THE EFFECT OF ENGLISH PROFICIENCY ON MATHEMATICS PERFORMANCE: A COMPARISON OF ITEM RESPONSE THEORY-BASED AREA AND MANTEL-HAENSZEL METHODS

BY

Pui Chi Chiu

University of Kansas

John Poggio

University of Kansas

Abstract

The purpose of the study is to examine the effect of English proficiency on students' mathematics performance by detecting differential item functioning (DIF) between English Language Learner (ELL) students and Non-ELL students in a state mathematics assessment. A total of 1088 seventh-grade students' data was investigated in the study; 921 were Non-ELL students and 167 were Non-ELL students. White, Asian and Hispanic students were included in the study. Item Responses Theory (IRT) –based area and Mantel-Haenszel methods were used to detect DIF in the study. Comparison between these two methods was conducted. The results showed that 18 items were identified as DIF in the IRT-based area method, and five items were identified as DIF in the Mantel-Haenszel method. Among these identified DIF items, three items were in common with both methods. A follow-up study with large sample sizes will be conducted in order to verify the inconsistent results found in this study.

The Effect of English Proficiency on Mathematics Performance: A Comparison of Item Response Theory-based Area and Mantel-Haenszel Methods

INTRODUCTION

The No Child Left Behind Act of 2001 (NCLB) requires that states; a) integrate scientifically-based reading research into comprehensive reading instructions for young children, b) implement annual standard-based assessments in reading and mathematics for 3^{rd} – 8^{th} grades by 2005-06, c) issue annual report cards on school performance and statewide test results and d) set and monitor Adequate Yearly Progress (AYP) goals in mathematics and reading based on 2001-02 baseline data.

According to the State Board of Education, the purpose of the state assessments are to measure specific indicators within the state curricular standards, provide a building total score that is used to measure AYP, report individual student scores along with the student's performance level, and provide subscale and total scores that can be used in conjunction with local assessment scores to assist in improving a building or district's programs.

The issue of whether language proficiency is related to learning ability and general academic achievement has been debated for many years (Gordon, 1981). However, Tate (1997) pointed out that most of the quantitative studies of mathematics performance have focused on various social groups, but very little research has been conducted on the association between language proficiency itself and trends in performance. This research may be lacking because language proficiency is difficult to disentangle from social and cultural factors (MacGregor & Price, 1999). The question of whether language proficiency affects mathematics learning "is a political question as well as an education question" (Tate & D'Ambrosio, 1997, p.650).

The U.S. Department of Education defines English Language Learner (ELL) as national-origin-minority students who are limited in English proficiency. The ELL term is often preferred over limited-English-proficient (LEP) since it highlights accomplishments rather than deficits. ELLs represent one of the fastest-growing groups among the school-aged population in the U.S. Estimates place the ELL population at over 9.9 million students. The ELL school-aged population has grown by more than 169 percent from 1979 to 2003. The percent of ELLs will be projected to 30 percent of school-aged population by 2015 in the U.S. In fact, many ELLs with academic challenges have been enrolled in U.S. schools since kindergarten, and by the upper elementary years do not have a formal designation to receive support services for language development (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006).

A study of detecting differential item functioning (DIF) in state assessment becomes a crucial issue to be considered and studied. According to Hambleton, Swaminathan and Rogers (1991), the accepted definition of DIF is that an item shows DIF if individuals with same level of ability but are from different groups, do not have the same probability of getting an item right. The operational definition of DIF in this study is that an item shows DIF on the mathematics assessment if individuals having the same mathematics ability but are from different levels of English proficiency, do not have the same probability of getting an item right. Students' mathematics performances are affected by their English proficiency, if the item shows DIF on the assessment. Two methods were used for detecting DIF in the study; Item Response Theory (IRT) –based method and Mantel-Haenszel method.

PURPOSE

In this study, the effect of English proficiency on students' mathematics performance was examined by investigating item responses between Non-ELL students and ELL students in a

Midwest state mathematics assessment. In an attempt to examine the effect of English proficiency on student's mathematics performance, the primary objective was to detect DIF between Non-ELL students and ELL students in mathematics assessment by using IRT-based area method and Mantel-Haenszel method. Students who had low English proficiency were assumed to perform differently from the students who had high English proficiency on the assessment. Specifically, Non-ELL students were assumed to have higher probability of getting item right than ELL students in the state mathematics assessment, even when their mathematics abilities were the same. The degree of agreement between the methods in identifying DIF was determined.

RATIONALE

The purpose of present study was to examine the effect of English proficiency on student's mathematics achievement. Previous studies found that language proficiency adversely affects students' mathematics performance, those scores could not be considered an accurate measurement of true ability. Policymakers must create mechanisms that allow ELL students to be tested alternatively (Brown, 2005).

HYPOTHESIS

The research hypothesis is:

H1: Students' mathematics performances are affected by their English proficiencies, if item shows DIF between Non-ELL students and ELL students in the mathematics assessment.

i. Non-ELL students have higher probability of getting item right than ELL students, even their mathematics abilities are the same.

LITERATURE REVIEW

Research has drawn attention to the importance of language proficiency in students' mathematics performances (see, for example, Abedi & Lord, 2001; Brown, 2005; Kiplinger, Haug, & Abedi, 2000). Students perform 10 percent to 30 percent worse on arithmetic word problems than on comparable problems presented in numeric format (Carpenter, Corbitt, Kepner, Linquist, & Reys, 1980). ELL students score lower than students who are proficient in English on standardized tests of mathematics achievement in elementary school as well as on the Scholastic Aptitude Test (SAT) and the quantitative and analytical sections of the Graduate Record Examination (GRE) (Abedi & Lord, 2001). The large gap between the performance of ELL students and native English speakers on mathematics items with high language demand strongly suggests that factors other than mathematical skill contribute to success in solving word problems (Cummins, Kintsch, Reusser, & Weimer, 1988).

Furthermore, the performance difference between ELL students and Non-ELL students was greater for tests of analytical mathematics that contained linguistically complex items than for computational mathematics (Abedi, Leon, & Mirocha, 2003). ELL students must filter their mathematics knowledge through a second language; they face an extra challenge to learn highly abstract mathematical concepts while they are still learning English (Brown, 2005).

Abedi and Lord (2001) found that students who were ELL scored lower on the mathematics test than proficient speakers of English. Linguistic modifications of test items resulted in significant differences in mathematics performance, and scores on the linguistically modified version were slightly higher. The large performance gap between ELL students and Non-ELL students may not be due mainly to lack of content knowledge. ELL students may

possess the content knowledge but may not be at the level of English language proficiency necessary to understand the linguistic structure of assessment tools (Abedi, 2004).

Moreover, Kiplinger, Haug, & Abedi (2000) found that simplification of linguistic structures and the addition of a glossary for non-mathematics vocabulary to a mathematics assessment results not only in better performance by ELLs and other students who are not good readers, but also virtually benefit to all students.

Two seemingly disconnected events; the increased importance of reading skills in mathematics and an increased emphasis on mathematical assessments, have become the focal point for mathematics education reform in the United States (Matteson, 2006). The link between the two is much stronger than it appears on the surface, and this connection can be confirmed by an examination of the test questions on standardized assessments. One problematic issue of mathematics education is that students must read and comprehend a variety of mathematical representations, which include critical elements that a) support students' mathematical understanding, b) aid the student in communicating mathematical knowledge, c) create connections among mathematical concepts and d) can be used in applying mathematical concepts in the real world.

The interdependent relationship between mathematics and language is acute for most students learning algebra because modeling problem situations requires translating from everyday language to algebraic expression (Driscoll, 1999), including the reorganization and reinterpretation of problem information (MacGregor & Stacey, 1993). The translation from everyday language to mathematical language becomes more perilous for ELL students. ELL students may have a misconception that mathematical symbolic language directly represents everyday language and vice versa, as the result errors occur (Lager, 2006).

Linguistic features which include passive voice constructions, comparative structures, prepositional phrases, sentence and discourse structure, subordinate clauses, conditional clauses, relative clauses, concrete versus abstract or impersonal presentations and negation, may cause difficulty for readers, and may interfere with concurrent task. Abedi (2004) claims that linguistics features slow down and add cognitive load to the readers, and make misinterpretation more likely. The performance of ELL students may be underestimated, as language factors introduce another source of measurement error in ELL students' test results that may not have much impact on native/fluent speakers of English (Abedi, 2002).

Cleary, ELL students' poor performance at mathematics problem-solving tasks can be a result of their level of English proficiency, which can mask their mathematics knowledge. Students' mathematics achievements are no longer accurately measured, and the entire mathematics assessment becomes unfair and invalid, if the mathematics assessment does not only measure students' mathematics abilities, but also their language proficiencies.

In order to study the effect of English proficiency on student's mathematics performance, detection of DIF between Non-ELL students and ELL students was conducted. A widely accepted definition of DIF was that an item is identified as DIF if examinees of equal ability, but from different subgroups do not have an equal probability of correctly responding to that item (Hambleton & Rogers, 1989). If the discrepancy in item performance between the subgroups of interest is equal across the entire range of abilities then the DIF is said to be "uniform". However, if the difference between the subgroups is not consistent across the entire range of abilities then the DIF is said to be "non-uniform" (Hambleton, Clauser, Mazor & Jones, 1993).

Again, the operational definition of DIF in this study was that an item shows DIF on mathematics assessment if individuals having the same mathematics ability, but from different

English proficiency, do not have the same probability of getting the item right. Students' mathematics performances are affected by their English proficiency, if the item shows DIF on the assessment. Two methods were used to detect DIF in this study; the IRT–based area method and the Mantel-Haenszel method.

IRT-based Method

In IRT, the function is identical at all points and the probabilities of a correct response are the same, if the parameters of two item characteristic functions are identical. DIF was investigated by comparing the item characteristic functions of two or more groups. Item characteristics functions may be compared in several ways. The first way was to compare parameters that describe item characteristics curves (ICCs). The second approach was to compare ICCs by evaluating the area between them (Rudner et al., 1980).

The most popular of the IRT-based DIF methods is the "area method" (Hambleton, Clauser, Mazor & Jones, 1993). The ICCs for the majority and minority groups are compared and the area between the ICCs over some internal on the ability scale is used as an indicator of DIF. The larger the area is, the more DIF is said to be present. However, one shortcoming of the area method was the lack of a critical value for interpreting the area statistics.

Mantel-Haenszel Method

In additional to the IRT-based method, Mantel-Haenszel method was also used to detect DIF in the study. The Mantel-Haenszel method was introduced into the measurement literature by ETS (Holland &Thayer, 1986, 1988). It compares the probabilities of a correct response in the two groups of interest for examinees of the same ability, although its calculation is very different form the IRT-based methods (Holland & Thayer, 1988).

The Mantel-Haenszel method works with item responses for two groups: the reference group and the focal group. Examinees are first sorted into score groups (k) according to total test score (m), such that $k = 1, 2, \ldots, m$; resulting in up to (n + 1) score groups, where n is the number of items in the test. After reference- and focal-group examinees are matched on total test score, a 2 x 2 x S contingency table is formed, where S is the number of different values of the total test score. At each score level, the data can be arranged as a 2 x 2 table, see Table 1.

In Table 1, A_j, B_j, C_j, and D_j, correspond to the numbers of examinees in the four cell; n_{Rj}, n_{Fj}, m_{1j}, and m_{0j}, are the marginals. T_j is the number of examinees in the jth score group who attempted the item under investigation. The Mantel-Haenszel chi-square statistic (Mantel, 1963) has the form:

$$\chi^2_{MH} = \frac{\left(\left| \sum_j A_j - \sum_j E\left(A_j\right) \right| - \frac{1}{2} \right)^2}{\sum_j Var\left(A_j\right)} \tag{1}$$

where

$$E\left(A_j\right) = \frac{n_{Rj} m_{1j}}{T_j} \tag{2}$$

and

$$Var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 \left(T_j - 1\right)} \tag{3}$$

This statistic is distributed as chi-square with one degree of freedom. The Mantel-Haenszel statistics exceeding the tabulated value of the chi-square distribution at a specific level of alpha indicate that item performance in the reference and focal groups over the (n + 1) score groups is consistently different (Hambleton & Rogers, 1989).

In addition, the Mantel-Haenszel DIF statistic, common log-odds ratio, was also determined. The Mantel-Haenszel statistic defines DIF in terms of the ratio of the odds of a correct response on the studied item for the reference group to the odds of a correct response for the focal group. In accordance with the definition of DIF, the groups are first matched in an estimate of ability, and then the odds ratio is calculated at each level of estimated ability. The logarithm of a weighted average of these separate odds ratios is computed to form the common log-odds ratio (Mantel & Haenszel, 1959). Positive values indicate DIF in favor of the reference group, and negative values indicate DIF in favor of the focal groups. The common log-odds ratio is given by:

$$\log\left(\hat{\alpha}_{MHj}\right) \tag{4}$$

where

$$\hat{\alpha}_{MHj} = \frac{\sum_{k=1}^{m} A_{jk} D_{jk} / T_k}{\sum_{k=1}^{m} B_{jk} C_{jk} / T_k} \tag{5}$$

METHOD

This study addressed the effect of English proficiency on student's mathematics performance. A sample of 1088 seventh-grade students was divided into two groups, Non-ELL and ELL, based on their English proficiencies. There were 921 students in the Non-ELL group and 167 students in the ELL group, those groups included White, Asian and Hispanic students. Univariate Analysis of Variance (ANOVA) was conducted to examine student's prior mathematics ability.

The IRT-based area method and the Mantel-Haenszel method were used to detect DIF between Non-ELL students and ELL students in mathematics assessment. The degree of agreement between the methods in identifying DIF was determined.

Participants

In this study, a previous collected dataset was used. The sample consisted of 1088 random samples of seventh-grade students who took the mathematic assessment in a Midwest state in Spring 2007. A total of 921 students were in the Non-ELL group, and 167 students were in the ELL group. In the Non-ELL group, there were 500 White students, 102 Asian students and 319 Hispanic students. In the ELL group, there were seven White students, 38 Asian students and 122 Hispanic students. Numbers of student in each group were summarized in Table 2.

Measures

Two test forms of mathematics assessment were given to the students: paper and pencil (P&P), and computer-based testing (CBT). In the assessment, there are 84 multiple-choice items; each item contains four response choices (A, B, C and D), and one point is given to each correct response. At the raw score scale, Cronbach's alphas are 0.923 for the overall sample, 0.925 for the Non-ELL group and 0.909 for the ELL group. Reliability of mathematics assessment in 2006 was measured by classification consistency and classification accuracy, the coefficients were 0.64 and 0.74 respectively. Validity of mathematics assessment in 2006 was examined by criterion-related validity evidence, which determines the degree to which examinees' performance on a test correlates at expected levels with one or more outcome criteria Results showed that the validity of 2006 state assessment scores was supported (Poggio, Yang, Irwin, Glasnapp, & Poggio, 2007).

Analytical Procedures

In order to detect DIF between Non-ELL and ELL groups, 1088 students were divided into two groups, Non-ELL group (reference group) and ELL group (focal group), based on their English proficiencies. There were 921 students in the reference group and 167 students in the focal group.

IRT-based Area Method

Parameter model fit analysis was conducted. In the parameter model fit analysis, the purpose was to determine which Parameter Logistic (PL) model, 3-PL, 2-PL or 1-PL model, fits the data. After conducting the model fit analysis, item parameters and ability parameters for mathematics assessments were estimated separately for each group, with common c-parameters, by using Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996).

For this IRT method, item parameters and latent trait scores (θ) were estimated by using marginal maximum likelihood estimation (MMLE) in Bilog-MG. Lord (1980) also suggested fixing c-parameters for each item to the combined-sample value in both groups, as the c-parameters are often poorly estimated and one does not want to identify DIF based solely on the c-parameter.

The fixed values of c-parameters were estimated in Bilog-MG. In this procedure, two groups of students were combined and item parameters were estimated at the same time. After c-parameters had been estimated, these c-parameters were used as fixed values in two other separated calibrations; one for the reference group and another for the focal group. In each calibration, a- and b-parameters were estimated.

Prior to the DIF analyses, two groups were placed onto the same scale by using Mean and Sigma Equating method. In this method, the scale of item parameters from focal group was

placed onto the scale of item parameters from reference group by using linear transformation. The equating constants, x and y, were determined:

Reference group: $\mu_{b\_reference} = -0.378$; $\sigma_{b\_reference} = 1.187$

Focal group: $\mu_{b\_focal} = -0.570$; $\sigma_{b\_focal} = 0.947$

$$x = \sigma_{b\_reference} / \sigma_{b\_focal}$$

$$= 1.187 / 0.947$$

$$= 1.253$$

$$y = \mu_{b\_reference} - (x * \mu_{b\_focal})$$

$$= -0.378 - (1.253 * -0.570)$$

$$= 0.336$$

$$b_{new\ focal} = x * b_{old\ focal} + y$$

$$= 1.253 * b_{old\ focal} + 0.336$$

$$a_{new\ focal} = a_{old\ focal} / x$$

$$= a_{old\ focal} / 1.253$$

$$c_{new\ focal} = c_{old\ focal} = c_{reference} = \text{common c}$$

After placing item parameters of focal group onto the scale of item parameters of reference group, the probability of correct response was calculated for each item in each group by using the equation of 3PL-model:

$$P\left(1\middle|\theta i\right)=cj+\left(1-cj\right)\frac{e^{1.7aj\left(\theta i-bj\right)}}{1+e^{1.7aj\left(\theta i-bj\right)}} \tag{6}$$

ICCs were plotted for each item. Each ICC represented the probability of obtaining a correct response across different mathematics ability levels, which ranged from -4 to 4. In order to easily compare ICCs between groups for each item, two groups' ICCs were placed together, and each group was indicated by different colored line, see Figure 1.

In this study, evaluating the area between ICCs was the main method for detecting DIF. DIF is not present, if after placing the parameter estimates on a common scale, the ICCs are identical and the area between the ICCs is zero. In computing the area, the value of c-parameters was assumed to be the same for both groups. Raju (1988) has shown that when the c-parameters are not the same, the area between the two curves is infinite if calculated over the entire range of ability. In this study, the closed form solution with common c-parameters was used to calculate the area between ICCs (Raju, 1988):

$$A_j=(1-c_j)\left|\frac{2(a_2-a_1)}{Da_1a_2}\ln\left[1+e^{Da_1a_2(b_2-b_1)/(a_2-a_1)}\right]-(b_2-b_1)\right| \tag{7}$$

Because there is no known sampling distribution for the area statistic under the null hypothesis of no group differences, items are typically ranked according to the values of the statistic, and those items with the highest values are flagged as potentially biased (Hambleton & Rogers, 1989). One possible solution to this problem is to obtain an approximate cutoff-score for identifying DIF presence. In this study, the cutoff value was obtained by carrying out an analysis

on two randomly equivalent groups; two groups of 1000 White Non-ELL samples. These two samples were assumed to have no bias present. The largest area statistic obtained served as an indicator of the greatest value of the statistic likely to occur by chance.

Mantel-Haenszel Method

In calculating the Mantel-Haenszel statistics, 69 score groups were formed based on the total test score. Mantel-Haenszel chi-square, common log-odds ratio and estimated standard error, Breslow-Day chi-square test (Breslow & Day, 1980) of trend in odds ratio heterogeneity, and Educational Testing Service (ETS) categorization scheme (Zieky, 1993), were calculated by using Differential Item Functioning Analysis System (DIFAS) (Penfield, 2005).

Breslow-Day chi-square test has been shown to be effective at detecting Non-Uniform DIF, whose calculations are similar to the Mantel-Haenszel chi-square test. Given a n-item test taken by two groups, the Breslow-Day test provides a statistics about the null hypothesis that states the homogeneity of the odds ratio across the k levels of the total test score, k = n – 1. The statistics is:

$$BD = \sum_{j=1}^{k} \frac{(A_j - E(A_j))^2}{Var(A_j)} \tag{8}$$

where

$$E(A_j) = \left(\hat{\alpha}(n_{Rj} + m_{1j}) + (n_{Fj} - m_{1j}) \pm \left\{\left[\hat{\alpha}(n_{Rj} + m_{1j}) + (n_{Fj} - m_{1j})\right]^2 - \left[4(\hat{\alpha} - 1)\hat{\alpha}(n_{Rj} m_{1j})\right]\right\}^{1/2}\right) / 2(\hat{\alpha} - 1) \tag{9}$$

and

$$Var(A_j) = \left(\frac{1}{E(A_j)} + \frac{1}{n_{Rj} - E(A_j)} + \frac{1}{m_{1j} - E(A_j)} + \frac{1}{n_{Fj} - m_{1j} + E(A_j)} \right)^{-1} \qquad (10)$$

$\hat{\alpha}$ is an estimate of the common odds ratio; A_j is the number of examinees with score j belonging to the reference group that responded correctly to the item under study; E and Var, respectively, are the expected value and the variance for the cell under the assumption of the homogeneity of the odds ratio, see Table 1. Under the null hypothesis, the statistic has an asymptotic chi-square distribution with k' – 1 degree of freedom, k' being the number of tables effectively considered.

ETS has developed a classification scheme – ETS categorization scheme – to flag items with DIF. Holland and Thayer (1985) transformed $\hat{\alpha}_{MH}$ to the ETS "delta scale" by defining $\hat{\Delta}$ as

$$\hat{\Delta} = -2.35 \ln[\hat{\alpha}_{MH}] \qquad (11)$$

The delta scale is an inverse normal distribution transformation of the percent correct to a linear scale with a mean of 13 and a standard deviation of 4 and is used as an index of item difficulty by ETS test development staff. A $\hat{\Delta}$ value for 0 indicates no difference in difficulty for the two groups. A positive value of $\hat{\Delta}$ indicates that the item was more difficult for the reference group, and a negative value indicates the item was more difficult for the focal group (Holland & Thayer, 1988).

The ETS categorization scheme categorizes items as having small (A), moderate (B), and large (C) levels of DIF. If $|\hat{\Delta}|$ for a particular item is at least 1.5 and is significantly greater than

1.0 at the 0.05 significance level, the item is classified as a "C" DIF item. If $|\hat{\Delta}|$ for a particular item is less than 1.0 or if $|\hat{\Delta}|$ is not significantly greater than 0.0 at the 0.05 significance level, the item is classified as an "A" DIF item. All other items are classified as "B" items.

In addition to the omnibus comparison between Non-ELL students and ELL students, two follow-up tests were also conducted. Hispanic Non-ELL students were compared with Hispanic ELL students, and White Non-ELL students were compared with Hispanic ELL students. DIFAS was also used to detect DIF for each of the follow-up tests.

Data Analysis

Descriptive statistics were used to examine within-group equality; mean and standard deviation of each subtest were calculated for each group. Univariate Analysis of Variance (ANOVA) was conducted to examine students' prior mathematics ability based on sixth-grade mathematics score.

IRT-based Area Method

Item parameters (a-, b- and c-parameters) were estimated by using Bilog-MG. Estimation of item parameters was based on the method of marginal maximum likelihood (MML). The MML solution employed two methods of solving the marginal likelihood equations: the EM method and the Newton-Gauss iterations. Approximation chi-square indices of fit were computed for each item following the estimation cycle. Output from the Bilog-MG program for the two groups of interest was inputted directly into Microsoft Excel®. Probability of getting an item right was calculated over an ability interval, -4 to 4. ICCs were plotted by using Statistical Package for the Social Sciences (SPSS), and the area between ICCs was calculated by using equation 7.

Mantel-Haenszel Method

In calculating the Mantel-Haenszel statistics, DIFAS was used. DIFAS is a Windows based program that performs a variety of functions related to assessing the presence of DIF in items (Penfield, 2005). The DIF procedures that DIFAS runs for dichotomously scored items include Mantel-Haenszel chi-square, common log-odds ratio and estimated standard error, Breslow-Day test of trend in odds ratio heterogeneity, and ETS categorization scheme.

The Mantel-Haenszel chi-square statistic is distributed as chi-square with one degree of freedom. Critical values of this statistic are 3.84 at the 0.05 significance level and 6.63 at the 0.01 significance level. The Mantel-Haenszel common log-odds ratio is asymptotically normally distributed. Positive values indicate DIF in favor of the reference group, and negative values indicate DIF in favor of the focal group. The Breslow-Day chi-square test of trend in odds ratio heterogeneity is a statistic for detecting Non-Uniform DIF, which is distributed as chi-square with one degree of freedom. Critical values for these statistics are 3.84 at the 0.05 significance level and 6.63 at the 0.01 significance level.

RESULTS

Descriptive Statistics

Table 3 shows the performance of seventh-grade students in a state mathematics assessment. The table presents descriptive statistical results (means and standard deviations) on the total responses correct for both Non-ELL students and ELL students (White, Asian and Hispanic). A review of these data demonstrates that descriptively, in both the Non-ELL group and the ELL group, Asian students have highest mean scores, whereas Hispanic students have lowest mean scores.

Statistical Inferential Analysis

Univariate Analysis of Variance (ANOVA)

In order to examine students' prior mathematics abilities, a univariate ANOVA test was conducted. In this study, there were three racial subgroups, White, Asian and Hispanic, among both Non-ELL and ELL groups; a total of six subgroups were formed. Tamhane's T2 Post Hoc test was used to analyze the difference between the ELL and Non-ELL groups for each racial subgroup. The results indicated that there was no significant difference between subgroups on sixth-grade mathematics scores at the 0.01 significance level; *p*-values were greater than 0.01 in all subgroup-comparisons, see Table 4.

DIF Analyses

IRT-based Area Method

The results indicated that 3-PL model fitted the data well; the E-M cycle converged and chi-square statistics test showed a statistically significant model fit, $p > 0.05$. This means that the observed data was not statistically significant different from the expected value in 3-PL model in IRT. Estimated item parameters are presented in Table 5.

Eighty-four ICCs were plotted, item parameters were estimated, and the area between ICCs was calculated, see Table 6. Since there is no known sampling distribution for the area statistic under the null hypothesis of no group differences, a cutoff value was obtained by carrying out an analysis on two randomly equivalent groups; two 1000 White Non-ELL samples. The assumption was that there was no bias present in these two equivalent groups, the largest area statistics obtained serves as an indicator of the greatest value of the statistic that likely occurred by chance. Although this was not an ideal approach, it provided an approximate cutoff-score for an area statistic. In this study, the cutoff value for the area statistic was 0.3414.

Test statistics showed that 18 items were identified as DIF and those calculated areas were greater than the cutoff value, 0.3414. Within these 18 DIF items, there were seven Uniform

DIF items (the discrepancy in term performance between the subgroups of interest was equal across the entire range of abilities), and 11 Non-Uniform DIF items (the difference between the subgroups was not consistent across the entire range of abilities). Numbers of No-DIF and DIF items are summarized in Table 7.

Mantel-Haenszel Method

As shown in Table 8, the results from DIFAS indicated that five items, items 27, 40, 41, 65, and 78, were identified as DIF. Among these five items, three items were identified as Uniform DIF and two items were identified as Non-Uniform DIF. In the Mantel-Haenszel chi-square test, chi-square values for items 27, 41 and 65 were greater than the critical value of 3.84 at the 0.05 significance level. In the Mantel-Haenszel common log-odds ratio, item 27 and item 41 showed positive values, whereas item 65 showed negative value. In the Breslow-Day chi-square test, two items were identified as Non-Uniform DIF; item 40 and item 78, those chi-square values were greater than 3.84 at the 0.05 significance level. ETS categorization scheme indicated that three items were in moderate level of DIF and the rest of the items were in small level of DIF. The numbers of DIF items are summarized in Table 9.

Two follow-up tests were conducted; comparison between Hispanic Non-ELL students and Hispanic ELL students; and comparison between White Non-ELL students and Hispanic ELL students. Table 10 presents the results obtained from DIFAS. In the Hispanic groups' comparison, seven items were identified as DIF; either the Mantel-Haenszel chi-square value or the Breslow-Day chi-square value was greater than the critical value of 3.84 at the 0.05 significance level. Among these seven items, two items were identified as Uniform DIF and five items were identified as Non-Uniform DIF, see Table 10a. In the White and Hispanic groups' comparison, four items were identified as DIF; the Breslow-Day chi-square values were greater

than 3.84 at the 0.05 significance level and they were identified as Non-Uniform DIF, see Table 10b. Numbers of DIF items are summarized in Table 9.

Comparison between the IRT-based Area Method and the Mantel-Haenszel Method

IRT-based area method and Mantel-Haenszel method were compared. Results of the comparison are summarized in Table 11. In the IRT-based area method, 18 items were identified as DIF and 66 items were identified as No-DIF. Among these 18 DIF items, seven items were identified as Uniform DIF and 11 items were identified as Non-Uniform DIF. In the Mantel-Haenszel method, five items were identified as DIF and 79 items were identified as No-DIF. Among these five DIF items, three items were identified as Uniform DIF and two items were identified as Non-Uniform DIF. Three items were in common in the IRT-based area method and the Mantel-Haenszel method; items 27, 65 and 78.

DISCUSSIONS

Students' prior mathematics abilities were examined by comparing students' sixth-grade mathematics scores between groups. Results showed that there was no statistically significant difference between Non-ELL students and ELL students on their sixth-grade mathematics scores, and there was no statistically significant difference between White, Asian and Hispanics students among those two groups. Therefore, mathematics ability was not a factor that caused differences between the Non-ELL group and the ELL group on their mathematics performances in the state mathematics assessment. This finding was critical to interpreting the following results.

Three items were identified as DIF in both the IRT-based area method and the Mantel-Haenszel method; two were Uniform DIF items and one was Non-Uniform DIF item. Among these two Uniform DIF items, item 27 and item 65, statistics showed that Non-ELL students outperformed ELL students on item 27, but not on item 65. Item 27 is a number sense question

that asks students to calculate the number of times a program spends on an entire show by giving students percentage and fraction information. The results showed that Non-ELL students had higher probability of getting the item right than ELL students, even though their mathematics abilities were the same. However, item 65 which is a computation question that asks students to indicate the location of digits in a product, showed opposite results. On item 65, Non-ELL students had lower probability of getting the item right than ELL students, even though their mathematics abilities were the same.

On the other hand, item 78 is a word problem that asks students the meaning of the numbers shown in a box-and-whiskers plot. The results of the Non-Uniform DIF item showed that Non-ELL students had a higher probability of getting the item right than ELL students when those students had medium to high mathematics abilities, $\theta > -0.80$. ELL students had a higher probability of getting the item right than Non-ELL students when those students' mathematics abilities were less than -0.80. These findings indicate that items identified as DIF in the study were not only biased against ELL students, but also Non-ELL students.

In addition, the findings indicated inconsistent results between the omnibus Mantel-Haenszel test and the follow-up tests in detecting DIF. The results showed that seven items were identified as DIF in the Hispanic groups' comparison, and four items were identified as DIF in the White and Hispanic groups' comparison. Among these identified DIF items, there was only one item, item 78, identified in both omnibus test and follow-up tests. The other items were only identified in the follow-up tests. More items were identified as DIF in the follow-up tests than in the omnibus test. Sample size and sample characteristic seem to be the issues related to this difference. Therefore, a simulation study with large sample sizes needs to be conducted in order to provide a solid answer to this question.

Furthermore, results showed that the Mantel-Haenszel method is more conservative than the IRT-based area method in detecting DIF; more items were identified as DIF in the IRT-based area method than in the Mantel-Haenszel method. Fifteen items that were identified as DIF in the IRT-based area method were not identified in the Mantel-Haenszel method, whereas two items that were identified as DIF in the Mantel-Haenszel method were not identified in the IRT-based area method. One possible reason for this discrepancy is that there was lack of critical value for interpreting area statistics in the IRT-based area method. The cutoff value that was used in this study was only an approximation cutoff score for the area statistics. Over-identification of DIF items may occur based on the approximated cutoff value of 0.3414. Thus, IRT-based area method would not be a reliable method in detecting DIF in the assessment, if there is no ideal cutoff score for the area statistics. A simulation study with large sample sizes needs to be conducted to obtain a sampling distribution for the area statistics in the future.

CONCLUSION

Several major points emerge from these findings. First, the effect of language proficiency on students' mathematics performance was examined in this study. Eighteen items were identified as DIF in the IRT-based area method and five items were identified as DIF in the Mantel-Haenszel method. Among these identified DIF items, three items were in common. Although item statistics indicated that not all identified DIF items are biased against ELL students, the effect of language proficiency on student's mathematics performance still needs to be considered. Policymakers should create mechanisms that allow ELL students to be tested alternatively, for instance, they may provide students a glossary of non-mathematical terms on the assessment, or make modifications of tests for assessing ELL students' mathematics abilities. Thus, the challenge of non-mathematical English vocabulary for ELL students can be reduced, the

language barrier can be minimized, and students' mathematics abilities can be meaningfully, equitably and accurately assessed.

Second, the results of the study indicate that there is a discrepancy between the IRT-based area method and the Mantel-Haenszel method in detecting DIF. Choice of one method or the other, therefore, seems to be a matter of reasonable indifference. One difficulty with the IRT-based area method is the lacking of a known sampling distribution for area statistics. The cutoff value that was used in this study was just an approximated cutoff score. Thus, decisions made based on this cutoff score may not be accurate enough in identifying the correct number of DIF items in the assessment.

Third, inconsistent results were found between the IRT-based area method and the Mantel-Haenszel method, and also found between the omnibus test and the follow-up tests in the Mantel-Haenszel method. A repeated study or a simulation study with large sample sizes needs to be conducted in the future, so that the results found in this study can be compared and evaluated.

Finally, no single method can be guaranteed to identify all of the DIF items in a test. Multiple methods may be used to address the instability problem which undermines the utility of current methods and can address the shortcomings found in particular methods.

References

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometric issues. *Educational Assessment*, *8*(3), 231–257.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researchers, 33*(1), 4-14.

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Lord, C. (2001). The Language Factor in Mathematics Tests. *Applied Measurement in Education, 14*(3), 219-234.

Breslow, N. E., & Day, N. E. (1980). *Statistical methods in cancer research: Volume 1 – The analysis of case-control studies*. Lyon: International Agency for Research on Cancer.

Brown, C. L. (2005). Equality of Literacy-Based Math Performance Assessments for English Language Learners. *Bilingual Research Journal, 29*(2), 337-363.

Bloom, B. S. (1987). *Taxonomy of Educational Objectives.* Longman Inc, NY: White Plains.

Carpenter, T. P., Corbitt, M. K., Kepner, H. S., Jr., Linquist, M. M., & Reys, R. E. (1980). Solving verbal problems: Results and implications from national assessment. *Arithmetic Teacher, 28*, 8-12.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20*, 405-438.

Driscoll, M. (1999). *Fostering algebraic thinking: A guide for teachers grades 6-10*. Portsmouth, NH: Heinemann.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical Guidelines for the Education of English Language Learners: Research-Based Recommendations for the Use of Accommodations in Large-Scale Assessments.* Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Gordon, J. C. B. (1981). *Verbal deficit: A critique.* London: Croom Helm.

Hambleton, B. K., & Rogers, H. J. (1989). Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel methods. *Applied Measurement in Education, 2*(4), 313-334.

Hambleton, B. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the Detection of Differentially Functioning Test Items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Hambleton, B. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publication.

Holland, P. W., & Thayer, D. T. (1985). An alternative definition of the ETS delta scale of item difficulty. ETS Research Report No. 85-43. Princeton, NJ: Educational Testing Service.

Holland, P. W., & Thayer, D. T. (1986). *Differential item performance and the Mantel-Haenszel procedure* (Tech. Rep. No 86-31). Princeton, NJ: Educational Testing Service.

Holland. P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Kiplinger, V. L., Haug, C. A., & Abedi, J. (2000, April). *Measuring Math – Not Reading – on a Math Assessment: A Language Accommodations Study of English Language Learners and Other Special Populations*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Lager, C. A. (2006). Types of Mathematics-Language Reading Interactions that Unnecessarily Hinder Algebra Learning and Assessment. *Reading Psychology, 27*(2), 165-204.

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

Lubienski, S. T., Camburn, E., & Shelley, M. C. (2004). *Reform-Oriented Mathematics Instruction, Achievement, and Equity: Examinations Of Race and SES in 2000 Main NAEP Data*. Report to the National Center for Education Statistics.

MacGregor, M. & Price, E. (1999). An Exploration of Aspects of Language Proficiency and Algebra Learning. *Journal for Research in Mathematics Education, 30* (4), 449-467.

MacGregor, M., & Stacey, K. (1993). Cognitive models underlying students' formulation of simple linear equations. *Journal for Research in Mathematics Education. 24*, 217–232.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extension of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, *58*, 690-700.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

Matteson, S. M. (2006). Mathematical Literacy and Standardized Mathematical Assessments. *Reading Psychology, 27*(2), 205-233.

No Child Left Behind Act of 2001. (2002), Pub. Law.107-110, 115 Stat. 1425.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29,* 150–151.

Poggio, A. J., Yang, X., Irwin, P. M., Glasnapp, D. R., & Poggio, J. P. (2007). *Kansas Assessments in Reading and Mathematics 2006 Technical Manual for the Kansas General Assessments, Kansas Assessments of Multiple Measures (KAMM), Kansas Alternate*

Assessments (KAA). Lawrence, KS: Center for Educational Testing and Evaluation, The University of Kansas.

Raju, N. S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two items response functions. *Applied Psychological Measurement, 14*(2), 197-207.

Rudner, L. M., Getson, P. R. & Knight, D. L. (1980). Biased item detection techniques. *Journal of Educational Statistics, 5*, 213-233.

Tate, W. F. (1997). Race, ethnicity SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education, 28*, 652-679.

Tate, W. F., & D'Ambrosio, B. S. (1997). Equity, mathematics reform, and research. *Journal for Research in Mathematics Education, 28*, 650-651.

Yang, Y. (2003). Dimensions of Socio-economic Status and their Relationship to Mathematics and Science Achievement at Individual and Collective Levels. *Scandinavian Journal of Educational Research, 47*(1), 21-41.

Zieky, M. (1993). Practical questions use of DIF statistics it item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-364). Hillsdale, NJ: Lawrence Erlbaum.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996).*Bilog MG: Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Table 1

2 x 2 Contingency Table for Score Levels.

	Item Score		
	Correct	Incorrect	Total
Reference Group (R)	$\mathbf{A_j}$	$\mathbf{B_j}$	n_{Rj}
Focal Group (F)	$\mathbf{C_j}$	$\mathbf{D_j}$	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

Table 2

Numbers of student in each group.

	Non-ELL Group	**ELL Group**	**Total**
White	**500**	**7**	507
Asian	**102**	**38**	140
Hispanic	**319**	**122**	441
Total	921	167	**1088**

Table 3

Descriptive Statistics for Non-ELL Students and ELL Students on Total Responses Correct.

Group	Race	N	Minimum	Maximum	Mean	Std. Deviation
Non-ELL	**White**	500	16	84	60.17	13.596
	Asian	102	26	83	64.78	12.760
	Hispanic	319	24	83	56.42	12.581
ELL	**White**	7	33	73	58.00	14.742
	Asian	38	31	84	58.21	14.697
	Hispanic	122	28	79	56.54	11.769

Table 4

Tamhane's T2 Post Hoc Test on Students' 6th Grade Mathematics Score.

(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	*p*-value
Asian Non-ELL	Hispanic Non-ELL	4.348	1.339	0.022
	White Non-ELL	2.885	1.345	0.401
	Asian ELL	5.367	2.473	0.403
	Hispanic ELL	4.518	1.402	0.023
	White ELL	4.014	5.635	1.000
Hispanic Non-ELL	Asian Non-ELL	-4.348	1.339	0.022
	White Non-ELL	-1.463	0.777	0.606
	Asian ELL	1.019	2.216	1.000
	Hispanic ELL	0.169	0.872	1.000
	White ELL	-0.334	5.527	1.000
White Non-ELL	Asian Non-ELL	-2.885	1.345	0.401
	Hispanic Non-ELL	1.463	0.777	0.606
	Asian ELL	2.483	2.220	0.991
	Hispanic ELL	1.633	0.881	0.635
	White ELL	1.129	5.529	1.000
Asian ELL	Asian Non-ELL	-5.367	2.473	0.403
	Hispanic Non-ELL	-1.019	2.216	1.000
	White Non-ELL	-2.483	2.220	0.991
	Hispanic ELL	-0.850	2.255	1.000
	White ELL	-1.353	5.906	1.000
Hispanic ELL	Asian Non-ELL	-4.518	1.402	0.023
	Hispanic Non-ELL	-0.169	0.872	1.000
	White Non-ELL	-1.633	0.881	0.635
	Asian ELL	0.850	2.255	1.000
	White ELL	-0.504	5.543	1.000
White ELL	Asian Non-ELL	-4.014	5.635	1.000
	Hispanic Non-ELL	0.334	5.527	1.000
	White Non-ELL	-1.129	5.529	1.000
	Asian ELL	1.353	5.906	1.000
	Hispanic ELL	0.504	5.543	1.000

Table 5

Estimated Item Parameters in IRT 3PL-Model.

Item Number	Non-ELL Group			ELL Group		
	a-parameter (slope)	b-parameter (threshold)	Common c-parameter (asymptote)	a-parameter (slope)	b-parameter (threshold)	Common c-parameter (asymptote)
1	1.083	-1.773	0.240	1.409	-1.755	0.240
2	0.764	-1.216	0.182	0.752	-1.506	0.182
3	0.669	-1.506	0.226	0.808	-1.430	0.226
4	0.889	-1.666	0.259	1.384	-1.445	0.259
5	1.056	-0.728	0.234	1.280	-0.629	0.234
6	0.596	-2.138	0.232	0.537	-2.578	0.232
7	1.408	-0.625	0.212	1.007	-0.702	0.212
8	0.726	1.502	0.208	0.904	1.035	0.208
9	0.875	-1.213	0.211	1.195	-1.172	0.211
10	1.007	-3.205	0.236	0.968	-3.017	0.236
11	0.709	-0.839	0.268	0.822	-0.519	0.268
12	0.501	-1.105	0.321	0.581	-1.415	0.321
13	0.635	-0.084	0.305	0.897	0.033	0.305
14	0.341	-0.777	0.386	0.423	-0.929	0.386
15	1.104	0.108	0.124	1.230	0.187	0.124
16	0.933	0.176	0.310	1.230	0.392	0.310
17	1.369	0.591	0.141	1.229	0.701	0.141
18	1.121	0.951	0.157	1.109	0.629	0.157
19	0.925	-2.176	0.222	0.883	-2.465	0.222
20	1.260	-0.284	0.137	1.372	-0.314	0.137
21	1.000	-0.969	0.304	1.146	-1.110	0.304
22	1.088	-0.471	0.147	1.040	-0.727	0.147
23	0.556	-1.862	0.236	0.857	-1.400	0.236
24	0.850	-1.444	0.266	1.222	-1.326	0.266
25	0.646	-0.411	0.270	0.966	-0.569	0.270
26	0.872	-0.572	0.180	0.901	-0.509	0.180
27	0.951	-0.036	0.351	1.136	0.606	0.351
28	1.047	0.738	0.216	0.992	0.611	0.216
29	0.770	-2.918	0.241	1.018	-2.917	0.241
30	0.338	-2.932	0.261	0.545	-1.999	0.261
31	0.685	-0.002	0.212	0.904	-0.219	0.212
32	0.797	-0.828	0.285	0.832	-1.114	0.285
33	0.792	-1.032	0.323	0.914	-1.032	0.323
34	0.522	0.496	0.366	0.661	0.654	0.366
35	0.963	-1.200	0.267	1.111	-1.184	0.267
36	1.059	0.310	0.150	1.025	0.222	0.150
37	0.933	-0.098	0.345	0.968	0.051	0.345
38	1.195	-0.365	0.266	0.745	-0.740	0.266
39	0.831	0.235	0.155	0.978	0.357	0.155
40	1.028	-0.260	0.313	0.725	-0.007	0.313
41	1.008	-0.536	0.292	1.315	-0.251	0.292
42	0.522	-1.818	0.223	0.949	-0.980	0.223

Item Number	Non-ELL Group			ELL Group		
	a-parameter (slope)	b-parameter (threshold)	Common c-parameter (asymptote)	a-parameter (slope)	b-parameter (threshold)	Common c-parameter (asymptote)
43	1.100	-0.987	0.238	1.027	-0.859	0.238
44	0.793	0.144	0.226	1.003	0.109	0.226
45	0.989	-0.152	0.251	1.100	-0.244	0.251
46	0.800	-0.644	0.321	0.878	-0.446	0.321
47	0.776	0.325	0.193	0.883	0.021	0.193
48	0.438	-1.242	0.298	0.513	-0.726	0.298
49	0.792	-0.112	0.436	0.652	-0.343	0.436
50	0.802	-0.780	0.382	0.668	-1.019	0.382
51	1.612	0.532	0.140	1.564	0.579	0.140
52	1.353	0.263	0.262	1.033	0.137	0.262
53	0.538	-1.946	0.268	0.723	-1.629	0.268
54	0.708	-0.931	0.231	1.074	-0.827	0.231
55	0.730	-0.652	0.336	0.785	-0.737	0.336
56	0.876	-0.249	0.186	1.271	-0.291	0.186
57	0.752	-0.565	0.128	1.213	-0.456	0.128
58	0.458	-2.025	0.264	0.525	-1.507	0.264
59	0.838	-0.129	0.487	1.038	-0.466	0.487
60	0.661	-1.082	0.180	0.865	-1.097	0.180
61	0.599	-1.779	0.259	1.169	-1.323	0.259
62	0.887	-0.643	0.234	1.124	-0.769	0.234
63	1.266	0.362	0.142	1.111	0.175	0.142
64	0.806	0.614	0.202	1.058	0.451	0.202
65	0.646	-0.272	0.212	0.717	-0.759	0.212
66	0.663	-0.931	0.246	0.716	-1.152	0.246
67	0.928	-0.969	0.315	0.831	-0.972	0.315
68	0.897	3.266	0.294	1.053	2.742	0.294
69	0.619	-0.260	0.229	0.448	-0.104	0.229
70	0.677	0.226	0.277	0.510	0.036	0.277
71	0.741	-0.059	0.320	0.829	0.161	0.320
72	0.577	-0.580	0.294	0.457	-0.562	0.294
73	0.751	-2.842	0.247	0.624	-4.468	0.247
74	0.576	-1.851	0.254	0.406	-2.346	0.254
75	1.014	0.604	0.296	0.719	0.462	0.296
76	0.664	1.083	0.244	0.893	1.160	0.244
77	0.544	-1.656	0.229	0.562	-1.514	0.229
78	0.723	-0.155	0.300	0.408	0.348	0.300
79	0.663	-0.968	0.260	0.672	-1.006	0.260
80	0.970	-0.611	0.270	1.062	-0.314	0.270
81	0.899	-0.463	0.249	1.014	-0.352	0.249
82	0.761	1.000	0.234	0.830	0.863	0.234
83	0.551	0.643	0.307	0.799	0.859	0.307
84	0.691	0.634	0.234	0.722	0.817	0.234

Table 6

Area (Closed Form Solution with Common c-parameters) between ICCs in IRT-based Method.

Item Number	Area	Item Number	Area	Item Number	Area	Item Number	Area
1	0.1327	**26**	0.0533	**51**	0.0400	**76**	0.2429
2	0.2372	**27**	0.4167	**52**	0.1589	**77**	0.1096
3	0.1703	**28**	0.1007	**53**	0.3458	**78**	0.6788
4	0.2804	**29**	0.1953	**54**	0.3095	**79**	0.0286
5	0.1220	**30**	0.9027	**55**	0.0709	**80**	0.2169
6	0.3406	**31**	0.2704	**56**	0.2374	**81**	0.1055
7	0.1889	**32**	0.2044	**57**	0.3686	**82**	0.1158
8	0.3835	**33**	0.0932	**58**	0.3917	**83**	0.3429
9	0.1987	**34**	0.2259	**59**	0.1839	**84**	0.1407
10	0.1434	**35**	0.0833	**60**	0.2384		
11	0.2437	**36**	0.0751	**61**	0.5689		
12	0.2410	**37**	0.0979	**62**	0.1698		
13	0.2694	**38**	0.3840	**63**	0.1667		
14	0.2953	**39**	0.1531	**64**	0.2218		
15	0.0897	**40**	0.2718	**65**	0.3850		
16	0.1946	**41**	0.2242	**66**	0.1703		
17	0.1030	**42**	0.7894	**67**	0.0702		
18	0.2707	**43**	0.0996	**68**	0.3707		
19	0.2242	**44**	0.1681	**69**	0.4010		
20	0.0505	**45**	0.0864	**70**	0.3079		
21	0.1127	**46**	0.1386	**71**	0.1578		
22	0.2191	**47**	0.2511	**72**	0.2620		
23	0.4962	**48**	0.3824	**73**	1.2242		
24	0.2269	**49**	0.1683	**74**	0.5446		
25	0.3200	**50**	0.1804	**75**	0.2468		

Table 7

Numbers of No-DIF and DIF Items in IRT-based Method.

	No-DIF	**DIF**
Numbers of Item	66	18

Types and Numbers of DIF Items in IRT-based Method.

	Uniform-DIF	**Non-Uniform DIF**
Item Number	8, 27, 48, 58, 65, 68, 73	23, 30, 38, 42, 53, 57, 61, 69, 74, 78, 83
Numbers of Item	7	11

Table 8

Detecting DIF using Mantel-Haenszel Method in DIFAS --- Non-ELL Students vs. ELL Students.

Item Number	Mantel-Haenszel χ^2	Mantel-Haenszel Common Log-Odds Ratio	Standard Error of Mantel-Haenszel Log-Odds Raito	Breslow-Day χ^2	ETS Categorization Scheme
1	0.048	-0.156	0.381	0.554	A
2	0.568	-0.216	0.248	0.034	A
3	0.006	0.052	0.254	0.316	A
4	0.106	-0.156	0.320	0.666	A
5	0.351	0.164	0.226	0.565	A
6	0.550	-0.262	0.301	0.441	A
7	0.007	-0.007	0.227	1.710	A
8	0.315	-0.123	0.186	0.578	A
9	0.207	-0.151	0.257	1.521	A
10	0.027	0.434	0.759	0.506	A
11	1.116	0.234	0.204	0.221	A
12	2.333	-0.373	0.229	0.036	A
13	0.308	0.126	0.189	0.309	A
14	0.301	-0.134	0.206	0.011	A
15	0.533	0.167	0.202	0.179	A
16	0.997	0.207	0.189	0.102	A
17	1.215	0.250	0.210	0.034	A
18	2.166	-0.315	0.198	0.944	A
19	0.834	-0.428	0.402	0.003	A
20	0.064	-0.077	0.213	0.055	A
21	2.087	-0.409	0.263	0.176	A
22	2.664	-0.375	0.217	0.196	A
23	0.005	0.015	0.253	0.829	A
24	0.145	-0.150	0.284	0.883	A
25	1.226	-0.262	0.212	0.803	A
26	0.084	0.081	0.205	0.041	A
27	**9.878 ****	**0.596**	**0.187**	**1.681**	**B**
28	0.064	-0.066	0.188	0.243	A
29	0.645	-0.711	0.665	0.633	A
30	0.009	0.009	0.260	0.752	A
31	0.882	-0.199	0.193	0.404	A
32	1.448	-0.306	0.234	0.103	A
33	0.027	-0.069	0.242	0.181	A
34	0.133	0.082	0.178	0.410	A
35	0.008	-0.011	0.268	0.297	A
36	0.011	-0.040	0.197	0.297	A
37	1.483	0.247	0.190	0.018	A
38	1.095	-0.235	0.210	0.527	A
39	1.500	0.251	0.191	0.003	A
40	**1.863**	**0.272**	**0.191**	**4.289 ***	**A**
41	**4.453 ***	**0.462**	**0.208**	**0.084**	**B**
42	1.485	0.317	0.232	1.464	A

** Significant at 0.05 level*
*** Significant at 0.01 level*

Item Number	Mantel-Haenszel χ^2	Mantel-Haenszel Common Log-Odds Ratio	Standard Error of Mantel-Haenszel Log-Odds Raito	Breslow-Day χ^2	ETS Categorization Scheme
43	0.790	0.244	0.239	0.354	A
44	0.005	0.004	0.189	0.022	A
45	0.053	-0.066	0.201	0.247	A
46	0.476	0.158	0.204	0.559	A
47	1.623	-0.263	0.190	0.381	A
48	1.528	0.271	0.203	0.073	A
49	0.077	-0.077	0.205	0.000	A
50	0.157	-0.118	0.232	0.922	A
51	0.545	0.173	0.208	0.029	A
52	0.339	-0.130	0.191	0.000	A
53	0.000	-0.037	0.263	0.251	A
54	0.198	-0.120	0.218	1.232	A
55	0.734	-0.215	0.222	0.053	A
56	0.001	-0.028	0.205	1.989	A
57	0.056	0.073	0.209	2.385	A
58	1.174	0.271	0.227	0.004	A
59	1.707	-0.307	0.218	1.795	A
60	0.415	-0.173	0.226	0.543	A
61	0.004	-0.058	0.279	3.615	A
62	1.394	-0.294	0.225	0.467	A
63	0.745	-0.197	0.204	0.125	A
64	0.064	0.064	0.185	1.116	A
65	**4.747 ***	**-0.463**	**0.204**	**0.090**	**B**
66	1.426	-0.288	0.223	0.088	A
67	0.015	0.055	0.232	0.481	A
68	0.003	-0.008	0.189	0.097	A
69	0.429	0.136	0.185	3.188	A
70	0.574	-0.152	0.182	0.790	A
71	1.283	0.228	0.186	0.117	A
72	0.053	0.064	0.198	2.805	A
73	3.166	-1.283	0.700	0.164	A
74	0.004	0.015	0.249	2.513	A
75	0.620	-0.162	0.184	0.002	A
76	1.604	0.257	0.187	0.000	A
77	0.029	0.068	0.234	0.304	A
78	**1.253**	**0.219**	**0.184**	**8.003****	**A**
79	0.002	-0.034	0.219	0.069	A
80	3.250	0.401	0.208	0.001	A
81	0.394	0.153	0.206	0.002	A
82	0.079	-0.067	0.180	0.000	A
83	1.497	0.234	0.178	0.206	A
84	0.086	0.070	0.182	0.573	A

* *Significant at 0.05 level*
** *Significant at 0.01 level*

Table 9

Types and Numbers of DIF Items in Mantel-Haenszel Method.

Omnibus Test --- Non-ELL Students vs. ELL Students.

	Uniform-DIF	**Non-Uniform DIF**
Item Number	27, 41, 65	40, 78
Numbers of Item	3	2

Follow-up Test --- Hispanic Non-ELL Students vs. Hispanic ELL Students.

	Uniform-DIF	**Non-Uniform DIF**
Item Number	32, 48	7, 18, 69, 78, 79
Numbers of Item	2	5

Follow-up Test --- White Non-ELL Students vs. Hispanic ELL Students.

	Uniform-DIF	**Non-Uniform DIF**
Item Number	NA	27, 69, 72, 78
Numbers of Item	0	4

Table 10a

Detecting DIF using Mantel-Haenszel Method in DIFAS --- Hispanic Non-ELL Students vs. Hispanic ELL Students.

Item Number	Mantel-Haenszel χ^2	Mantel-Haenszel Common Log-Odds Ratio	Standard Error of Mantel-Haenszel Log-Odds Raito	Breslow-Day χ^2	ETS Categorization Scheme
1	0.033	-0.200	0.480	0.386	A
2	0.128	-0.156	0.307	1.148	A
3	0.465	0.285	0.329	0.050	A
4	0.538	-0.375	0.398	1.107	A
5	0.000	-0.043	0.285	0.689	A
6	0.101	-0.183	0.374	1.044	A
7	**0.000**	**-0.044**	**0.286**	**4.279***	**A**
8	0.002	0.040	0.244	0.016	A
9	0.172	-0.196	0.334	0.349	A
10	0.176	-0.047	0.899	1.134	A
11	0.003	-0.050	0.267	0.282	A
12	0.629	-0.270	0.287	0.319	A
13	0.449	0.195	0.242	0.005	A
14	0.043	-0.089	0.263	0.084	A
15	0.899	0.272	0.253	0.599	A
16	0.603	0.204	0.233	0.026	A
17	0.029	0.079	0.265	0.168	A
18	**0.807**	**-0.269**	**0.259**	**4.189***	**A**
19	1.676	-0.707	0.499	1.905	A
20	0.317	0.195	0.276	0.010	A
21	0.596	-0.314	0.337	0.928	A
22	0.004	-0.052	0.265	0.622	A
23	0.001	-0.047	0.337	0.399	A
24	0.383	-0.271	0.354	2.414	A
25	1.291	-0.365	0.283	0.409	A
26	0.375	0.199	0.266	0.021	A
27	3.348	0.452	0.235	2.571	A
28	0.053	-0.082	0.237	0.245	A
29	1.426	-1.300	0.911	2.661	A
30	0.030	-0.121	0.348	3.280	A
31	0.259	-0.161	0.251	0.973	A
32	**7.098***	**-0.847**	**0.311**	**0.945**	**C**
33	0.252	0.220	0.327	0.315	A
34	0.036	0.074	0.238	2.047	A
35	0.007	0.029	0.339	0.430	A
36	0.043	-0.085	0.253	0.027	A
37	0.494	0.210	0.256	0.327	A
38	0.158	-0.145	0.269	0.005	A
39	1.711	0.360	0.251	0.124	A
40	0.101	0.113	0.251	0.577	A

** Significant at 0.05 level*
*** Significant at 0.01 level*

Item Number	Mantel-Haenszel χ^2	Mantel-Haenszel Common Log-Odds Ratio	Standard Error of Mantel-Haenszel Log-Odds Raito	Breslow-Day χ^2	ETS Categorization Scheme
41	3.828	0.575	0.273	0.04	A
42	3.057	0.661	0.338	1.39	A
43	3.403	0.573	0.287	0.001	A
44	0.189	0.136	0.243	0.627	A
45	0.002	-0.023	0.258	0.551	A
46	0.782	0.250	0.256	1.439	A
47	2.009	-0.389	0.251	0.675	A
48	**5.078***	**0.605**	**0.263**	**1.714**	**B**
49	0.101	0.122	0.265	0.169	A
50	0.231	0.190	0.300	0.674	A
51	1.336	0.356	0.281	1.323	A
52	0.001	-0.041	0.251	0.039	A
53	0.092	-0.160	0.336	0.237	A
54	0.252	-0.164	0.265	3.644	A
55	1.164	-0.348	0.284	0.016	A
56	0.002	0.020	0.250	1.101	A
57	0.005	0.014	0.258	1.954	A
58	0.000	-0.057	0.319	0.46	A
59	1.213	-0.338	0.278	0.512	A
60	1.396	-0.375	0.286	0.272	A
61	0.091	0.174	0.357	1.16	A
62	0.923	-0.304	0.276	1.822	A
63	0.204	0.152	0.261	0.064	A
64	0.258	0.146	0.236	0.302	A
65	0.590	-0.228	0.254	0.05	A
66	0.002	-0.031	0.294	0.049	A
67	0.149	0.168	0.308	0.011	A
68	0.109	0.113	0.247	2.229	A
69	**1.196**	**0.280**	**0.238**	**9.810****	**A**
70	0.334	-0.165	0.237	0.039	A
71	0.004	-0.042	0.237	0.005	A
72	0.001	-0.025	0.245	0.398	A
73	1.335	-1.452	1.049	0.745	A
74	0.785	-0.376	0.353	0.93	A
75	0.059	-0.087	0.238	0.195	A
76	0.042	-0.083	0.250	0.218	A
77	0.012	0.079	0.300	1.768	A
78	**0.250**	**0.148**	**0.239**	**4.360***	**A**
79	**1.161**	**-0.359**	**0.291**	**4.100***	**A**
80	0.426	0.209	0.263	0.759	A
81	0.188	-0.146	0.259	0.001	A
82	0.141	0.114	0.231	0.074	A
83	0.030	0.068	0.234	0.58	A
84	1.504	-0.323	0.238	1.445	A

** Significant at 0.05 level*
*** Significant at 0.01 level*

Table 10b

Detecting DIF using Mantel-Haenszel Method in DIFAS --- White Non-ELL Students vs. Hispanic ELL Students.

Item Number	Mantel-Haenszel χ^2	Mantel-Haenszel Common Log-Odds Ratio	Standard Error of Mantel-Haenszel Log-Odds Raito	Breslow-Day χ^2	ETS Categorization Scheme
1	0.002	0.080	0.446	0.616	A
2	0.641	-0.292	0.304	0.000	A
3	0.003	0.031	0.304	0.388	A
4	0.210	0.266	0.392	1.177	A
5	0.006	0.058	0.272	1.718	A
6	0.130	-0.198	0.364	0.210	A
7	0.144	0.138	0.271	2.095	A
8	0.066	-0.089	0.235	0.426	A
9	0.141	-0.184	0.334	0.242	A
10	1.192	2.094	1.209	1.420	A
11	0.050	0.093	0.260	0.460	A
12	1.871	-0.396	0.268	0.015	A
13	0.101	0.106	0.235	0.034	A
14	0.153	-0.127	0.247	0.489	A
15	0.010	0.055	0.247	1.271	A
16	0.125	0.112	0.236	0.158	A
17	2.894	0.457	0.255	0.066	A
18	1.003	-0.277	0.244	1.107	A
19	0.028	-0.046	0.529	0.651	A
20	0.305	-0.173	0.256	0.414	A
21	0.825	-0.326	0.312	0.867	A
22	0.595	-0.227	0.256	0.007	A
23	0.002	-0.062	0.314	1.043	A
24	0.010	-0.025	0.343	0.589	A
25	0.769	-0.254	0.250	2.127	A
26	0.005	-0.049	0.250	0.230	A
27	**9.107****	**0.707**	**0.231**	**6.783****	**B**
28	1.247	-0.281	0.231	0.361	A
29	0.000	-0.388	0.927	0.428	A
30	0.749	-0.364	0.351	1.032	A
31	2.774	-0.419	0.239	1.191	A
32	3.070	-0.579	0.308	0.129	A
33	0.115	-0.138	0.288	0.060	A
34	0.280	0.133	0.212	0.304	A
35	0.237	-0.209	0.323	0.657	A
36	0.000	0.032	0.239	0.562	A
37	0.489	0.190	0.232	0.000	A
38	1.219	-0.307	0.252	0.185	A
39	0.239	0.138	0.231	0.098	A
40	0.508	0.193	0.238	3.148	A

** Significant at 0.05 level*
*** Significant at 0.01 level*

Item Number	Mantel-Haenszel χ^2	Mantel-Haenszel Common Log-Odds Ratio	Standard Error of Mantel-Haenszel Log-Odds Raito	Breslow-Day χ^2	ETS Categorization Scheme
41	2.267	0.429	0.261	0.313	A
42	0.143	0.149	0.283	1.069	A
43	3.701	0.637	0.302	0.934	A
44	0.029	-0.067	0.232	0.009	A
45	1.205	-0.293	0.245	0.050	A
46	1.616	0.342	0.248	1.395	A
47	1.034	-0.263	0.229	2.793	A
48	0.738	0.252	0.253	0.075	A
49	0.004	-0.015	0.243	0.029	A
50	0.046	-0.094	0.271	1.888	A
51	0.056	0.092	0.253	0.164	A
52	0.578	-0.200	0.231	0.098	A
53	0.000	-0.046	0.324	0.229	A
54	0.127	0.131	0.266	0.792	A
55	0.004	0.018	0.265	0.298	A
56	0.331	0.176	0.248	1.522	A
57	0.613	0.242	0.255	1.778	A
58	0.066	-0.122	0.302	0.464	A
59	0.900	-0.290	0.268	1.133	A
60	0.522	-0.236	0.275	0.224	A
61	0.001	-0.049	0.341	2.354	A
62	0.419	-0.208	0.267	0.075	A
63	0.757	-0.247	0.250	0.062	A
64	0.094	-0.090	0.218	0.581	A
65	2.398	-0.393	0.237	0.052	A
66	1.267	-0.333	0.270	0.586	A
67	0.026	-0.085	0.285	2.265	A
68	0.058	0.084	0.235	0.001	A
69	**2.477**	**0.365**	**0.223**	**6.114***	**A**
70	0.215	-0.128	0.224	0.406	A
71	2.800	0.403	0.230	1.886	A
72	**0.050**	**0.083**	**0.243**	**4.108***	**A**
73	2.461	-1.487	0.933	0.324	A
74	0.752	-0.350	0.343	1.284	A
75	0.088	-0.092	0.225	0.243	A
76	1.713	0.325	0.224	0.360	A
77	0.000	0.039	0.276	0.336	A
78	**1.586**	**0.291**	**0.221**	**11.831****	**A**
79	1.040	-0.341	0.289	0.456	A
80	2.533	0.455	0.260	0.092	A
81	0.749	0.252	0.252	0.110	A
82	0.005	-0.009	0.221	0.002	A
83	0.671	0.201	0.216	0.017	A
84	1.271	0.281	0.224	0.082	A

* *Significant at 0.05 level*
** *Significant at 0.01 level*

Table 11

Comparison Between IRT-based Area and Mantel-Haenszel Methods

Item	**IRT-based Area Method**			**Mantel-Haenszel Method**				**Agreement between IRT-based Area and Mantel-Haenszel Methods**
	Area between ICCs	Uniform DIF	Non-Uniform DIF	Mantel-Haenszel χ^2	Breslow-Day χ^2	Uniform DIF	Non-Uniform DIF	
1	0.133			0.048	0.554			
2	0.237			0.568	0.034			
3	0.170			0.006	0.316			
4	0.280			0.106	0.666			
5	0.122			0.351	0.565			
6	0.341			0.550	0.441			
7	0.189			0.007	1.710			
8	0.384*	✓		0.315	0.578			
9	0.199			0.207	1.521			
10	0.143			0.027	0.506			
11	0.244			1.116	0.221			
12	0.241			2.333	0.036			
13	0.269			0.308	0.309			
14	0.295			0.301	0.011			
15	0.090			0.533	0.179			
16	0.195			0.997	0.102			
17	0.103			1.215	0.034			
18	0.271			2.166	0.944			
19	0.224			0.834	0.003			
20	0.051			0.064	0.055			
21	0.113			2.087	0.176			
22	0.219			2.664	0.196			
23	0.496*		✓	0.005	0.829			
24	0.227			0.145	0.883			
25	0.320			1.226	0.803			
26	0.053			0.084	0.041			
27	0.417*	✓		9.878**	1.681	✓		✓
28	0.101			0.064	0.243			
29	0.195			0.645	0.633			
30	0.903*		✓	0.009	0.752			
31	0.270			0.882	0.404			
32	0.204			1.448	0.103			
33	0.093			0.027	0.181			
34	0.226			0.133	0.410			
35	0.083			0.008	0.297			
36	0.075			0.011	0.297			
37	0.098			1.483	0.018			
38	0.384*		✓	1.095	0.527			
39	0.153			1.500	0.003			
40	0.272			1.863	4.289**		✓	
41	0.224			4.453**	0.084	✓		
42	0.789*		✓	1.485	1.464			

** Area is greater than cutoff value, 0.3414*
***χ^2 is greater than critical value, 3.84, at 0.05 significance level*

Item	IRT-based Area Method			Mantel-Haenszel Method				Agreement between IRT-based Area and Mantel-Haenszel Methods
	Area between ICCs	Uniform DIF	Non-Uniform DIF	Mantel-Haenszel χ^2	Breslow-Day χ^2	Uniform DIF	Non-Uniform DIF	
43	0.100			0.790	0.354			
44	0.168			0.005	0.022			
45	0.086			0.053	0.247			
46	0.139			0.476	0.559			
47	0.251			1.623	0.381			
48	0.382*	✓		1.528	0.073			
49	0.168			0.077	0.000			
50	0.180			0.157	0.922			
51	0.040			0.545	0.029			
52	0.159			0.339	0.000			
53	0.346*		✓	0.000	0.251			
54	0.310			0.198	1.232			
55	0.071			0.734	0.053			
56	0.237			0.001	1.989			
57	0.369*		✓	0.056	2.385			
58	0.392*	✓		1.174	0.004			
59	0.184			1.707	1.795			
60	0.238			0.415	0.543			
61	0.569*		✓	0.004	3.615			
62	0.170			1.394	0.467			
63	0.167			0.745	0.125			
64	0.222			0.064	1.116			
65	0.385*	✓		4.747**	0.090	✓		✓
66	0.170			1.426	0.088			
67	0.070			0.015	0.481			
68	0.371*	✓		0.003	0.097			
69	0.401*		✓	0.429	3.188			
70	0.308			0.574	0.790			
71	0.158			1.283	0.117			
72	0.262			0.053	2.805			
73	1.224*	✓		3.166	0.164			
74	0.545*		✓	0.004	2.513			
75	0.247			0.620	0.002			
76	0.243			1.604	0.000			
77	0.110			0.029	0.304			
78	0.679*		✓	1.253	8.003**		✓	✓
79	0.029			0.002	0.069			
80	0.217			3.250	0.001			
81	0.105			0.394	0.002			
82	0.116			0.079	0.000			
83	0.343*		✓	1.497	0.206			
84	0.141			0.086	0.573			

* *Area is greater than cutoff value, 0.3414*

** *χ^2 is greater than critical value, 3.84, at 0.05 significance level*

Figure Caption

Figure 1: Item Characteristics Curves, for example, Item 27 and Item 78.