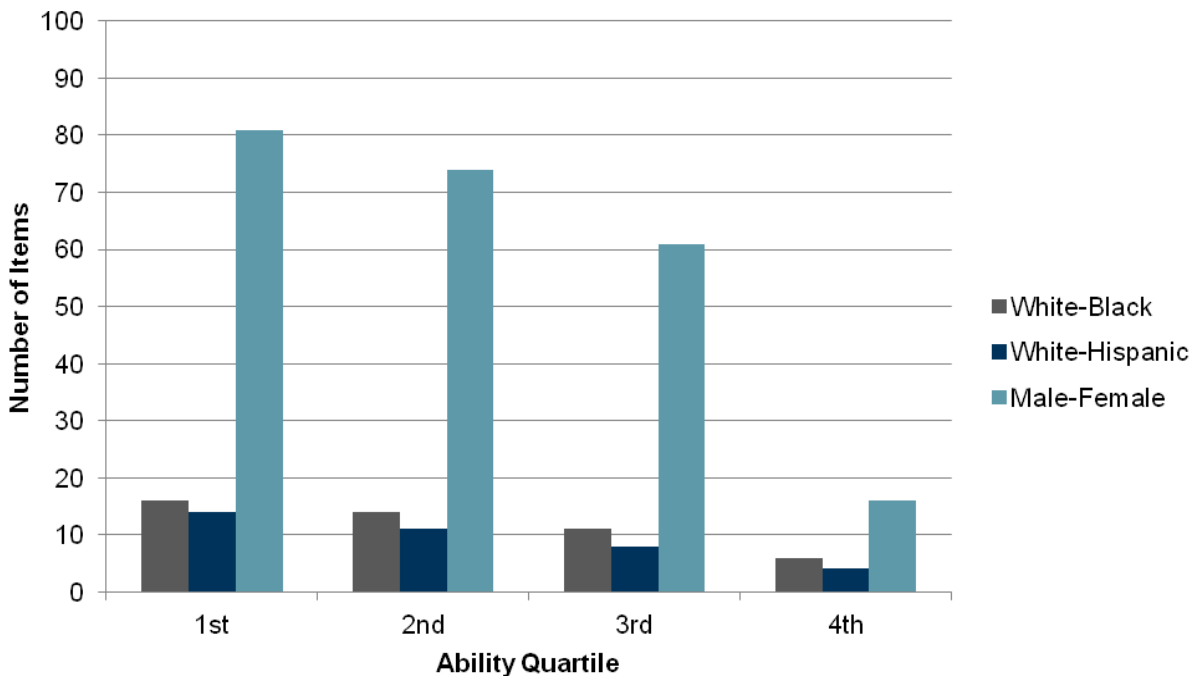


Figure 1. Items demonstrating statistically significant M-H DIF values between comparison groups displayed by ability quartiles.

The results of the Mantel-Haenszel procedure suggest that some DIF may be occurring across group comparisons and ability levels (see Figure 1). This is particularly true, according to these results, for the gender comparison, which had a significantly greater number of items identified at all abilities when compared to the other comparison groups. These results support further exploration of these data in order to gather more information on the interaction between ability and performance on the assessment items to determine if the significant differences seen here are true DIF or if they can be attributed to variations in ability occurring within the arbitrary quartiles that were selected for this analysis.

Logistic Regression

The DIF analysis that used logistic regression produced a great deal of data. First, the model chi-square values and Nagelkerke R-square values were obtained for all three models. Recall that the models were the ability model, the ability and uniform DIF model, and the ability, uniform and non-uniform DIF model. The chi-square and Nagelkerke R-square values for these individual models are available in Appendix D.

The data of most interest are the calculations that allow for the evaluation of the contribution of uniform and non-uniform DIF. Figure 2 provides a visual summary of the number of items that had significant chi-square values for the test of uniform DIF and the simultaneous test of uniform and non-uniform DIF. The statistics for individual items are available in Appendix D.

As seen in Figure 2, a portion of the items for the White-Black and White-Hispanic comparisons demonstrated statistically significant uniform and non-uniform DIF. The results of the gender comparison analyses suggest that all of the items in this comparison are demonstrating uniform and non-uniform DIF. The results seen in Figure 2 are somewhat consistent with those seen in the Mantel-Haenszel procedure, with a number of items from each comparison demonstrating some DIF and all 100 items demonstrating gender DIF.

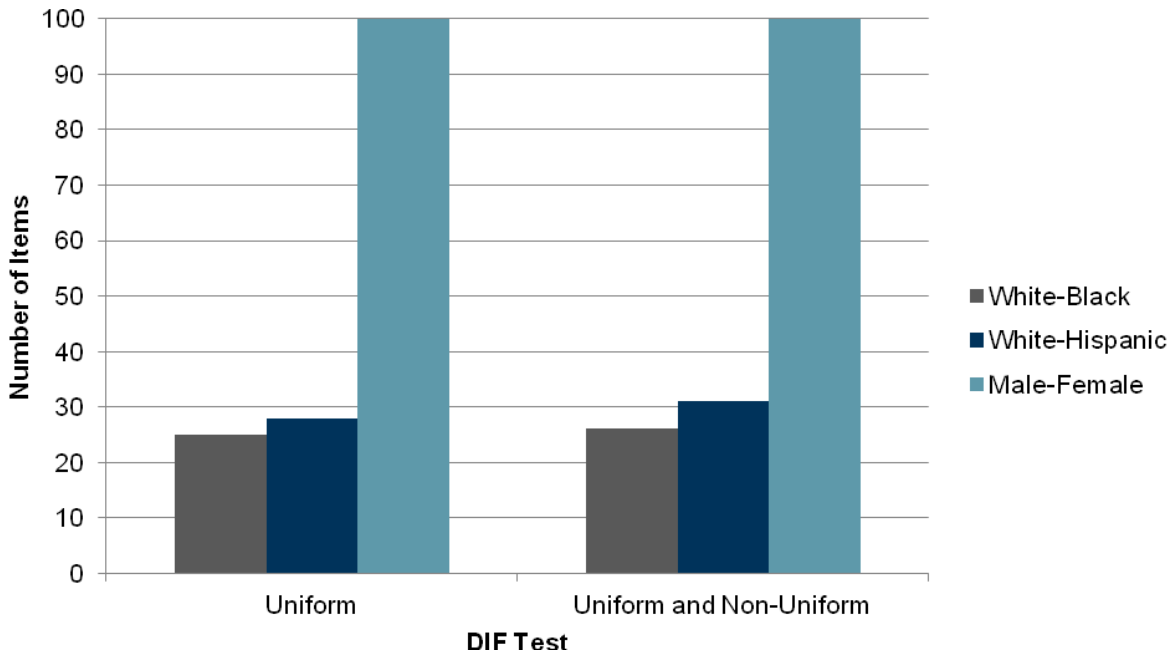
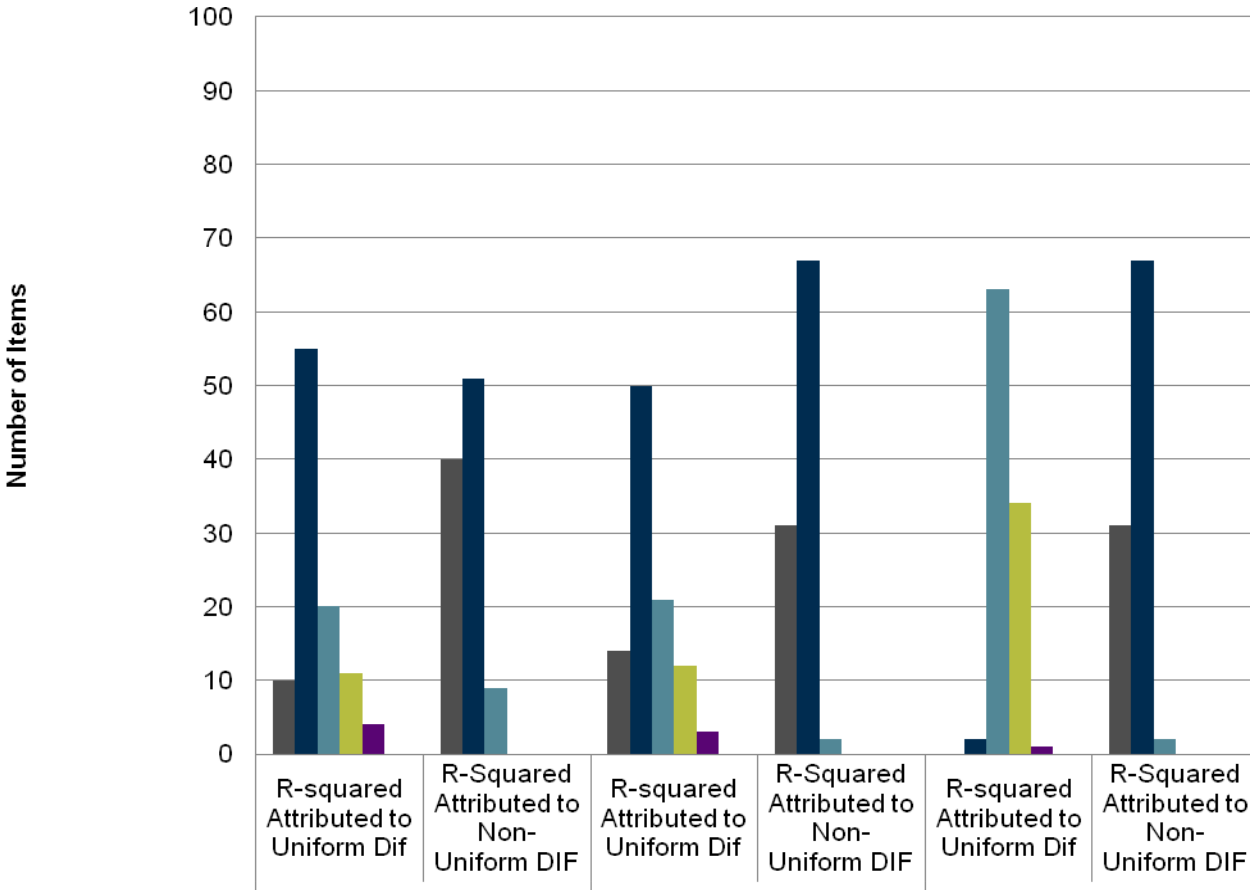


Figure 2. Number of items with significant chi-square values when testing for uniform DIF only and uniform and non-uniform DIF simultaneously.

In order to provide better context for the results seen in Figure 2, Figure 3 provides the Nagelkerke R-squared value ranges for all the items evaluated for uniform and non-uniform DIF, by comparison groups. This greater context is needed because although items may have a significant value based on a statistical test, closer inspection of the results, typically in the form of an effect size like an r-squared, is needed to provide an indication of the practical significance of the results. Thus, it should be noted that only the Nagelkerke values above 0.05 are considered to have a value that is not negligible for practical purposes. According to these data, the Nagelkerke R-square value for the contributions of either uniform or non-uniform DIF are less than .01 for more than 50% of the items, and less than .025 for almost all of the items. Furthermore, the contribution of non-uniform DIF was less than .001 for the

majority of the items. Thus, it can be concluded that once ability is adequately controlled for statistically, DIF between race or gender groups is reduced to a negligible value. In other words, a student's race is not related to the probability of a student being able to correctly answer any of the items analyzed. Thus, this final analysis has demonstrated that the items analyzed do not demonstrate a noteworthy bias toward a particular gender or race.



	White-Black		White-Hispanic		Male-Female	
	R-squared Attributed to Uniform Dif	R-Squared Attributed to Non-Uniform DIF	R-squared Attributed to Uniform Dif	R-Squared Attributed to Non-Uniform DIF	R-squared Attributed to Uniform Dif	R-Squared Attributed to Non-Uniform DIF
Less than .001	10	40	14	31	0	31
.001-.005	55	51	50	67	2	67
.005-.010	20	9	21	2	63	2
.010-.025	11	0	12	0	34	0
.025-.050	4	0	3	0	1	0
Greater than .050	0	0	0	0	0	0

Figure 3. Summary of Nagelkerke R-squared values attributed to uniform and non-uniform DIF for 100 items from the STAR Reading assessment across comparison groups.

LIMITATIONS

The results of the analyses presented should be considered in light of a few limitations. Due to sample size limitations, DIF could not be evaluated for Asian and Native American sub-groups. Therefore, the results presented are limited to the sub-groups evaluated and the inference should not be drawn that no significant DIF exist for other sub-groups. In addition, the analysis is limited to the sample of items that were included: SRRS Batches 5 and 6. Analyses of all other SRRS Batches, and of STAR Reading Service items as well, would be needed in order to be confident that the results reported here apply to STAR Reading in general – both the Service and Enterprise versions.

The approach taken, of including items that demonstrated the greatest differences between groups (i.e. the informal DIF procedure) does provide some confidence about the overall pool of items in Batches 5 and 6. However, there are a number of items from other STAR Reading batches that were not included in any of the analyses. Thus, it is advisable to replicate the procedures on other Batches of STAR Readng items to gather more evidence on the extent to which DIF may be occurring between these groups and others.

CONCLUSION

In general, the results of the analyses reported here are encouraging; they suggest that although statistically detectab;le DIF may be occurring for some of the items contained in the STAR Reading assessments, its effect is basically negligible in terms of its actual contribution to differences in the probability of one group answering an item correctly over another. Thus,

there is no need for an extensive discussion of how the reference group was favored on a given item relative to the focal group in the analyses. Furthermore, it appears that no further editorial analysis of the content of items in SRRS Batches 5 and 6 is needed, as it is unnecessary to consider making any changes to the items included in the analysis to reduce their bias for certain sub-groups.

The thorough, three-step procedure used to examine the influence of DIF provided a great deal of information by reducing the potentially confounding influence of ability in the interpretation of DIF results. Furthermore, the use of the Nagelkerke R-squared provided a numeric value of the extent to which the effect of DIF was influencing item responses, for those items with statistically significant uniform or non-uniform DIF. If it were not for these pieces of information, it is likely that substantially different recommendations would have been made. In summary, although a number of items appeared to demonstrate DIF in preliminary analyses, the analyses in Steps 2 and 3 led to the conclusion that the influence of DIF on the assessment items included here is not statistically significant or, for the cases that were statistically significant, the DIF contribution was negligible. Thus, it can be concluded that the items analyzed herein are not biased toward one racial group or another and are providing information on differences in ability between students taking the assessments. This result provides evidence to support a strong validity argument for the STAR Reading assessments by demonstrating that variations in scores are due to ability rather than construct irrelevant variance that is caused by gender or racial bias

This process not only provides validity evidence for the STAR Reading assessments, but also provides support for the item development process currently in place. Although the item development process was not evaluated directly in this analysis, the items resulting from this process appear to be very robust in not demonstrating any gender or racial bias. Moreover, the

contribution of ability in demonstrating differences in the probability of correctly answering a given item on a STAR Reading assessment is what would be expected of this measure, given that the STAR Reading assessments, much like most assessments, are designed to detect differences in ability between individuals. Thus, the item development process currently in place appears to be producing strong items that are free gender or racial bias.

In conclusion, this report provides validity evidence for the use of STAR Enterprise products. In general, educators can be reassured that the items that are being presented to their students are not biased toward a particular sub-group and that the results of the assessments taken by their students will provide them with meaningful information about every student's ability level, which will consequently inform their instructional practices.

REFERENCES

American Psychological Association, National Council on Measurement in Education, American Educational Research Association (1999). Standards for Educational and Psychological Testing, Second Edition. American Educational Research Association, Washington, D.C., 194 pages.

Messick, S. (1995a). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice,*

Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Milone, M. (2008). *The development of ATOS: The Renaissance readability formula*. Wisconsin Rapids, WI: Renaissance Learning Incorporated.

Nagelkerke, N.J.D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika, 78*(3), 691-692.

Osterlind, S.J. (1983). Test item bias. In Lewis-Beck, M.S. (Series Ed.), *Quantitative Applications in the Social Sciences Series.* Newbury Park, CA: Sage Publications.

Osterlind, S.J., & Everson, H.T. (2009). Differential Item Functioning, 2nd Edition. In Liao, T.F. (Series Ed.), *Quantitative Applications in the Social Sciences Series.* Thousand Oaks, CA: Sage Publications.

Peck, S.C., Roeser, R.W., Zarrett, N., & Eccles, J.S. (2008). Exploring the roles of extracurricular activity quantity and quality in the educational resilience of vulnerable adolescents: Variable- and pattern-centered approaches. *Journal of Social Issues, 64*(1), 135-155.

Renaissance Learning Incorporated (2012). *STAR Reading Technical Manual*. Wisconsin Rapids, WI: Renaissance Learning Incorporated.

Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type item scores.* Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, National Defense Headquarters.

Appendix A: Asian and Native American DIF Analyses

PROCEDURE

Due to concerns related to the limited sample sizes for a number of the items, the cut-score rule that was used for the DIF analyses for both of these groups used a minimum focal group sample size of 300 cases instead of using the results from the informal DIF procedure to guide the selection of the items most likely to be demonstrating significant DIF between groups. The following section, however, provides summary statistics from the informal DIF in order to provide evidence for the generalizability of the results from a subset of these data.

INFORMAL DIF RESULTS

As seen in Table 1, the results of the informal DIF procedure suggest that the mean p-value difference statistics calculated across all STAR Reading items included in the dataset were consistent across sample sizes. In other words, on average, p-value differences did not appear to be related to sample size. In fact, for the White-Native comparison, the highest mean p-value difference statistic is seen in items that have a sample size of more than 300 cases for the focal group (e.g. Native Americans). This is beneficial as the logistic regression DIF will only be performed on the items that have reached the aforementioned threshold of 300 cases per group. In summary, the results of the informal DIF procedure appear to suggest that it is safe to generalize the results of the analyses that are only conducted on a subset of the sample, to other subsets of the overall sample.

Table 1

STAR Reading Informal DIF Summary Statistics

Comparison	Range for Number of Cases	Number of Items within Range	Mean P-Value Difference Statistic	*SD*
White-Asian				
	300+	58	0.02	0.03
	200-300	114	0.03	0.02
	150-199	326	0.03	0.03
	100-149	552	0.04	0.03
	0-99	173	0.05	0.04
White-Native American				
	300+	5	0.12	0.05
	200-300	35	0.05	0.03
	150-199	34	0.06	0.03
	100-149	214	0.08	0.05
	0-99	934	0.09	0.06

As seen in Table 2, the results of the informal DIF procedure for STAR Math are very similar to those seen for STAR Reading. Specifically, they suggest that the mean p-value difference statistics calculated across all STAR Math items included in the dataset were consistent across sample sizes. Again we see that, on average, differences in p-values do not appear to be related to sample size. Thus, we can again make some generalizations from the results of the analyses that are only conducted on a subset of the sample.

Table 2

STAR Math Informal DIF Summary Statistics

Comparison	Range for Number of Cases	Number of Items within Range	Mean P-Value Difference Statistic	*SD*
White-Asian				
	300+	7	0.03	0.02
	200-300	79	0.04	0.04
	150-199	25	0.06	0.04
	100-149	97	0.07	0.04

	0-99	346	0.09	0.07
White-Native American				
	300+	22	0.09	0.06
	200-300	63	0.08	0.05
	150-199	29	0.04	0.04
	100-149	131	0.08	0.05
	0-99	309	0.09	0.07

LOGISTIC REGRESSION

Chi-Square. The Chi-Square test for significant uniform and non-uniform DIF yielded similar results across groups for both STAR Reading and STAR Math. The White-Asian comparison for STAR Reading identified only 8 items (13.8%) that had a significant Chi-Square value when testing for Uniform DIF and 6 items (10.3%) that were demonstrating statistically significant uniform and non-uniform DIF. The Native American-White comparison for STAR Reading had no items that demonstrated a statistically Chi-Square value for both uniform and non-uniform DIF.

The White-Asian comparison for STAR Math identified no items that demonstrated a statistically significant Chi-Square value for both uniform and non-uniform DIF. The Native American-White comparison for STAR Math identified only 1 item (4.5%) that had a significant Chi-Square value when testing for Uniform DIF and 2 items (9.0%) that were demonstrating statistically significant uniform and non-uniform DIF.

Nagelkerke R-Squared. The results of the Nagelkerke R-squared calculations to determine the effect size of the DIF that may be occurring for the items analyzed were very similar to those seen with other groups, such as White-Black and White-Hispanic comparisons. The Nagelkerke R-Squared values were calculated for all of the items with sample sizes over 300. Overall, across all comparisons and for both STAR Reading and STAR

Math the R-squared values ranged from less than .001 to .005, which can be interpreted as an effect size that has no practical significance in terms of race explaining any difference in the probability of answering an item correctly.

Table 5

Uniform and Combined Uniform and Non-Uniform DIF Results for STAR Reading and STAR Math by Comparison Group

			Item Counts for Nagelkerke R-Squared Ranges		
		N	< .001	.001-.005	< .01
Uniform DIF					
STAR Reading					
	Asian	58	23	31	4
	Native American	5	2	3	0
STAR Math					
	Asian	7	5	2	0
	Native American	22	11	11	0
Uniform & Non-Uniform DIF					
STAR Reading		58	31	27	0
	Asian				
	Native American	5	2	3	0
STAR Math		7	3	4	0
	Asian				
	Native American	22	10	12	0

CONCLUSION

The results of these analyses are consistent with those seen in other comparisons (e.g. White-Black and White-Hispanic). In brief, once ability is controlled for statistically, any DIF that is observed in the informal DIF procedure is non-significant. Therefore, despite the analyses being conducted on only a small subset of items that met the sample size threshold, the

results of the analyses suggest that no DIF is occurring when comparing the performance of Asian or Native American students to White students. In summary, the results of these analyses further support the use of the STAR Reading and STAR Math assessments with students from culturally and linguistically diverse backgrounds.

APPENDIX B

Table A-1

Top 100 Items with Largest Informal DIF Statistics for the White-Black Comparison Group

Count	Item ID	Mean P-Value White	Mean P-Value Black	Absolute P-Value Difference Statistic
1	39836	0.507	0.237	0.270
2	39370	0.817	0.555	0.261
3	39065	0.750	0.506	0.244
4	38597	0.704	0.464	0.240
5	38601	0.810	0.570	0.239
6	39104	0.674	0.436	0.238
7	39818	0.572	0.343	0.229
8	38362	0.754	0.527	0.227
9	39936	0.723	0.501	0.221
10	39861	0.817	0.605	0.212
11	39862	0.493	0.281	0.212
12	38496	0.738	0.528	0.211
13	39835	0.781	0.572	0.209
14	39874	0.938	0.730	0.208
15	39949	0.560	0.356	0.204
16	39856	0.646	0.443	0.203
17	38532	0.745	0.545	0.201
18	39859	0.426	0.226	0.200
19	40031	0.776	0.576	0.200
20	39883	0.570	0.374	0.196
21	38437	0.519	0.323	0.196
22	38564	0.700	0.505	0.195
23	39304	0.861	0.666	0.194
24	38851	0.566	0.374	0.193
25	39352	0.767	0.575	0.192
26	39832	0.474	0.283	0.191
27	39972	0.812	0.623	0.190
28	39013	0.718	0.529	0.188
29	39875	0.727	0.541	0.186
30	39387	0.575	0.390	0.186
31	39944	0.691	0.505	0.185
32	38489	0.656	0.473	0.184
33	38812	0.639	0.455	0.183
34	38637	0.796	0.613	0.183
35	39390	0.740	0.557	0.183

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		White	Black	
36	39494	0.663	0.482	0.182
37	39120	0.697	0.516	0.181
38	39066	0.573	0.393	0.181
39	38815	0.605	0.425	0.180
40	36786	0.321	0.143	0.178
41	37973	0.686	0.509	0.177
42	39894	0.931	0.754	0.177
43	39359	0.805	0.629	0.176
44	38392	0.653	0.477	0.176
45	38626	0.663	0.487	0.176
46	38408	0.780	0.607	0.173
47	38260	0.829	0.656	0.173
48	39882	0.612	0.439	0.173
49	39841	0.715	0.543	0.172
50	37805	0.836	0.665	0.171
51	37797	0.646	0.475	0.171
52	39825	0.563	0.392	0.170
53	38761	0.801	0.630	0.170
54	38616	0.678	0.508	0.169
55	38664	0.550	0.381	0.169
56	38213	0.790	0.621	0.169
57	38628	0.755	0.586	0.169
58	38420	0.470	0.301	0.169
59	38599	0.492	0.323	0.168
60	38378	0.791	0.622	0.168
61	38814	0.671	0.503	0.168
62	38375	0.667	0.499	0.168
63	38632	0.725	0.557	0.167
64	38795	0.591	0.424	0.167
65	39469	0.778	0.611	0.167
66	38825	0.695	0.528	0.167
67	39824	0.718	0.551	0.167
68	37876	0.683	0.517	0.167
69	38778	0.700	0.534	0.167
70	38806	0.725	0.559	0.166
71	39132	0.803	0.637	0.166
72	39112	0.885	0.719	0.166
73	38743	0.742	0.576	0.166
74	38465	0.507	0.341	0.166
75	39358	0.845	0.679	0.166

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		White	Black	
76	38767	0.787	0.622	0.165
77	38537	0.630	0.465	0.165
78	39382	0.741	0.576	0.164
79	36572	0.843	0.680	0.164
80	39833	0.573	0.410	0.163
81	38524	0.543	0.380	0.162
82	37824	0.680	0.519	0.162
83	38766	0.606	0.444	0.162
84	38449	0.732	0.571	0.161
85	37671	0.697	0.536	0.161
86	39067	0.690	0.529	0.161
87	37239	0.611	0.450	0.161
88	37802	0.822	0.661	0.160
89	38855	0.821	0.661	0.160
90	39071	0.814	0.654	0.160
91	39105	0.761	0.602	0.159
92	39354	0.812	0.653	0.159
93	40032	0.700	0.542	0.159
94	37795	0.839	0.681	0.158
95	37813	0.686	0.529	0.158
96	38247	0.684	0.526	0.158
97	39338	0.746	0.589	0.157
98	39846	0.867	0.710	0.157
99	38588	0.771	0.615	0.157
100	38400	0.774	0.617	0.157

Table A-2

Top 100 Items with Largest Informal DIF Statistics for the White-Hispanic Comparison Group

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		White	Hispanic	
1	37861	0.744	0.436	0.308
2	39856	0.646	0.354	0.292
3	39065	0.750	0.459	0.291
4	40031	0.776	0.518	0.258
5	38597	0.704	0.465	0.238
6	39874	0.938	0.708	0.231
7	39883	0.570	0.344	0.226
8	38489	0.656	0.431	0.226
9	39352	0.767	0.541	0.225
10	39965	0.839	0.614	0.225
11	38532	0.745	0.521	0.224
12	38825	0.695	0.479	0.217
13	38766	0.606	0.390	0.216
14	39104	0.674	0.460	0.214
15	39836	0.507	0.300	0.208
16	39936	0.723	0.515	0.207
17	39868	0.642	0.435	0.207
18	38795	0.591	0.385	0.207
19	39066	0.573	0.367	0.207
20	38815	0.605	0.402	0.203
21	39304	0.861	0.667	0.194
22	39305	0.835	0.641	0.194
23	39862	0.493	0.301	0.192
24	38725	0.764	0.574	0.190
25	36786	0.321	0.131	0.190
26	38496	0.738	0.550	0.189
27	39944	0.691	0.503	0.187
28	37239	0.611	0.424	0.187
29	38373	0.550	0.364	0.186
30	38826	0.775	0.593	0.182
31	39387	0.575	0.395	0.180
32	37677	0.589	0.409	0.180
33	39894	0.931	0.752	0.179
34	39825	0.563	0.385	0.178

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		White	Hispanic	
35	38823	0.767	0.591	0.177
36	38437	0.519	0.343	0.177
37	39833	0.573	0.396	0.176
38	39824	0.718	0.542	0.176
39	38392	0.653	0.478	0.175
40	38854	0.731	0.557	0.174
41	38378	0.791	0.618	0.173
42	39846	0.867	0.696	0.171
43	38616	0.678	0.507	0.170
44	39015	0.752	0.581	0.170
45	38867	0.872	0.702	0.170
46	39972	0.812	0.643	0.170
47	39132	0.803	0.635	0.168
48	39499	0.691	0.524	0.166
49	39949	0.560	0.394	0.166
50	39131	0.453	0.287	0.166
51	39063	0.774	0.611	0.163
52	39110	0.889	0.727	0.163
53	39354	0.812	0.650	0.162
54	38402	0.571	0.409	0.161
55	37812	0.713	0.552	0.160
56	38760	0.639	0.479	0.160
57	38465	0.507	0.348	0.160
58	38769	0.803	0.644	0.159
59	37268	0.721	0.563	0.158
60	39841	0.715	0.557	0.158
61	39378	0.729	0.571	0.158
62	39818	0.572	0.415	0.158
63	38851	0.566	0.409	0.157
64	38731	0.550	0.393	0.157
65	38531	0.724	0.567	0.157
66	38793	0.766	0.609	0.157
67	39921	0.671	0.514	0.157
68	39119	0.785	0.628	0.157
69	38744	0.618	0.462	0.156
70	38808	0.755	0.598	0.156
71	38664	0.550	0.394	0.156
72	39861	0.817	0.661	0.155
73	37802	0.822	0.667	0.155
74	38746	0.864	0.710	0.154

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		White	Hispanic	
75	39449	0.753	0.599	0.154
76	37794	0.716	0.563	0.154
77	38483	0.677	0.524	0.153
78	38785	0.697	0.544	0.152
79	37942	0.646	0.494	0.152
80	38520	0.743	0.592	0.152
81	38737	0.649	0.497	0.152
82	38442	0.854	0.703	0.152
83	39108	0.608	0.456	0.151
84	39389	0.589	0.438	0.151
85	38431	0.710	0.559	0.151
86	39370	0.817	0.666	0.151
87	38833	0.443	0.293	0.150
88	38362	0.754	0.604	0.150
89	39877	0.543	0.394	0.149
90	38602	0.693	0.545	0.148
91	39317	0.598	0.450	0.148
92	39116	0.759	0.612	0.147
93	39105	0.761	0.614	0.147
94	39875	0.727	0.580	0.146
95	39379	0.678	0.532	0.146
96	37602	0.647	0.501	0.146
97	39325	0.736	0.590	0.146
98	39805	0.693	0.548	0.145
99	38529	0.625	0.481	0.144
100	38423	0.587	0.443	0.144

Table A-3

Top 100 Items with Largest Informal DIF Statistics for the Gender Comparison Group

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		Male	Female	
1	38709	0.700	0.815	0.115
2	39834	0.533	0.647	0.114
3	36786	0.291	0.178	0.113
4	37671	0.567	0.679	0.112
5	39763	0.678	0.776	0.098
6	39493	0.709	0.806	0.096
7	38460	0.619	0.712	0.093
8	39748	0.751	0.843	0.091
9	38629	0.697	0.787	0.090
10	38459	0.703	0.792	0.089
11	38599	0.469	0.380	0.089
12	39865	0.579	0.668	0.088
13	39016	0.487	0.575	0.087
14	39880	0.582	0.669	0.087
15	36535	0.854	0.940	0.086
16	39972	0.700	0.785	0.085
17	38817	0.514	0.599	0.084
18	38604	0.553	0.637	0.084
19	39390	0.631	0.714	0.084
20	37993	0.646	0.729	0.083
21	39478	0.708	0.792	0.083
22	38350	0.760	0.842	0.083
23	39864	0.586	0.668	0.082
24	38672	0.602	0.683	0.081
25	38763	0.557	0.637	0.080
26	39735	0.716	0.796	0.080
27	39850	0.643	0.722	0.079
28	38377	0.706	0.785	0.079
29	38581	0.601	0.680	0.079
30	38764	0.691	0.769	0.078
31	38391	0.741	0.819	0.078
32	38805	0.682	0.760	0.078
33	37794	0.613	0.691	0.078
34	39464	0.773	0.850	0.077
35	39903	0.564	0.640	0.076

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		Male	Female	
36	38393	0.519	0.594	0.075
37	39863	0.524	0.599	0.075
38	39456	0.668	0.742	0.075
39	38491	0.574	0.648	0.074
40	37894	0.637	0.711	0.074
41	38864	0.638	0.712	0.073
42	38607	0.667	0.740	0.073
43	38033	0.744	0.817	0.073
44	39132	0.701	0.774	0.073
45	39374	0.525	0.598	0.073
46	39866	0.525	0.597	0.073
47	38602	0.589	0.661	0.072
48	38446	0.758	0.830	0.072
49	39467	0.836	0.908	0.072
50	37767	0.725	0.797	0.072
51	38829	0.531	0.603	0.072
52	39385	0.817	0.888	0.071
53	38521	0.777	0.848	0.071
54	39120	0.590	0.661	0.071
55	39483	0.722	0.793	0.071
56	38238	0.597	0.667	0.071
57	39477	0.632	0.703	0.070
58	39855	0.690	0.760	0.070
59	39883	0.433	0.503	0.070
60	37959	0.636	0.706	0.070
61	39969	0.737	0.807	0.070
62	39017	0.713	0.783	0.070
63	38368	0.816	0.886	0.070
64	39380	0.638	0.708	0.070
65	39949	0.452	0.522	0.070
66	37133	0.721	0.790	0.069
67	38808	0.666	0.735	0.068
68	39389	0.493	0.561	0.068
69	38640	0.675	0.743	0.068
70	39833	0.463	0.531	0.068
71	38684	0.797	0.730	0.068
72	37906	0.646	0.714	0.068
73	38771	0.574	0.642	0.068
74	38359	0.562	0.630	0.067
75	39373	0.538	0.605	0.067

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		Male	Female	
76	37975	0.745	0.812	0.067
77	39010	0.567	0.499	0.067
78	38452	0.715	0.782	0.067
79	39018	0.737	0.804	0.067
80	38157	0.691	0.758	0.067
81	39848	0.700	0.766	0.067
82	37753	0.726	0.792	0.066
83	39485	0.380	0.446	0.066
84	39730	0.771	0.837	0.066
85	38614	0.578	0.645	0.066
86	39490	0.566	0.632	0.066
87	38246	0.760	0.826	0.066
88	38601	0.762	0.696	0.066
89	39372	0.525	0.591	0.065
90	39008	0.380	0.445	0.065
91	38836	0.658	0.723	0.065
92	39452	0.654	0.719	0.065
93	39986	0.631	0.695	0.065
94	38605	0.808	0.873	0.065
95	39944	0.579	0.643	0.065
96	39128	0.724	0.788	0.064
97	39961	0.705	0.769	0.064
98	38717	0.737	0.801	0.064
99	37770	0.825	0.889	0.064
100	38559	0.414	0.478	0.064

APPENDIX C

Table C-1

Items with Significant Results from Black-White Mantel-Haenszel DIF Procedure

	Ability Quartile			
Count	1ST	2ND	3RD	4TH
1	39972	39972	39874	39818
2	39949	39936	39836	39132
3	39936	39894	39818	38825
4	39894	39874	39370	38496
5	39874	39859	39304	38437
6	39861	39818	39065	38408
7	39856	39370	38825	-
8	39835	38851	38806	-
9	39370	38778	38626	-
10	39354	38626	38601	-
11	39352	38601	38496	-
12	39304	38532	-	-
13	39112	38496	-	-
14	39065	38213	-	-
15	38601	-	-	-
16	37795	-	-	-

Table C-2

Items with Significant Results from Hispanic-White Mantel-Haenszel DIF Procedure

	Ability Quartile			
Count	1ST	2ND	3RD	4TH
1	40031	39883	39846	38437
2	39949	39856	39317	38423
3	39894	39304	39132	37677
4	39874	39131	39065	37602
5	39861	39065	38825	-
6	39856	38833	38483	-
7	39352	38825	38431	-
8	39304	38823	38402	-
9	39110	38744	-	-
10	39065	38489	-	-
11	38867	38442	-	-
12	38825	-	-	-
13	38815	-	-	-
14	37861	-	-	-

Table C-3

Items with Significant Results from Gender Mantel-Haenszel DIF Procedure

	Ability Quartile			
Count	1ST	2ND	3RD	4TH
1	36535	36535	36535	36786
2	36786	36786	36786	38459
3	37133	37133	37133	38460
4	37671	37671	37671	38491
5	37753	37767	37794	38629
6	37767	37770	37959	38763
7	37770	37794	37993	38829
8	37894	37906	38157	38836
9	37959	37975	38246	39008
10	37975	37993	38377	39373
11	38033	38033	38391	39485
12	38157	38246	38393	39490
13	38238	38350	38452	39493
14	38350	38368	38459	39735
15	38359	38377	38460	39834
16	38368	38391	38491	39865
17	38377	38446	38521	39880
18	38391	38452	38599	-
19	38393	38459	38602	-
20	38446	38460	38604	-
21	38452	38521	38607	-
22	38459	38581	38629	-
23	38460	38599	38672	-
24	38491	38601	38684	-
25	38521	38602	38709	-
26	38559	38604	38763	-
27	38581	38605	38808	-
28	38599	38614	38817	-
29	38601	38629	38836	-
30	38604	38640	39008	-
31	38605	38672	39010	-
32	38614	38684	39016	-
33	38629	38709	39128	-
34	38640	38717	39132	-
35	38672	38764	39373	-
36	38684	38771	39374	-
37	38709	38805	39385	-
38	38717	38817	39389	-
39	38763	38829	39390	-
40	38764	39016	39452	-

Count	Ability Quartile			
	1ST	2ND	3RD	4TH
41	38805	39120	39464	-
42	38808	39372	39478	-
43	38836	39374	39483	-
44	38864	39380	39485	-
45	39016	39385	39490	-
46	39017	39389	39493	-
47	39018	39390	39748	-
48	39120	39456	39763	-
49	39128	39464	39833	-
50	39374	39467	39834	-
51	39380	39477	39848	-
52	39385	39478	39863	-
53	39389	39483	39864	-
54	39390	39485	39865	-
55	39452	39493	39866	-
56	39456	39730	39883	-
57	39464	39735	39944	-
58	39467	39748	39949	-
59	39477	39763	39961	-
60	39478	39833	39972	-
61	39483	39834	39986	-
62	39493	39848	-	-
63	39730	39850	-	-
64	39735	39855	-	-
65	39748	39863	-	-
66	39763	39864	-	-
67	39833	39865	-	-
68	39834	39866	-	-
69	39848	39880	-	-
70	39850	39883	-	-
71	39855	39944	-	-
72	39863	39949	-	-
73	39865	39972	-	-
74	39866	39986	-	-
75	39880	-	-	-
76	39903	-	-	-
77	39944	-	-	-
78	39949	-	-	-
79	39961	-	-	-
80	39969	-	-	-
81	39972	-	-	-

APPENDIX D

Table D-1

Logistic Regression DIF Results for the Top 100 Items from the White-Black Comparison Group

	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
Item ID	χ^2	ρ	χ^2	ρ	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
36572	4.396	0.036	12.715	0.002	0.002	0.002
36786	9.468	0.002	10.546	0.005	0.004	0.000
37239	2.746	0.097	9.082	0.011	0.002	0.004
37671	0.001	0.975	2.872	0.238	0.000	0.001
37795	6.185	0.013	15.120	0.001	0.005	0.006
37797	1.984	0.159	3.462	0.177	0.001	0.001
37802	0.172	0.678	0.257	0.879	0.000	0.000
37805	7.223	0.007	7.452	0.024	0.005	0.000
37813	6.318	0.012	6.341	0.042	0.003	0.000
37824	4.801	0.028	5.450	0.066	0.002	0.000
37876	8.386	0.004	8.537	0.014	0.006	0.000
37973	2.027	0.155	2.130	0.345	0.001	0.000
38213	6.879	0.009	13.762	0.001	0.005	0.005
38247	6.155	0.013	6.424	0.040	0.002	0.000
38260	7.425	0.006	7.843	0.020	0.005	0.000
38362	7.623	0.006	14.049	0.001	0.004	0.004
38375	1.204	0.273	2.833	0.243	0.001	0.001
38378	3.553	0.059	3.572	0.168	0.001	0.000
38392	6.508	0.011	7.802	0.020	0.005	0.001
38400	4.975	0.026	5.143	0.076	0.002	0.000
38408	24.425	0.000	25.958	0.000	0.014	0.001
38420	4.254	0.039	4.543	0.103	0.002	0.001
38437	25.922	0.000	26.008	0.000	0.010	0.000
38449	14.387	0.000	14.723	0.001	0.006	0.000
38465	5.963	0.015	6.691	0.035	0.003	0.000
38489	18.511	0.000	20.200	0.000	0.008	0.001
38496	28.189	0.000	33.864	0.000	0.010	0.001
38524	5.056	0.025	5.062	0.080	0.002	0.000
38532	16.200	0.000	32.664	0.000	0.007	0.007
38537	1.992	0.158	4.905	0.086	0.001	0.002

Item ID	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
	χ^2	ρ	χ^2	ρ	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
38564	10.913	0.001	11.118	0.004	0.006	0.000
38588	3.621	0.057	4.767	0.092	0.003	0.001
38597	11.887	0.001	12.034	0.002	0.007	0.000
38599	5.104	0.024	5.376	0.068	0.004	0.000
38601	41.194	0.000	41.761	0.000	0.028	0.001
38616	6.120	0.013	13.659	0.001	0.004	0.005
38626	17.072	0.000	21.353	0.000	0.007	0.002
38628	1.418	0.234	5.220	0.074	0.001	0.002
38632	6.361	0.012	6.975	0.031	0.003	0.000
38637	4.423	0.035	4.561	0.102	0.002	0.001
38664	6.462	0.011	9.985	0.007	0.002	0.002
38743	2.415	0.120	3.836	0.147	0.001	0.001
38761	6.777	0.009	7.213	0.027	0.004	0.001
38766	14.202	0.000	20.118	0.000	0.010	0.005
38767	4.623	0.032	6.864	0.032	0.003	0.001
38778	5.474	0.019	5.506	0.064	0.004	0.000
38795	9.079	0.003	10.175	0.006	0.005	0.001
38806	5.385	0.020	13.563	0.001	0.003	0.004
38812	10.116	0.001	10.192	0.006	0.006	0.000
38814	4.192	0.041	4.193	0.123	0.002	0.000
38815	10.734	0.001	13.942	0.001	0.007	0.003
38825	10.219	0.001	10.267	0.006	0.003	0.000
38851	16.247	0.000	28.011	0.000	0.007	0.005
38855	6.786	0.009	7.633	0.022	0.004	0.001
39013	3.744	0.053	12.100	0.002	0.002	0.005
39065	33.152	0.000	34.462	0.000	0.018	0.000
39066	5.307	0.021	6.906	0.032	0.003	0.001
39067	1.842	0.175	4.090	0.129	0.001	0.002
39071	2.966	0.085	3.526	0.172	0.002	0.000
39104	15.961	0.000	17.435	0.000	0.010	0.001
39105	4.595	0.032	4.625	0.099	0.002	0.000
39112	6.851	0.009	10.767	0.005	0.005	0.003
39120	5.869	0.015	6.668	0.036	0.003	0.001
39132	15.551	0.000	28.021	0.000	0.007	0.005
39304	43.055	0.000	45.331	0.000	0.022	0.001
39338	0.014	0.906	0.154	0.926	0.000	0.000
39352	15.502	0.000	18.126	0.000	0.007	0.001
39354	9.902	0.002	11.467	0.003	0.004	0.001

Item ID	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
	χ^2	ρ	χ^2	ρ	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
39358	7.213	0.007	9.590	0.008	0.003	0.002
39359	4.863	0.027	8.578	0.014	0.002	0.002
39370	59.443	0.000	59.629	0.000	0.025	0.000
39382	3.683	0.055	6.700	0.035	0.002	0.001
39387	6.007	0.014	8.634	0.013	0.003	0.001
39390	1.224	0.269	3.390	0.184	0.000	0.001
39469	5.227	0.022	7.058	0.029	0.002	0.001
39494	4.882	0.027	4.976	0.083	0.002	0.000
39818	35.161	0.000	39.853	0.000	0.013	0.002
39824	7.323	0.007	7.804	0.020	0.003	0.000
39825	6.876	0.009	15.729	0.000	0.003	0.004
39832	6.765	0.009	6.893	0.032	0.002	0.000
39833	0.201	0.654	0.201	0.904	0.000	0.000
39835	15.065	0.000	15.065	0.001	0.006	0.000
39836	26.571	0.000	31.708	0.000	0.008	0.002
39841	0.745	0.388	0.745	0.689	0.000	0.000
39846	11.071	0.001	16.023	0.000	0.009	0.004
39856	26.538	0.000	31.869	0.000	0.012	0.002
39859	6.753	0.009	12.706	0.002	0.002	0.002
39861	15.875	0.000	20.986	0.000	0.010	0.003
39862	1.507	0.220	3.388	0.184	0.000	0.001
39874	50.898	0.000	58.210	0.000	0.046	0.007
39875	4.115	0.043	6.093	0.048	0.003	0.002
39882	8.015	0.005	8.018	0.018	0.004	0.000
39883	7.111	0.008	7.267	0.026	0.003	0.000
39894	54.688	0.000	56.942	0.000	0.030	0.001
39936	29.927	0.000	30.025	0.000	0.012	0.000
39944	0.022	0.882	0.256	0.880	0.000	0.000
39949	17.223	0.000	17.525	0.000	0.008	0.000
39972	9.836	0.002	26.502	0.000	0.004	0.006
40031	0.445	0.505	2.047	0.359	0.000	0.001
40032	0.001	0.975	0.182	0.913	0.000	0.000

Table D-2

Logistic Regression DIF Results for the Top 100 Items from the White-Hispanic Comparison Group

	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
Item ID	χ^2	ρ	χ^2	ρ	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
36786	15.879	0.000	16.900	0.000	0.006	0.006
37239	1.123	0.289	2.234	0.327	0.001	0.000
37268	5.158	0.023	5.719	0.057	0.003	0.001
37602	4.295	0.038	4.297	0.117	0.003	0.001
37677	11.847	0.001	18.703	0.000	0.006	0.000
37794	0.640	0.424	1.976	0.372	0.000	0.000
37802	0.101	0.751	0.442	0.802	0.000	0.000
37812	2.348	0.125	3.759	0.153	0.002	0.003
37861	39.150	0.000	39.299	0.000	0.024	0.001
37942	1.171	0.279	1.222	0.543	0.000	0.002
38362	2.212	0.137	2.346	0.309	0.002	0.000
38373	9.466	0.002	16.233	0.000	0.006	0.002
38378	1.139	0.286	2.208	0.332	0.001	0.002
38392	4.017	0.045	9.991	0.007	0.003	0.001
38402	10.921	0.001	11.954	0.003	0.003	0.001
38423	16.070	0.000	17.571	0.000	0.007	0.000
38431	9.099	0.003	9.127	0.010	0.004	0.002
38437	15.419	0.000	19.085	0.000	0.005	0.002
38442	16.168	0.000	16.448	0.000	0.007	0.000
38465	7.722	0.005	13.236	0.001	0.004	0.002
38483	13.244	0.000	19.082	0.000	0.004	0.002
38489	34.982	0.000	40.266	0.000	0.013	0.001
38496	3.276	0.070	3.817	0.148	0.001	0.001
38520	4.087	0.043	4.513	0.105	0.002	0.001
38529	0.002	0.964	0.146	0.930	0.000	0.002
38531	0.199	0.656	5.717	0.057	0.000	0.004
38532	17.902	0.000	19.149	0.000	0.006	0.002
38597	7.045	0.008	13.489	0.001	0.004	0.000
38602	0.287	0.592	0.387	0.824	0.000	0.002
38616	5.343	0.021	12.156	0.002	0.003	0.000

Item ID	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
	χ^2	ρ	χ^2	ρ	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
38664	2.092	0.148	2.771	0.250	0.001	0.000
38725	23.622	0.000	23.808	0.000	0.010	0.001
38731	4.388	0.036	4.533	0.104	0.002	0.001
38737	7.632	0.006	9.341	0.009	0.004	0.004
38744	7.052	0.008	11.477	0.003	0.003	0.002
38746	12.407	0.000	16.359	0.000	0.005	0.002
38760	5.087	0.024	5.109	0.078	0.002	0.004
38766	21.685	0.000	22.711	0.000	0.015	0.000
38769	5.332	0.021	5.355	0.069	0.003	0.003
38785	1.700	0.192	2.714	0.257	0.001	0.002
38793	9.909	0.002	10.832	0.004	0.005	0.002
38795	29.477	0.000	37.873	0.000	0.015	0.000
38808	0.199	0.656	0.479	0.787	0.000	0.001
38815	9.302	0.002	9.318	0.009	0.007	0.000
38823	35.001	0.000	45.534	0.000	0.011	0.001
38825	38.161	0.000	38.202	0.000	0.011	0.001
38826	35.117	0.000	41.034	0.000	0.010	0.001
38833	5.966	0.015	12.944	0.002	0.002	0.000
38851	3.098	0.078	4.736	0.094	0.002	0.002
38854	9.529	0.002	10.732	0.005	0.007	0.002
38867	5.147	0.023	5.579	0.061	0.004	0.002
39015	2.045	0.153	4.200	0.122	0.001	0.000
39063	4.539	0.033	7.811	0.020	0.002	0.001
39065	87.350	0.000	87.370	0.000	0.042	0.004
39066	12.240	0.000	17.243	0.000	0.005	0.000
39104	4.043	0.044	5.281	0.071	0.002	0.002
39105	0.409	0.522	3.455	0.178	0.000	0.002
39108	7.096	0.008	10.165	0.006	0.004	0.000
39110	7.231	0.007	7.742	0.021	0.005	0.001
39116	5.663	0.017	5.699	0.058	0.004	0.001
39119	6.015	0.014	11.158	0.004	0.003	0.002
39131	23.252	0.000	26.166	0.000	0.010	0.000
39132	11.853	0.001	14.874	0.001	0.004	0.001
39304	31.313	0.000	32.594	0.000	0.014	0.000
39305	17.578	0.000	18.107	0.000	0.006	0.000
39317	14.045	0.000	16.566	0.000	0.006	0.001
39325	5.965	0.015	6.899	0.032	0.002	0.001

Item ID	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
	χ^2	ρ	χ^2	ρ	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
39352	28.578	0.000	29.983	0.000	0.010	0.002
39354	6.609	0.010	6.890	0.032	0.003	0.002
39370	1.403	0.236	3.093	0.213	0.001	0.001
39378	11.181	0.001	11.195	0.004	0.005	0.001
39379	4.553	0.033	4.715	0.095	0.002	0.000
39387	4.582	0.032	6.452	0.040	0.002	0.000
39389	4.100	0.043	4.217	0.121	0.002	0.002
39449	4.333	0.037	6.511	0.039	0.002	0.001
39499	14.506	0.000	15.278	0.000	0.006	0.001
39805	7.373	0.007	8.100	0.017	0.006	0.003
39818	3.930	0.047	4.045	0.132	0.002	0.002
39824	2.165	0.141	2.319	0.314	0.000	0.000
39825	6.270	0.012	8.284	0.016	0.003	0.000
39833	1.575	0.209	2.691	0.260	0.001	0.000
39836	0.890	0.345	2.955	0.228	0.000	0.001
39841	2.595	0.107	7.647	0.022	0.002	0.000
39846	8.476	0.004	17.176	0.000	0.008	0.005
39856	82.888	0.000	89.645	0.000	0.037	0.001
39861	4.257	0.039	9.438	0.009	0.003	0.001
39862	1.835	0.176	2.260	0.323	0.000	0.001
39868	5.165	0.023	7.926	0.019	0.005	0.001
39874	29.693	0.000	31.500	0.000	0.031	0.004
39875	0.049	0.825	0.106	0.948	0.000	0.001
39877	0.467	0.494	2.760	0.252	0.000	0.001
39883	21.316	0.000	23.010	0.000	0.009	0.001
39894	32.117	0.000	34.714	0.000	0.018	0.003
39921	8.459	0.004	8.899	0.012	0.004	0.001
39936	19.028	0.000	23.272	0.000	0.006	0.002
39944	0.643	0.423	3.285	0.193	0.000	0.001
39949	3.465	0.063	9.232	0.010	0.001	0.000
39965	6.178	0.013	8.040	0.018	0.005	0.000
39972	6.216	0.013	12.570	0.002	0.003	0.002
40031	9.066	0.003	9.068	0.011	0.005	0.001

Table D-3

Logistic Regression DIF Results for the Top 100 Items from the Gender Comparison Group

Item ID	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
	χ^2	P	χ^2	ρ	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
36535	241.654	0.000	284.753	0.000	0.039	0.006
36786	158.548	0.000	158.614	0.000	0.018	0.000
37133	48.515	0.000	53.385	0.000	0.010	0.001
37671	108.599	0.000	115.188	0.000	0.014	0.001
37753	51.693	0.000	53.923	0.000	0.009	0.000
37767	49.711	0.000	49.848	0.000	0.011	0.000
37770	52.420	0.000	52.494	0.000	0.013	0.000
37794	52.797	0.000	69.318	0.000	0.008	0.003
37894	47.215	0.000	48.115	0.000	0.009	0.001
37906	33.039	0.000	42.756	0.000	0.007	0.002
37959	62.340	0.000	62.755	0.000	0.008	0.000
37975	29.694	0.000	39.405	0.000	0.005	0.002
37993	51.630	0.000	64.257	0.000	0.008	0.002
38033	53.200	0.000	54.023	0.000	0.011	0.001
38157	41.798	0.000	51.194	0.000	0.006	0.001
38238	34.686	0.000	34.898	0.000	0.004	0.000
38246	43.263	0.000	51.241	0.000	0.006	0.002
38350	73.219	0.000	82.603	0.000	0.016	0.002
38359	35.752	0.000	35.942	0.000	0.008	0.000
38368	54.064	0.000	62.379	0.000	0.013	0.002
38377	87.716	0.000	104.080	0.000	0.011	0.002
38391	74.721	0.000	81.199	0.000	0.012	0.001
38393	32.316	0.000	34.314	0.000	0.007	0.001
38446	46.315	0.000	49.823	0.000	0.011	0.001
38452	60.604	0.000	74.115	0.000	0.008	0.002
38459	107.634	0.000	137.034	0.000	0.016	0.004
38460	89.111	0.000	99.655	0.000	0.013	0.002
38491	82.748	0.000	85.331	0.000	0.008	0.000
38521	65.059	0.000	82.149	0.000	0.010	0.002
38559	25.641	0.000	26.603	0.000	0.006	0.000
38581	68.165	0.000	69.602	0.000	0.009	0.000
38599	52.336	0.000	53.338	0.000	0.010	0.001

38601	33.999	0.000	35.563	0.000	0.007	0.001
38602	40.671	0.000	61.439	0.000	0.008	0.004
38604	94.388	0.000	117.401	0.000	0.011	0.002
38605	53.364	0.000	61.523	0.000	0.011	0.002
38607	34.822	0.000	55.011	0.000	0.007	0.004
38614	32.725	0.000	32.898	0.000	0.007	0.000
38629	83.845	0.000	99.623	0.000	0.019	0.003
38640	50.249	0.000	62.526	0.000	0.008	0.002
38672	77.109	0.000	100.057	0.000	0.007	0.002
38684	137.579	0.000	138.233	0.000	0.014	0.000
38709	108.891	0.000	110.785	0.000	0.021	0.001
38717	42.956	0.000	43.276	0.000	0.008	0.000
38763	53.187	0.000	63.198	0.000	0.009	0.001
38764	52.521	0.000	54.776	0.000	0.008	0.001
38771	34.124	0.000	39.301	0.000	0.006	0.001
38805	48.138	0.000	48.149	0.000	0.008	0.000
38808	45.837	0.000	55.526	0.000	0.008	0.002
38817	51.291	0.000	59.906	0.000	0.011	0.002
38829	51.532	0.000	76.223	0.000	0.005	0.002
38836	53.694	0.000	57.368	0.000	0.006	0.000
38864	36.494	0.000	37.899	0.000	0.007	0.001
39008	34.014	0.000	65.960	0.000	0.006	0.004
39010	31.052	0.000	33.160	0.000	0.006	0.000
39016	60.978	0.000	67.539	0.000	0.011	0.002
39017	37.208	0.000	44.966	0.000	0.008	0.002
39018	35.524	0.000	35.525	0.000	0.008	0.000
39120	38.965	0.000	46.148	0.000	0.007	0.001
39128	44.922	0.000	52.208	0.000	0.006	0.001
39132	43.248	0.000	59.767	0.000	0.005	0.002
39372	47.515	0.000	48.129	0.000	0.006	0.000
39373	41.705	0.000	53.992	0.000	0.005	0.001
39374	51.249	0.000	53.197	0.000	0.007	0.000
39380	49.884	0.000	55.175	0.000	0.007	0.000
39385	103.877	0.000	110.056	0.000	0.017	0.001
39389	44.063	0.000	46.257	0.000	0.006	0.001
39390	94.891	0.000	116.743	0.000	0.010	0.002
39452	50.082	0.000	62.906	0.000	0.005	0.002
39456	60.634	0.000	63.494	0.000	0.008	0.001
39464	69.494	0.000	77.166	0.000	0.015	0.001
39467	88.412	0.000	88.437	0.000	0.015	0.000
39477	61.423	0.000	63.550	0.000	0.008	0.000
39478	121.506	0.000	137.185	0.000	0.016	0.002

39483	60.251	0.000	73.277	0.000	0.008	0.001
39485	52.232	0.000	63.149	0.000	0.007	0.001
39490	42.884	0.000	69.083	0.000	0.006	0.003
39493	116.590	0.000	133.004	0.000	0.015	0.002
39730	62.342	0.000	63.877	0.000	0.009	0.000
39735	74.673	0.000	74.988	0.000	0.009	0.000
39748	143.662	0.000	149.513	0.000	0.022	0.000
39763	138.764	0.000	141.364	0.000	0.018	0.001
39833	60.591	0.000	60.608	0.000	0.006	0.000
39834	157.856	0.000	200.254	0.000	0.020	0.005
39848	43.144	0.000	51.227	0.000	0.009	0.001
39850	42.619	0.000	46.586	0.000	0.008	0.001
39855	49.002	0.000	53.637	0.000	0.007	0.001
39863	65.418	0.000	71.311	0.000	0.008	0.001
39864	90.250	0.000	119.361	0.000	0.011	0.004
39865	89.119	0.000	96.515	0.000	0.012	0.001
39866	65.139	0.000	74.052	0.000	0.009	0.001
39880	54.391	0.000	60.318	0.000	0.011	0.001
39883	55.638	0.000	76.433	0.000	0.007	0.003
39903	36.542	0.000	40.943	0.000	0.007	0.001
39944	69.525	0.000	82.586	0.000	0.007	0.002
39949	55.813	0.000	62.152	0.000	0.007	0.001
39961	36.747	0.000	36.755	0.000	0.007	0.000
39969	35.408	0.000	36.135	0.000	0.009	0.000
39972	103.738	0.000	124.124	0.000	0.013	0.002
39986	42.948	0.000	47.080	0.000	0.005	0.001

APPENDIX E

LOGISTIC REGRESSION DIF SYNTAX FOR SPSS V.19

GENDER ANALYSIS

```
LOGISTIC REGRESSION VARIABLES=ItemCorrect
/METHOD=ENTER iScaledScore /METHOD=ENTER GenderMale
/CONTRAST (GenderMale)=Indicator
/CRITERIA PIN(.05) POUR(.10) ITERATE(20) CUT(.5).
EXECUTE.
LOGISTIC REGRESSION VARIABLES=ItemCorrect
/METHOD=ENTER iScaledScore /METHOD=ENTER GenderMale GenderMale*iScaledScore
/CONTRAST (GenderMale)=Indicator
/CRITERIA PIN(.05) POUR(.10) ITERATE(20) CUT(.5).
EXECUTE.
```

RACE ANALYSIS

```
LOGISTIC REGRESSION VARIABLES=ItemCorrect
/METHOD=ENTER iScaledScore /METHOD=ENTER RaceID
/CONTRAST (RaceID)=Indicator
/CRITERIA PIN(.05) POUR(.10) ITERATE(20) CUT(.5).
EXECUTE.
LOGISTIC REGRESSION VARIABLES=ItemCorrect
/METHOD=ENTER iScaledScore /METHOD=ENTER RaceID RaceID*iScaledScore
/CONTRAST (RaceID)=Indicator
/CRITERIA PIN(.05) POUR(.10) ITERATE(20) CUT(.5).
EXECUTE.
```