Evaluating the Influence of Differential Item Functioning for Race and Gender on STAR Math 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 Math 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 Math, 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

Contents 2
Executive Summary 3
Introduction 8
  Star Math 8
    Design 8
    Test Administration Procedures 9
    Content Development 10
  Validity and Differential Item Functioning 11
    Uniform versus Non-Uniform DIF 13
  Purpose 16
Method 16
  Sample 17
  Variables 18
    Item Correct 18
    Gender 19
    Race 19
    STAR Math Scaled Score 20
    Assessment Items 20
  Procedure 21
    Reference and Focal Groups 21
    Step 1: Informal DIF Procedure 21
    Step 2: Mantel-Haenszel Procedure for Evaluating DIF 23
    Step 3: Logistic Regression Procedure 25
Results 28
    Informal DIF 28
    Mantel-Haenszel Procedure 29
    Logistic Regression 31
Limitations 33
Conclusion 34
References 36
Appendix A: Asian and Native American DIF Analyses 38
  Procedure 39
  Informal DIF Results 39
Logistic Regression 41

Conclusion 42
Appendix B 44
Appendix C 51
Appendix D 55
Appendix E 62
Logistic Regression DIF Syntax for SPSS v.19 63
Gender Analysis 63
Race Analysis 63

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 Math 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 Math 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

Items. The items of interest here were 554 new items written for use in STAR Math Enterprise, a new version of STAR Math designed as a standards-referenced math achievement test, STAR Math Enterorise items measure 4 major content domains, 36 general skill sets, and more than 500 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 Math tests administered nationwide.

Tests. Students took STAR Math tests in the normal course of their math studies. Each STAR Math test consisted of 24 adaptively administered scored test items, with another 3 to 5 randomly selected, unscored items embedded in the STAR Math test. Student scores on the STAR Math tests served as the criterion measure for the DIF analyses; the new items were not counted in calculating the STAR Math scores.

Data. A total of 4,748,414 cases (item response records) were included in the original data file, Gender information was recorded for 3,353,243 (70.6%) of those cases; demographic information on race was recorded for 1,583,164 (33.3%).

Data analyses. The purpose of the analyses was to systematically evaluate the prospective new STAR Math 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

The initial informal DIF analysis, which was limited to subgroups with large student samples (males vs. females, and white-black-Hispanic students) included all of the 554 items from the STAR Math 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 top 10% of items demonstrating with the largest p-value difference statistics were used for the remaining analyses.

The Mantel-Haenszel procedure was completed for three comparisons: White-Black, White-Hispanic, and Male-Female. Students were grouped into four quartiles, based on their STAR Math 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.

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 Math assessments with students from culturally and linguistically diverse backgrounds.

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 554 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 MATH

STAR Math is a computer-adaptive test assessment designed to provide educators with periodic measures of student achievement 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 Math assessments have two purposes, which are to provide: (a) data that allows for progress monitoring of student performance across time; and (b) a measure of math achievement relative to a national normative sample. For more about the advantages of these purposes see the STAR Math Technical Manual (Renaissance Learning Inc.; RLI, 2008).

DESIGN

STAR Math is a computer-adaptive test, which has the ability of taking 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 reliability between tests. 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 math 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.

The STAR Math assessment items are tailored to each individual by created an assessment that pulls 24 items from a large item bank. Items are selected from 8 content strands. The content strands *Computation Processes* and *Numeration Concepts* are weighed more heavily in the number of items that are presented to the student because of the represent fundamental math abilities. Specifically, 8 items from each of these strands are presented to the student for grades 1 through 12. The other content strands (*Algebra, Data Analysis and Statistics, Estimation, Geometry, Measurement, and Word Problems*) vary in the number of items that are presented to students, based on their grade and ability. These content strands cover 214 math learning objectives. The current version of the Renaissance Place STAR Math software pulls from an item bank totaling more than 2,000 items to create a 24-item test for a given student. For more about the design, improvements over time in STAR Math software, and test security procedures, see the STAR Reading Technical Manual (RLI, 2008).

Test Administration Procedures

The STAR Math 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 Math 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, 2008). 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 Math software, there is a feature built-in to the practice sessions that ensures 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 Math software uses adaptive branching; beginning with items that are relatively low relative to the students expected level of ability. The test administration involves 24 items and most students complete the assessment in less than 15 minutes (RLI, 2008). To prevent the over-use of similar items, the software keeps track of items that are presented to students and will not present the same item twice within a 75-day period.

Content Development

The content of the assessment items contained in the STAR Math program was developed for students from grades kindergarten through grade 12. This means that the ability ranges from a kindergarten level to beyond high school to account for students who have exceeded the norm to not reach a ceiling when their achievement is assessed. The content was designed to capture the typical curriculum provided to students in modern schools, with a focus on content delivered across the United States. A number of sources were reviewed in creating the content. For more information on these sources and on the aforementioned strands that provide the over-arching structure for the 214 math learning objectives, see the STAR Math Technical Manual (RLI, 2008).

The STAR Math 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 Math assessments. However, the item and scale calibration are outside the scope of this report – the analyses described in detail later are conducted on the entire item bank at a particular point in time (development of items is ongoing). 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 achievement , 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 fo 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 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 differences 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 Math that 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 Math 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 Math items appears to have significant implications for this assessment, given its computer-adaptive delivery system.

Uniform versus 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).

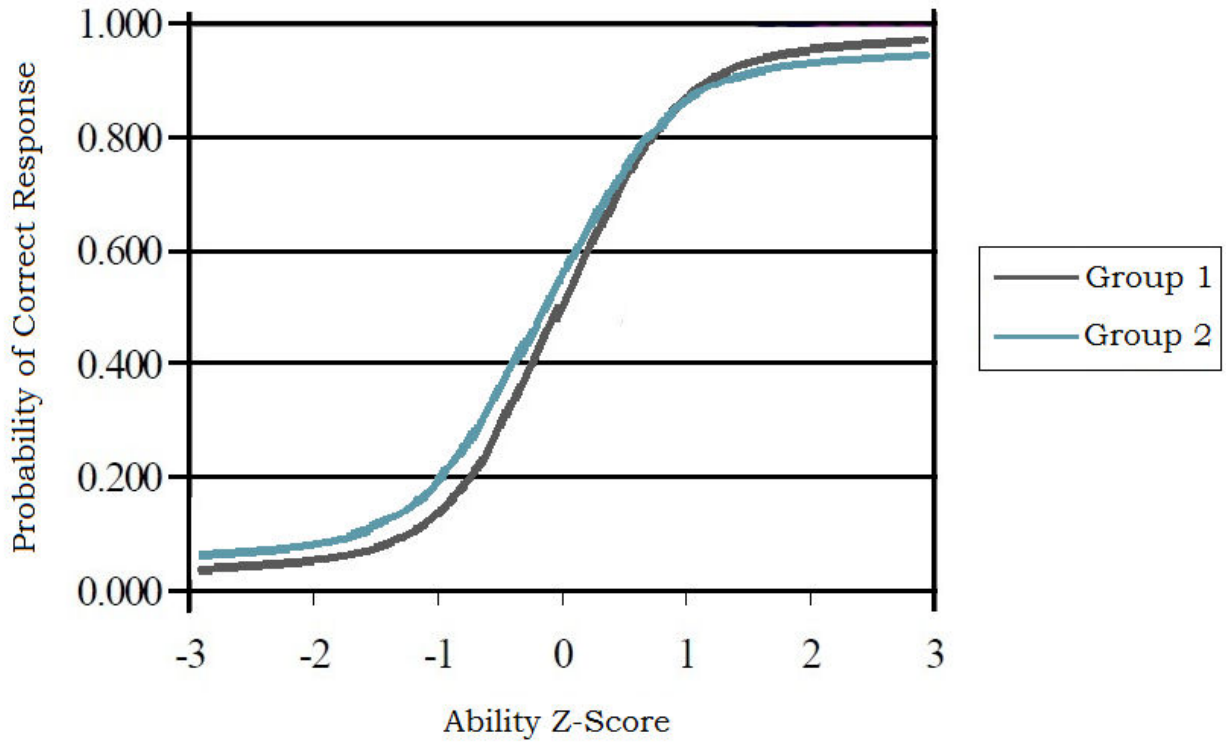


Figure 1. Item characteristic curves for reference and focal groups not demonstrating any type of DIF.

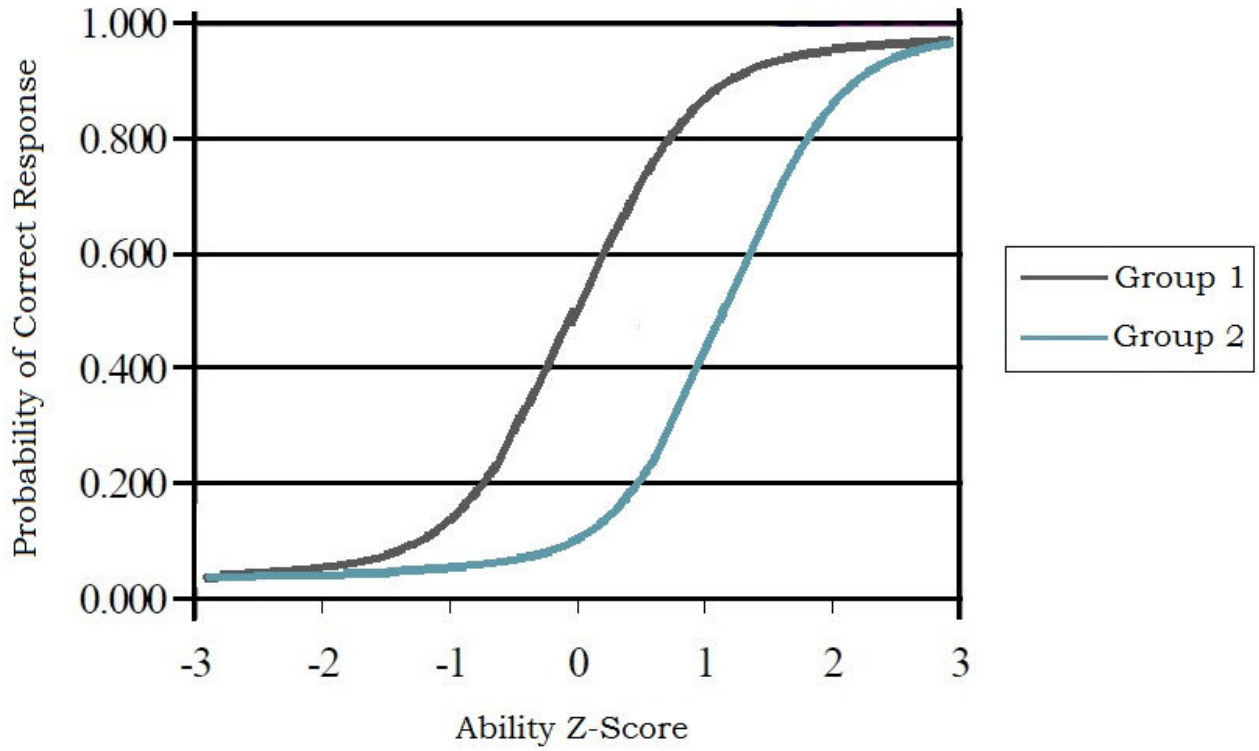


Figure 2. Item characteristic curves for reference and focal groups demonstrating uniform DIF.

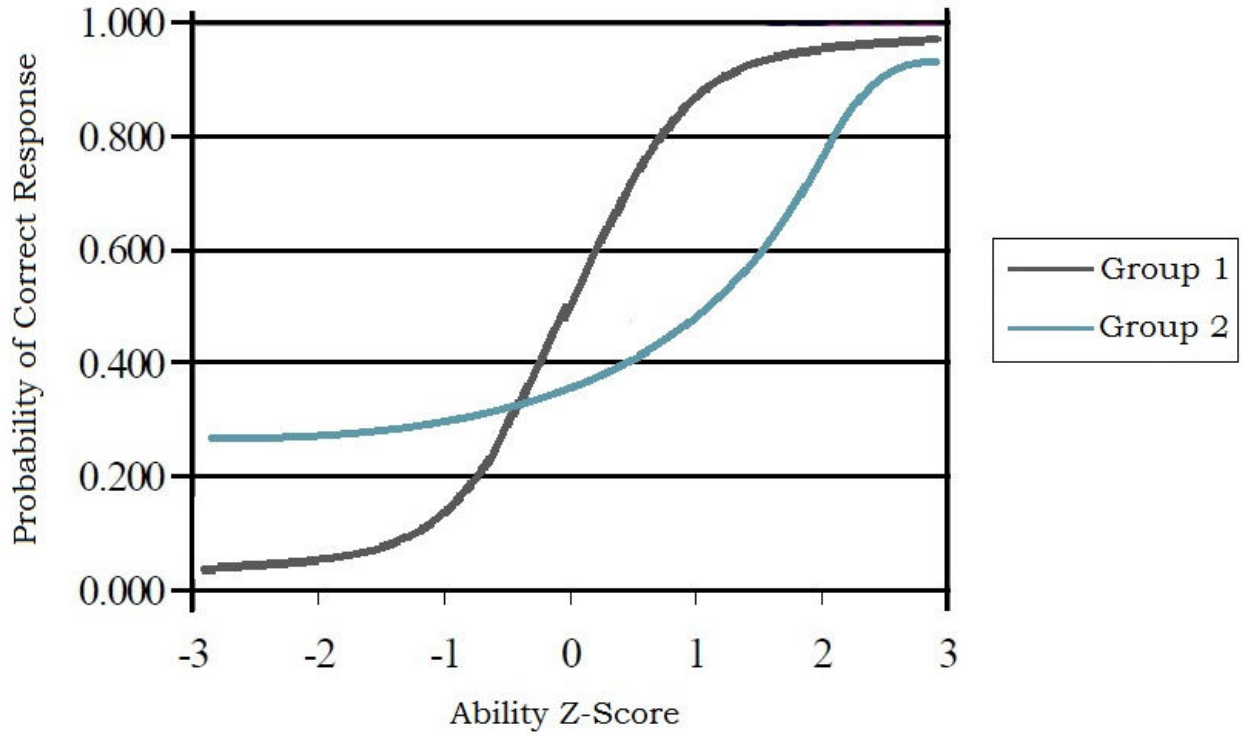


Figure 3. Item characteristic curves for reference and focal groups demonstrating non-uniform DIF.

PURPOSE

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

METHOD

SAMPLE

Star Math Batch 9 was used for analysis. Batch 9 consisted of 554 new items being evaluated for operational use; it went live on March 30, 2010. A total of 4,748,414 cases (item responses) were included in the original data file. The number of cases that had demographic information on gender was equal to 3,353,243 (70.6% of the total number of cases) and the number of cases that had demographic information on race was equal to 1,583,164 (33.3% 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 554. 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 this STAR Math batch.

Table 1

Sample Demographics

	Dataset Cases Summary		Sample Size Summary Statistics Across Items			
	Count[1]	Percent	Mean	SD	Min.	Max.
Gender						
Male	1,739,884	51.9	3,141	2,672	591	10,395
Female	1,613,359	48.1	2,912	2,573	422	9,911
Race						
White	749,038	47.3	1,352	1,192	185	4,578
Black	390,268	24.7	704	502	142	1,875
Hispanic	335,550	21.2	606	494	114	1,847
Asian	48,329	3.1	87	85	4	339
Native American	55,115	3.5	99	82	9	354

Note. The number of items that had race or gender demographic information is 554.

There were a number of cases that were filtered out of the data files during their original creation, which could be considered a 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.

Although the sample sizes do attain the threshold of minimum sample size for some items, the samples for the Asian and Native American 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. Fortunately, a number of subgroups (male, female, White, Black, and Hispanic) have satisfactory sample sizes to conduct comprehensive evaluations of the influence of DIF on assessment scores.

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 Math Scaled Score

STAR Math 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 Math Technical Manual (RLI, 2008). Star Math 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 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 Math data file. Therefore, a 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, the 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.

---

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

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:

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

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 55 items, which is roughly 10 percent of the items, demonstrating the greatest difference in p-value 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 based on their STAR Math scaled score. 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 ones 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 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:

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

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

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \tag{3}$$

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 Math 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:

---

[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

=(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)

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)

---

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.

[3] Nagelkerke r-squared

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

$$=(\text{R-squared for Step 3}) - (\text{R-squared for Step 2}) \qquad (8)$$

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).

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 554 items from the STAR Math dataset. Summary statistics from the p-value difference statistics can be found in Table 3. Overall, item p-value difference statistics 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 difference statistics equal or less than 0.001 across all comparisons. This indicates that the percentages of students in different demographic groups answering these items correctly were basically the same for these items. Despite some of the items demonstrating very low p-value difference statistics, others demonstrated a relatively high difference, with some p-value difference statistics as large as 0.25 between groups. As planned, the top 10% of items demonstrating with the largest p-value difference

statistics 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 D.

Table 3

Informal DIF P-Value Summary Statistics

	P-Value Statistics			
Comparison Groups (Reference-Focal)	*M*	*SD*	Min	Max
Complete Sample				
White-Black	0.081	0.053	0.000	0.260
White-Hispanic	0.057	0.043	0.000	0.230
Male-Female	0.027	0.022	0.000	0.150
Top 10% of Items with Largest P-Value Difference				
White-Black	0.186	0.023	0.160	0.260
White-Hispanic	0.146	0.027	0.120	0.230
Male-Female	0.073	0.023	0.050	0.150

Mantel-Haenszel Procedure

The Mantel-Haenszel procedure to assess DIF for STAR Math items was completed for three comparisons, which were White-Black, White-Hispanic, and Male-Female. The four ability quartiles allowed us to statistically control for ability to some extent. The results of this analysis are summarized in Figure 4 and shown in more detail in Appendix D.

Overall, there were items identified to be 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 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.

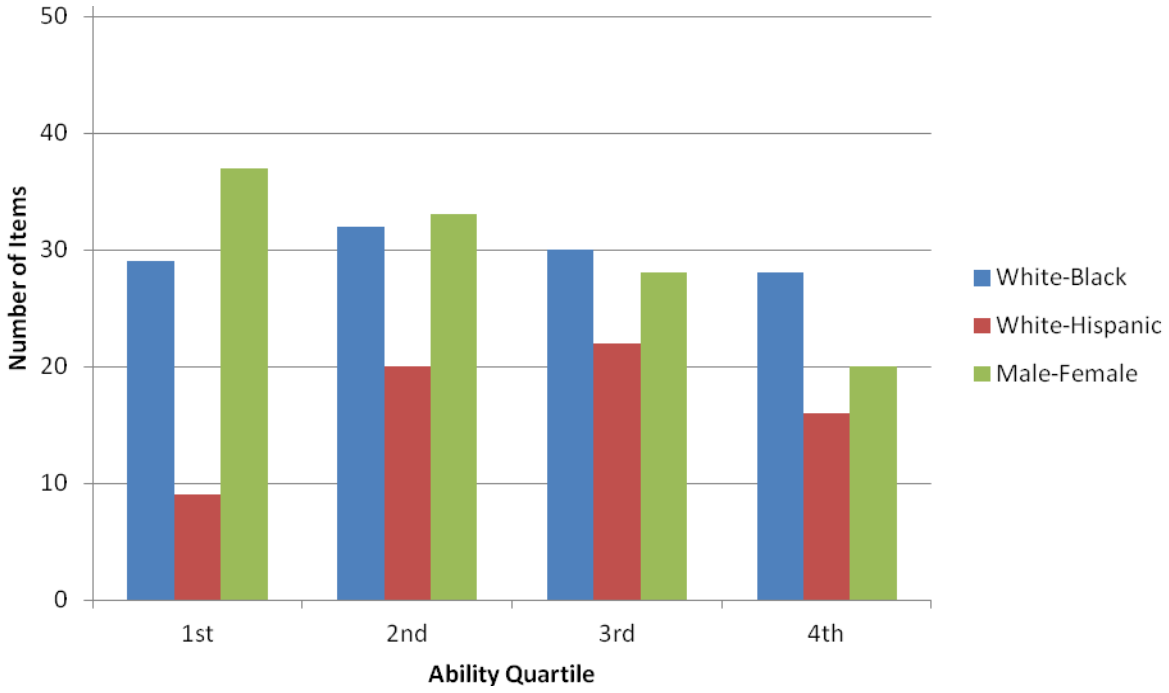


Figure 4. Items demonstrating statistically significant DIF values between comparison groups displayed by ability quartiles.

The results of the Mantel-Haenszel procedure suggest that some DIF may be occurring across group comparisons and ability levels. 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 in fact 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 5 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 5, a portion of the items for the White-Black and White-Hispanic comparisons were significantly demonstrating 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 trend seen in Figure 5 is somewhat consistent with those seen in the Mantel-Haenszel procedure.

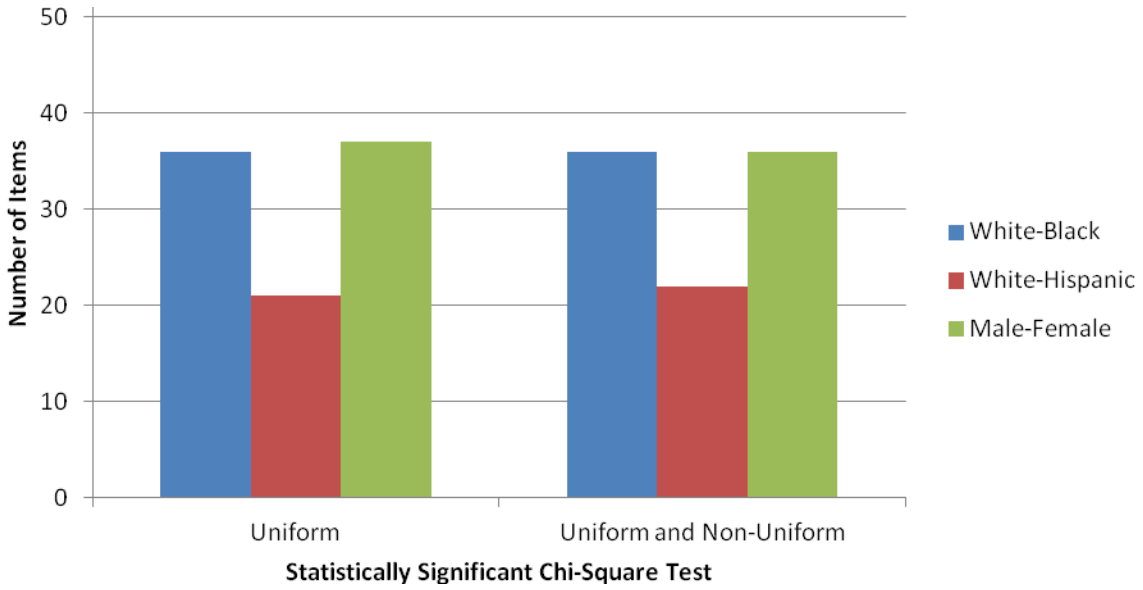
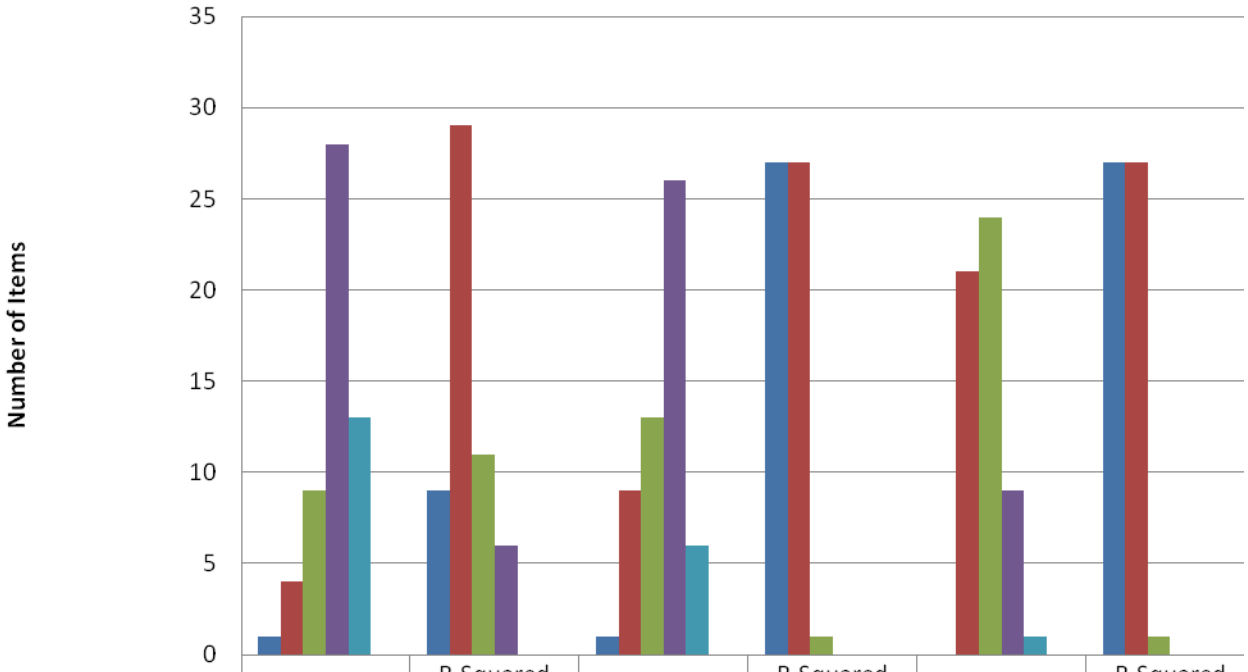


Figure 5. Number of items with significant chi-square values when testing for uniform DIF only and uniform and non-uniform DIF simultaneously.

In order to provide better context for the results seen in Figure 5, Figure 6 provides the Nagelkerke R-squared value ranges for all the items evaluated for uniform and non-uniform DIF by comparison groups. 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 is 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	1	9	1	27	0	27
.001-.005	4	29	9	27	21	27
.005-.010	9	11	13	1	24	1
.010-.025	28	6	26	0	9	0
.025-.050	13	0	6	0	1	0
Greater than .050	0	0	0	0	0	0

Figure 6. Summary of Nagelkerke R-squared values attributed to uniform and non-uniform DIF for 100 items from the STAR Math 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. The approach taken to include items that demonstrated the greatest difference in the probability of a correct answer between groups (i.e. informal DIF procedure) does allow for some degree of certainty in drawing inferences about the pool of items contained in the entire dataset. However, there are a number of items from other STAR Math batches that were not included in any of the analyses. Thus, there may be a need to replicate the procedures described here to gather more evidence on the extent to which DIF may be occurring between these groups and others.

CONCLUSION

In general, the analyses provided results suggesting that although DIF may be occurring for some of the items contained in the STAR Math 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 content analysis 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 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, given the number of items that appeared to be demonstrating DIF in preliminary analyses. However, given the information that was gathered through this analysis, it can be concluded that the influence of DIF on the assessment items included here is negligible; while there were some items for which DIF measures were statistically significant, there was little or no practical significance.

This process not only provides validity evidence for the STAR Math assessments, but also provides support for the item development process currently in place, due to the individual items functioning exactly as expected. Specifically, the items do not appear to be biased toward one racial group or another and are providing information on differences in ability between students taking the assessments.

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.

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 (2008). *STAR Math 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 B-1

Top 10% of Items with Largest Informal DIF Statistics for the White-Black Comparison Group

		Mean P-Value		Absolute P-Value Difference
Count	Item ID	White	Black	Statistic
1	12067	0.51	0.25	0.26
2	12559	0.62	0.38	0.24
3	12024	0.48	0.24	0.24
4	11927	0.41	0.18	0.23
5	12082	0.5	0.28	0.22
6	12549	0.51	0.31	0.21
7	12557	0.59	0.38	0.21
8	10233	0.55	0.35	0.2
9	12036	0.53	0.32	0.2
10	12554	0.63	0.43	0.2
11	12582	0.37	0.17	0.2
12	10220	0.34	0.14	0.2
13	12802	0.76	0.56	0.2
14	12065	0.54	0.34	0.2
15	12075	0.63	0.44	0.2
16	12133	0.51	0.32	0.2
17	12764	0.52	0.32	0.2
18	10245	0.36	0.16	0.2
19	12085	0.51	0.32	0.19
20	12579	0.79	0.6	0.19
21	12578	0.43	0.24	0.19
22	12583	0.48	0.29	0.19
23	12580	0.6	0.41	0.19
24	12973	0.5	0.31	0.19
25	12066	0.55	0.36	0.19
26	12020	0.58	0.39	0.19
27	13094	0.41	0.23	0.18
28	12052	0.55	0.38	0.18
29	12551	0.32	0.14	0.18
30	12602	0.58	0.39	0.18
31	12115	0.37	0.19	0.18
32	13090	0.56	0.38	0.17
33	12970	0.61	0.43	0.17
34	12699	0.46	0.3	0.17

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		White	Black	
35	12129	0.35	0.18	0.17
36	12673	0.59	0.42	0.17
37	12960	0.6	0.42	0.17
38	13032	0.55	0.38	0.17
39	12791	0.53	0.35	0.17
40	12792	0.53	0.36	0.17
41	10242	0.33	0.16	0.17
42	12967	0.57	0.4	0.17
43	12556	0.75	0.58	0.17
44	12890	0.47	0.3	0.17
45	12972	0.45	0.29	0.16
46	13050	0.48	0.32	0.16
47	12782	0.37	0.21	0.16
48	12070	0.45	0.29	0.16
49	12614	0.65	0.48	0.16
50	12084	0.44	0.28	0.16
51	12068	0.56	0.4	0.16
52	12786	0.31	0.15	0.16
53	12541	0.74	0.58	0.16
54	12598	0.51	0.36	0.16
55	12619	0.56	0.39	0.16

Table B-2

Top 10% of Items with Largest Informal DIF Statistics for the White-Hispanic Comparison Group

		Mean P-Value		Absolute P-Value Difference
Count	Item ID	White	Hispanic	Statistic
1	10233	0.55	0.32	0.23
2	13094	0.41	0.19	0.22
3	13092	0.57	0.34	0.22
4	13090	0.56	0.35	0.21
5	12082	0.5	0.31	0.19
6	13088	0.56	0.37	0.18
7	10873	0.44	0.26	0.18
8	12559	0.62	0.45	0.17
9	12970	0.61	0.44	0.17
10	12049	0.68	0.51	0.17
11	12549	0.51	0.36	0.16
12	12085	0.51	0.36	0.16
13	11140	0.52	0.36	0.16
14	12063	0.46	0.3	0.16
15	12130	0.4	0.24	0.16
16	12557	0.59	0.43	0.15
17	12038	0.6	0.44	0.15
18	13083	0.45	0.29	0.15
19	10235	0.43	0.28	0.15
20	12036	0.53	0.39	0.14
21	12554	0.63	0.49	0.14
22	12582	0.37	0.24	0.14
23	10220	0.34	0.2	0.14
24	12579	0.79	0.65	0.14
25	12578	0.43	0.29	0.14
26	12052	0.55	0.41	0.14
27	13095	0.39	0.25	0.14
28	13193	0.53	0.4	0.14
29	11927	0.41	0.28	0.13
30	12802	0.76	0.63	0.13
31	12551	0.32	0.19	0.13
32	12602	0.58	0.45	0.13
33	12699	0.46	0.34	0.13
34	13058	0.64	0.51	0.13

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		White	Hispanic	
35	12548	0.33	0.2	0.13
36	12127	0.53	0.4	0.13
37	11159	0.29	0.15	0.13
38	13190	0.36	0.23	0.13
39	12674	0.55	0.43	0.13
40	11165	0.34	0.21	0.13
41	10236	0.52	0.39	0.13
42	12086	0.7	0.57	0.13
43	13066	0.34	0.21	0.13
44	11157	0.3	0.17	0.13
45	11163	0.47	0.34	0.13
46	12065	0.54	0.42	0.12
47	12075	0.63	0.51	0.12
48	12583	0.48	0.37	0.12
49	12129	0.35	0.23	0.12
50	12673	0.59	0.47	0.12
51	12960	0.6	0.48	0.12
52	12558	0.41	0.29	0.12
53	13091	0.47	0.35	0.12
54	13062	0.27	0.16	0.12
55	10226	0.31	0.2	0.12

Table B-3

Top 10% of 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	12088	0.57	0.42	0.15
2	12127	0.54	0.39	0.15
3	13086	0.27	0.39	0.12
4	12120	0.71	0.60	0.11
5	12924	0.28	0.38	0.1
6	12022	0.32	0.42	0.1
7	12082	0.36	0.46	0.1
8	12922	0.28	0.39	0.1
9	12086	0.7	0.60	0.1
10	12921	0.29	0.38	0.1
11	12043	0.49	0.41	0.09
12	12582	0.27	0.36	0.09
13	12111	0.48	0.40	0.08
14	13052	0.69	0.61	0.08
15	12105	0.47	0.40	0.07
16	12578	0.33	0.40	0.07
17	12113	0.42	0.35	0.07
18	12677	0.74	0.67	0.07
19	12549	0.41	0.48	0.07
20	10870	0.58	0.51	0.07
21	12954	0.48	0.42	0.07
22	13240	0.22	0.29	0.07
23	12073	0.42	0.36	0.07
24	13032	0.49	0.43	0.07
25	12691	0.72	0.65	0.07
26	12671	0.69	0.62	0.06
27	12551	0.24	0.30	0.06
28	12553	0.27	0.34	0.06
29	13092	0.5	0.44	0.06
30	13242	0.22	0.28	0.06
31	13039	0.36	0.30	0.06
32	12617	0.73	0.67	0.06
33	12755	0.69	0.63	0.06
34	13193	0.5	0.44	0.06
35	12688	0.89	0.83	0.06

Count	Item ID	Mean P-Value		Absolute P-Value Difference Statistic
		Male	Female	
36	12609	0.61	0.54	0.06
37	13059	0.73	0.67	0.06
38	13185	0.65	0.59	0.06
39	13008	0.26	0.20	0.06
40	13272	0.26	0.32	0.06
41	12762	0.80	0.74	0.06
42	12765	0.53	0.58	0.06
43	12090	0.56	0.50	0.06
44	12100	0.53	0.47	0.06
45	13189	0.47	0.42	0.06
46	12698	0.24	0.29	0.06
47	12096	0.46	0.40	0.06
48	12651	0.74	0.68	0.06
49	12080	0.53	0.48	0.06
50	12036	0.46	0.40	0.06
51	12037	0.68	0.62	0.05
52	12958	0.56	0.50	0.05
53	12119	0.26	0.31	0.05
54	12635	0.83	0.77	0.05
55	13187	0.54	0.49	0.05

Appendix C

Table C-1

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

	Ability Quartile			
Count	1ST	2ND	3RD	4TH
1	10220	10233	11917	10220
2	11927	11927	12020	11927
3	12024	12020	12024	12020
4	12066	12036	12067	12024
5	12067	12068	12082	12036
6	12075	12082	12115	12065
7	12082	12128	12541	12067
8	12129	12133	12549	12082
9	12541	12541	12551	12129
10	12549	12549	12554	12549
11	12551	12551	12556	12551
12	12554	12554	12557	12554
13	12556	12556	12559	12556
14	12557	12557	12578	12557
15	12559	12559	12579	12559
16	12579	12578	12580	12578
17	12580	12579	12582	12579
18	12582	12580	12583	12580
19	12583	12582	12598	12582
20	12598	12583	12602	12598
21	12602	12602	12614	12602
22	12692	12614	12619	12614
23	12764	12619	12764	12619
24	12782	12673	12782	12782
25	12802	12692	12802	12890
26	12890	12699	12890	12967
27	12970	12764	12970	12970
28	12973	12782	12973	12973
29	13094	12802	13032	
30		12890	13090	
31		12967		
32		12970		

Table C-2

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

	Ability Quartile			
Count	1ST	2ND	3RD	4TH
1	11927	10233	10873	12082
2	12548	11159	11927	12548
3	12549	12063	12038	12549
4	12551	12075	12063	12551
5	12579	12548	12082	12554
6	12802	12549	12130	12557
7	12960	12551	12548	12558
8	13090	12554	12549	12559
9	13094	12557	12551	12578
10		12558	12557	12579
11		12559	12559	12582
12		12578	12578	12602
13		12579	12579	12674
14		12582	12582	12960
15		12583	12583	12970
16		12602	12673	13066
17		12673	12699	
18		12802	12802	
19		12970	12970	
20		13094	13083	
21			13092	
22			13190	

Table C-3

Items with Statistically Significant Results from Gender Mantel-Haenszel DIF Procedure

	Ability Quartile			
Count	1ST	2ND	3RD	4TH
1	12043	12043	12022	12022
2	12073	12073	12036	12082
3	12080	12082	12082	12090
4	12086	12088	12088	12111
5	12088	12096	12105	12113
6	12111	12105	12113	12127
7	12120	12127	12119	12549
8	12127	12549	12549	12551
9	12549	12551	12551	12553
10	12551	12553	12553	12578
11	12553	12578	12578	12582
12	12578	12582	12582	12609
13	12582	12609	12609	12671
14	12609	12617	12635	12921
15	12617	12635	12651	12922
16	12635	12677	12671	12924
17	12651	12688	12677	12954
18	12671	12691	12691	13008
19	12677	12755	12755	13039
20	12688	12762	12762	13086
21	12691	12765	12765	
22	12698	12921	12921	
23	12755	12922	12922	
24	12762	12924	12924	
25	12921	12954	12954	
26	12922	12958	13008	
27	12924	13008	13086	
28	12954	13032	13240	
29	12958	13039		
30	13008	13052		
31	13032	13086		
32	13039	13193		
33	13052	13242		
34	13059			
35	13086			
36	13189			
37	13240			

Appendix D

Table D-1.

Logistic Regression DIF Results for the Top 10% of Items from the White-Black Comparison Group.

	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
Item ID	X^2	p-value	X^2	p-value	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
10220	15.057	0.000	16.701	0.000	0.044	0.005
10233	8.217	0.004	9.235	0.010	0.021	0.002
10242	7.582	0.006	13.459	0.001	0.021	0.016
10245	6.117	0.013	6.846	0.033	0.015	0.002
11927	114.702	0.000	119.820	0.000	0.047	0.002
12020	14.218	0.000	18.823	0.000	0.022	0.006
12024	30.355	0.000	33.614	0.000	0.045	0.005
12036	9.340	0.002	20.014	0.000	0.012	0.014
12052	0.004	0.950	0.029	0.986	0.000	0.000
12065	9.427	0.002	9.478	0.009	0.014	0.000
12066	7.984	0.005	7.987	0.018	0.010	0.000
12067	18.907	0.000	22.077	0.000	0.027	0.004
12068	7.161	0.007	7.240	0.027	0.010	0.000
12070	1.675	0.196	5.493	0.064	0.002	0.005
12075	17.537	0.000	17.695	0.000	0.027	0.000
12082	26.566	0.000	28.469	0.000	0.042	0.003
12084	6.695	0.010	10.813	0.004	0.011	0.006
12085	7.267	0.007	8.152	0.017	0.010	0.001
12115	6.396	0.011	9.195	0.010	0.021	0.009
12128	0.825	0.364	2.054	0.358	0.003	0.004
12129	9.318	0.002	9.377	0.009	0.034	0.000
12133	6.351	0.012	10.987	0.004	0.019	0.014
12541	35.649	0.000	36.484	0.000	0.007	0.000
12549	138.042	0.000	163.794	0.000	0.028	0.005
12551	103.062	0.000	149.786	0.000	0.027	0.006
12554	80.355	0.000	91.223	0.000	0.015	0.002
12556	101.557	0.000	116.493	0.000	0.023	0.003
12557	92.546	0.000	98.308	0.000	0.016	0.001
12559	140.755	0.000	149.052	0.000	0.024	0.002
12578	101.201	0.000	128.955	0.000	0.026	0.007

	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
Item ID	X^2	p-value	X^2	p-value	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
12579	109.561	0.000	139.325	0.000	0.030	0.007
12580	62.974	0.000	64.512	0.000	0.014	0.001
12582	125.169	0.000	167.369	0.000	0.033	0.011
12583	64.756	0.000	68.196	0.000	0.014	0.001
12598	12.435	0.000	16.647	0.000	0.003	0.001
12602	32.933	0.000	49.257	0.000	0.009	0.004
12614	25.177	0.000	39.094	0.000	0.011	0.005
12619	17.181	0.000	25.517	0.000	0.007	0.004
12673	16.404	0.000	24.368	0.000	0.006	0.003
12699	12.164	0.000	12.646	0.002	0.023	0.001
12764	26.936	0.000	27.794	0.000	0.013	0.001
12782	32.347	0.000	33.428	0.000	0.016	0.001
12786	30.136	0.000	30.320	0.000	0.016	0.000
12791	3.934	0.047	9.489	0.009	0.007	0.011
12792	6.691	0.010	6.807	0.033	0.012	0.000
12802	92.797	0.000	99.246	0.000	0.021	0.002
12890	59.489	0.000	73.044	0.000	0.016	0.003
12960	21.985	0.000	25.455	0.000	0.011	0.001
12967	18.579	0.000	19.421	0.000	0.008	0.001
12970	26.631	0.000	31.142	0.000	0.013	0.002
12972	19.305	0.000	38.461	0.000	0.010	0.011
12973	25.921	0.000	35.737	0.000	0.013	0.005
13032	4.333	0.037	7.892	0.019	0.003	0.002
13090	5.909	0.015	6.348	0.042	0.009	0.001
13094	15.480	0.000	18.467	0.000	0.025	0.005

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	X^2	p-value	X^2	p-value	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
10220	4.400	0.036	6.960	0.031	0.014	0.001
10233	10.347	0.001	10.615	0.005	0.029	0.002
10235	5.205	0.023	5.553	0.062	0.016	0.007
10236	2.764	0.096	2.796	0.247	0.007	0.001
10873	8.417	0.004	9.165	0.010	0.037	0.000
11140	3.826	0.050	4.721	0.094	0.006	0.000
11157	5.317	0.021	5.321	0.070	0.021	0.001
11159	4.017	0.045	5.508	0.064	0.015	0.004
11163	2.666	0.103	4.089	0.129	0.008	0.000
11165	2.929	0.087	5.254	0.072	0.011	0.001
11927	23.800	0.000	24.211	0.000	0.011	0.002
12036	4.395	0.036	7.759	0.021	0.007	0.001
12038	5.721	0.017	6.314	0.043	0.010	0.000
12049	8.526	0.004	8.548	0.014	0.013	0.002
12052	2.064	0.151	3.594	0.166	0.003	0.001
12063	16.711	0.000	16.838	0.000	0.027	0.001
12065	2.750	0.097	3.203	0.202	0.004	0.002
12075	6.734	0.009	6.914	0.032	0.012	0.001
12082	21.641	0.000	26.585	0.000	0.038	0.004
12085	5.146	0.023	5.222	0.073	0.007	0.000
12086	2.400	0.121	2.707	0.258	0.003	0.000
12127	1.186	0.276	1.186	0.553	0.004	0.000
12129	4.106	0.043	4.946	0.084	0.016	0.000
12130	4.217	0.040	5.923	0.052	0.015	0.000
12548	47.699	0.000	56.486	0.000	0.011	0.000
12549	57.722	0.000	68.000	0.000	0.012	0.002
12551	55.077	0.000	70.755	0.000	0.012	0.000
12554	118.616	0.000	23.747	0.000	0.003	0.000
12557	40.521	0.000	42.964	0.000	0.007	0.001
12558	25.712	0.000	34.316	0.000	0.005	0.001

	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
Item ID	X^2	p-value	X^2	p-value	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
12559	57.984	0.000	62.339	0.000	0.011	0.000
12578	57.731	0.000	74.228	0.000	0.015	0.000
12579	70.901	0.000	73.117	0.000	0.019	0.001
12582	54.115	0.000	60.605	0.000	0.014	0.000
12583	10.491	0.001	15.636	0.000	0.003	0.000
12602	22.661	0.000	28.939	0.000	0.007	0.001
12635	2.128	0.145	2.525	0.283	0.001	0.000
12673	11.035	0.001	19.744	0.000	0.005	0.001
12674	7.423	0.006	15.479	0.000	0.003	0.000
12699	5.450	0.020	6.168	0.046	0.012	0.001
12802	35.201	0.000	35.299	0.000	0.009	0.001
12960	13.417	0.000	17.267	0.000	0.007	0.000
12970	34.524	0.000	41.104	0.000	0.017	0.002
13058	4.431	0.035	4.727	0.094	0.010	0.000
13066	6.705	0.010	8.954	0.011	0.019	0.003
13083	2.842	0.092	4.085	0.130	0.006	0.001
13088	9.618	0.002	11.127	0.004	0.018	0.000
13090	11.407	0.001	11.500	0.003	0.021	0.000
13091	1.719	0.190	1.987	0.370	0.003	0.000
13092	13.410	0.000	16.641	0.000	0.027	0.001
13094	22.465	0.000	23.472	0.000	0.041	0.001
13095	6.309	0.012	6.789	0.034	0.010	0.001
13187	0.005	0.944	3.229	0.199	0.000	0.000
13190	6.926	0.008	11.027	0.004	0.014	0.001
13193	3.760	0.052	4.299	0.117	0.006	0.002

Table D-3

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

	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
Item ID	X^2	p-value	X^2	p-value	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
10870	4.784	0.029	5.835	0.000	0.006	0.001
12022	32.282	0.000	35.508	0.000	0.017	0.002
12036	4.217	0.040	18.007	0.000	0.002	0.007
12037	4.768	0.029	5.613	0.000	0.002	0.001
12043	11.836	0.001	12.280	0.009	0.006	0.000
12073	10.970	0.001	10.973	0.000	0.006	0.000
12080	4.590	0.032	6.576	0.000	0.002	0.001
12082	31.514	0.000	37.791	0.000	0.017	0.004
12086	12.953	0.000	12.980	0.000	0.006	0.000
12088	46.929	0.000	49.323	0.000	0.021	0.001
12090	4.425	0.035	7.905	0.000	0.002	0.002
12096	6.631	0.010	8.906	0.000	0.004	0.001
12100	6.857	0.009	7.138	0.000	0.004	0.000
12105	12.700	0.000	15.142	0.000	0.006	0.002
12111	13.508	0.000	14.468	0.000	0.007	0.001
12113	11.423	0.001	14.751	0.002	0.006	0.001
12119	4.396	0.036	6.123	0.000	0.006	0.002
12120	9.912	0.002	10.357	0.010	0.010	0.001
12127	17.226	0.000	21.101	0.010	0.016	0.004
12549	108.244	0.000	109.730	0.001	0.007	0.000
12551	100.544	0.000	101.194	0.000	0.007	0.000
12553	91.266	0.000	98.452	0.004	0.007	0.000
12578	92.069	0.000	93.074	0.000	0.007	0.000
12582	135.505	0.000	142.971	0.000	0.011	0.000
12609	28.156	0.000	28.240	0.000	0.005	0.000
12617	23.566	0.000	30.780	0.000	0.003	0.002
12635	21.890	0.000	23.468	0.000	0.003	0.000
12651	22.738	0.000	23.677	0.033	0.004	0.000
12671	30.859	0.000	33.970	0.009	0.004	0.001
12677	24.939	0.000	29.187	0.000	0.004	0.001
12688	37.554	0.000	37.881	0.000	0.008	0.000
12691	26.813	0.000	27.081	0.000	0.005	0.000

	Uniform DIF Only		Uniform and Non-Uniform DIF		Nagelkerke R-Squared	
Item ID	X^2	p-value	X^2	p-value	Attributed to Uniform DIF	Attributed to Non-Uniform DIF
12698	8.009	0.005	8.771	0.000	0.006	0.001
12755	19.781	0.000	21.740	0.000	0.004	0.000
12762	27.061	0.000	28.183	0.033	0.005	0.000
12765	32.940	0.000	34.782	0.000	0.005	0.001
12921	161.363	0.000	165.456	0.004	0.015	0.000
12922	186.371	0.000	190.524	0.000	0.017	0.001
12924	186.073	0.000	189.544	0.000	0.019	0.000
12954	29.580	0.000	32.393	0.018	0.004	0.001
12958	15.146	0.000	20.798	0.000	0.003	0.001
13008	34.692	0.000	36.255	0.027	0.006	0.000
13032	14.911	0.000	22.687	0.017	0.003	0.002
13039	19.399	0.000	20.596	0.042	0.006	0.000
13052	8.866	0.003	12.624	0.000	0.006	0.003
13059	5.860	0.015	6.352	0.000	0.004	0.001
13086	37.805	0.000	37.874	0.009	0.027	0.000
13092	4.030	0.045	4.034	0.000	0.003	0.000
13185	6.168	0.013	6.192	0.000	0.005	0.000
13187	4.687	0.030	6.194	0.000	0.003	0.001
13189	5.516	0.019	6.257	0.358	0.004	0.001
13193	5.583	0.018	5.957	0.019	0.003	0.001
13240	13.689	0.000	13.942	0.000	0.009	0.000
13242	12.005	0.001	13.214	0.064	0.008	0.001
13272	5.995	0.014	7.816	0.986	0.006	0.002

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.
```