Caribbean Curriculum
Vol. 14, 2007, 49–71

DIFFERENTIAL ITEM FUNCTIONING AND MALE-FEMALE DIFFERENCES IN A LARGE-SCALE MATHEMATICS ASSESSMENT IN TRINIDAD AND TOBAGO
An Examination of Standard 1 Mathematics Assessment

Launcelot I. Brown and *Gibbs Y. Kanyongo*

This study investigates gender differences and the existence of gender-related differential item functioning (DIF) in a large-scale Standard 1 mathematics assessment in Trinidad and Tobago. Although research consistently shows that mathematics scores for male students are usually higher than for female students at the secondary and tertiary level, the differences are not very clear at the primary level. Actually, results from this study show that female students performed slightly better than male students on this examination. Logistic regression procedure was used to detect DIF items, and the results show that about 17% of the items in the examination displayed gender-related DIF; however, for all DIF items the effect sizes were negligible.

Gender Differences in Mathematics Achievement

Research on differences in mathematics performance related to gender, although waning at times, has continued over the years (Fennema, Carpenter, Jacobs, Franke, & Levi, 1998; Friedman, 1989; Hyde, Fennema, & Lamon, 1990; Maccoby & Jacklin, 1974). Performance differences in mathematics have been postulated to be partly due to attitudinal differences regarding mathematics (Duffy, Gunther, & Walters, 1997; Fennema & Sherman, 1977; Forgasz & Leder, 1996; Meyer & Koehler, 1990; Stipek & Gralinski, 1991; Tocci & Engelhard, 1991), and parental expectations that suggest a belief in the mathematical superiority of boys over girls (Blevins-Knabe, Austin, Musun, Eddy, & Jones, 2000). This parental belief has been found to be predictive of students' achievement in mathematics courses (Parsons, Adler, & Kaczala, 1982) and students' belief in their own mathematical abilities (Tiedemann, 2000).

It has also been noted that teachers' attitudes toward students' performance in mathematics classes parallel those of the parents. Girls are seen as successful due to their hard work (Jussim & Eccles, 1992; Siegle & Reis, 1998; Tiedemann, 2000), while boys' success is attributed to their talent (Jussim & Eccles, 1992). With regard to the English-speaking Caribbean, there is some evidence that teachers also hold differing perceptions of students' mathematical ability based on gender. However, these perceptions contrast with those articulated in the existing literature. The limited research conducted in this area has shown that, for various reasons, teachers respond more positively to girls than to boys (Kutnick, 2000; Kutnick, Jules & Layne, 1997; Parry, 2000). Additionally, on mandated exams and on teacher-made tests, girls perform better than boys at all ages between 8 and 16, in all curriculum areas, and across all subjects including subjects within the sciences (Kutnick et al., 1997).

Brown (2005) observed that among seven- and eight-year-olds in Trinidad and Tobago, on the mathematics component of the 2000 Continuous Assessment Programme (CAP), overall, girls performed better than boys, a higher proportion of boys than girls was in the lower tail of the distribution, and non-response on items was higher among boys than girls. However, this general finding did not apply to students at or above the 84th percentile where boys had the higher mean.

Hanna (2003) noted that gender differences have decreased to the point that several countries have in effect achieved gender equity in mathematics. Tapia and Marsh (2004), in their review of the relevant literature, also report the non-existence of measurable differences in the achievement scores of boys and girls on math tests at age nine. Furthermore, differences that began to become apparent at age 13 seem to have all but disappeared since 1994. While these findings are in the desired direction, the Trinidad and Tobago and Caribbean data suggest a continuing differential in favour of girls at the primary level but allow for less definitive interpretations at the secondary level (Brown, 2005).

Explaining gender differences in mathematics achievement

Gallagher et al. (2000) presented a taxonomy of content and cognitive characteristics that attempts to account for gender differences in mathematics. The taxonomy was based on outcomes reported in the educational and psychological literature. According to the Gallagher et al. taxonomy, females should perform better than males on items with contextual characteristics likely to be more familiar or interesting to females, on items that require a high level of verbal skill, and on items that require mastery of mathematical content. Males should perform better than females on items that have contextual characteristics likely to be more familiar or interesting to males, on items that are likely to place heavy demands on spatial skills, and on items that have multiple solution paths.

In their study, Gierl, Bisanz, Bisanz, and Boughton (2003) found that males perform better than females on items that require significant spatial processing. This finding was consistent with previous research (e.g., Casey, Nuttall, Pezaris, & Benbow, 1995; Halpern, 1997). They also found that females outperformed males on items requiring memorization, but the differences were small.

Differential Item Functioning (DIF)

Differential item functioning (DIF) occurs when one group is more likely than another to answer an item correctly when both groups are similar on the trait that is being assessed. In DIF terminology, a sample from the population of interest is referred to as the focal group and the sample from the comparable population is referred to as the reference group. Group similarity is usually established by use of total raw score or some measure of the trait (Meyer, Huynh, & Seaman, 2004). In this study, girls comprise the reference group and boys the focal group.

DIF occurs when individuals in a focal group respond differently to a test item than do individuals in a reference group, even when comparisons are restricted to individuals with similar overall skill levels on the trait in question (Johanson, 1997). Dodeen and Johanson (1997) note that the Educational Testing Service (ETS) classifies DIF into three categories: Category A contains items with negligible DIF; Category B for

intermediate DIF items; and Category C for large DIF items. This classification is usually used for investigating DIF in testing situations.

DIF is important in detecting multidimensionality in a test. Multidimensionality occurs when an item assesses a competency other than that being assessed by the test. The source of the multidimensionality for a differentially functioning item may or may not be relevant to the psychological construct being assessed by a particular test in question. Such items unfairly punish those students who might not be familiar with the competency that is not relevant to the test, while unfairly rewarding those who might be familiar with it. It is therefore important to identify and flag items displaying multidimensionality in a test because all items on a test should be one-dimensional, that is, measuring only one construct.

There are a number of reasons that lead to DIF occurring in a test. The test may have items which have negative wording that is confusing to one group. In addition, some test items present the content in ways that prevent certain groups of examinees from demonstrating their knowledge or traits. The way an item is worded or structured may introduce a source of difficulty that is not relevant to the construct being measured. Thus, the item may be measuring two or more constructs, of which only one is relevant to the purpose of the instrument. The extraneous construct may influence the responses of one group differently. However, how it influences the primary construct may not be explicit (Camilli & Shepard, 1994). For example, on a test of math achievement, a weakness in reading comprehension may prevent a group of students from accurately completing math word problems. In this instance, reading comprehension is not the construct of interest, but the examinees' reading level influences their responses.

Test and item bias

At this juncture, it is important that we distinguish between DIF and test bias. Camilli and Shepard (1994) have defined test bias "as invalidity or systematic error in how a test measures for members of a particular group.” They further state that “bias is systematic in the sense that it creates a distortion in test results for members of a particular group" (p.

8). That is, it allows for invalid conclusions to be drawn with regard to the performance of a particular group.

But test bias is as a result of item bias. Items are biased when they exhibit DIF. It is important to note that the existence of substantial mean differences between two groups on a test is not sufficient evidence to consider a test biased (Camilli & Shepard, 1994), because true differences could exist between the groups on the underlying construct being assessed by the instrument. The empirical evidence that leads to the conclusion that an item is biased is the extent to which there is *differential item functioning* (DIF) (Hambleton, Swaminathan, & Rogers, 1991).

DIF and gender

Gender DIF is a major concern on large-scale achievement tests in mathematics because differences between females and males are often found (e.g., Bielinski & Davison, 2001; Boughton, Gierl, & Khaliq, 2000; DeMars, 1998; Garner & Engelhard, 1999; Scheuneman & Grima, 1997; Willingham & Cole, 1997). Over the years, many studies have looked at DIF and gender on mathematics items, for example, Doolittle & Cleary (1987), Ryan & Chiu (2001), Scheuneman & Grima, (1997), and Wang & Lane (1996). Harris and Carlton (1993), controlling on the total test score using the Mantel-Haenszel (M-H) procedure, observed that on the Scholastic Aptitude Test (SAT), there were "identifiable patterns of gender differences" (p. 137) in the overall performance of male and female students. While girls performed differentially better than boys in algebra and on textbook-like items, confirming the previous findings of Doolittle and Cleary, the opposite was found for geometry items, real-life non-textbook-like items, and items requiring higher-level mental processing. With the exception of Wang and Lane, whose sample comprised sixth and seventh grade students (age range 11–12 years), all the other studies examined DIF for items on examinations taken at the senior high school or college level.

DIF Detection Procedures

Some of the commonly used psychometric procedures that detect DIF in dichotomously scored test items are the Mantel-Haenszel (MH), the Simultaneous Item Bias Test (SIBTEST) and the logistic regression (LR) procedures.

The Mantel-Haenszel (MH)

The MH statistic has been one of the most widely used procedures to evaluate DIF (Clauser & Mazor, 1998). The MH procedure was proposed by Holland and Thayer (1988) as a technique for detecting DIF. Since its proposal, there have been many studies about its efficacy in the identification of items with DIF under different conditions, such as DIF magnitude, DIF type, sample size, and sample ability differences (Hidalgo & Lopez-Pina, 2004). Despite its wide use, results from previous studies have indicated that it lacks power in detecting non-uniform DIF (Swaminathan & Rogers, 1990).

SIBTEST

SIBTEST is a nonparametric procedure that both estimates the amount of DIF in an item and statistically tests whether that amount is different from 0 (Bolt, 2000). SIBTEST uses a regression correction method to match examinees from the two groups at the same latent ability levels so as to compare their performances on the studied items. It requires an initial distinction between two non-overlapping subsets of items in the test: (1) a valid subtest, which contains items that are assumed to measure the ability the test is designed to measure; and (2) a suspect subtest, which contains the items being tested for DIF. Scores on the valid subtest are used to match examinees having the same ability levels across groups in order to test items from the suspect subtest for DIF (Bolt, 2000).

Logistic regression procedure

This is the DIF procedure that is used in this study. In DIF detection with logistic regression, the probability of item score (correct vs. incorrect) is predicted from a constant, total test score, group membership, and the

interaction between group membership and total score (Swaminathan & Rogers, 1990). Logistic regression is model-based, therefore, uniform and non-uniform DIF can be modelled in the same equation and coefficients for each can be tested separately (French & Miller, 1996).

The logistic regression equation is expressed in general terms as:

$$p(u = 1/x) = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

where p (u = 1/**x**) is the conditional probability of obtaining a correct answer given **x** (vector of independent variables), and f(**x**) is the function that defines the linear combination of the predictor variables. In DIF analysis, the dependent variable is the item score, and the independent variables are the group variable *G* (usually gender or ethnicity), the observed ability θ as a matching criterion (usually the total test score), and the group by ability interaction θG. The logistic regression model is therefore expressed as:

$$f(x) = \tau_0 + \tau_1\theta + \tau_2 G + \tau_3\theta G$$

where τ_0 is the intercept, τ_1 is the ability regression coefficient, τ_2 is the coefficient for the group variable, and τ_3 is the coefficient for the ability*group interaction parameter. Logistic regression is a highly effective technique for detecting DIF in dichotomous items (French & Miller, 1996; Swaminathan & Rogers, 1990; Zumbo, 1999).

Zumbo and Thomas (1997) proposed ΔR^2, a weighted least squares effect size measure when using logistic regression in DIF detection, to quantify the magnitude of uniform or non-uniform DIF. The guidelines they provided were:

Type A items – negligible DIF: $\Delta R^2 < 0.13$
Type B items – moderate DIF: $0.13 \leq \Delta R^2 \leq 0.26$
Type C items – large DIF: $\Delta R^2 > 0.26$

However, Jodoin and Gierl (2001), based on their simulation study, have advised caution when interpreting the effect size based on the Zumbo

and Thomas (1997) guidelines, and have instead suggested the following guidelines as more accurately quantifying the magnitude of DIF:

Type A items – negligible DIF: $\Delta R^2 < .035$

Type B items – moderate DIF: $.035 \leq \Delta R^2 \leq .070$

Type C items – large DIF: $\Delta R^2 > .070$

In this study, we use the Jodoin and Gierl criteria in interpreting the effect size of the DIF.

Purpose of the Study

A common feature of most of the previous studies on gender DIF is that none of them were done in the Caribbean context using Caribbean data. It is therefore the objective of the current study to accomplish this.

The National Test is administered annually to students in Standard 1 (Std 1)—age range 7 to 8 years—in public and private primary schools. It is based on the national curriculum, and assesses two curriculum areas—Mathematics and Language Arts. The purpose of the test is diagnostic, but as has happened with many large-scale examinations, in the eyes of the public and many school administrators, the test is seen as rating the performance of the schools and, as a result, classifying schools into good schools and bad schools.

Based on student performance on the tests, schools are assigned various support personnel, but also receive greater critical oversight from school supervisors. Unfortunately, this has created an atmosphere of anxiety and competition around the tests, and has resulted in similar kinds of pressure on students, teachers, and administrators as seen in cases where tests results are used for high-stakes decisions. Because decisions are made based on students' test scores, it is prudent for examination boards to demonstrate that any inferences made on the basis of the tests are valid for examinees of all groups.

Therefore, the purpose of this study is to investigate gender differences and the existence of DIF in the Std. 1 mathematics assessment in Trinidad and Tobago. The need to investigate DIF and gender differences in mathematics achievement was motivated by an ongoing quest to explain gender differences in mathematics achievement in the

Caribbean. We felt that it was important to look at gender differences through the lens of test quality, where test quality was measured by the number of DIF items. Of interest is whether items function the same for boys and girls of comparable ability, and the prevalence of the gender-related DIF. Specifically, this study attempts to answer the following questions:

1. Do male examinees perform significantly different from female examinees on a Standard 1 large-scale mathematics assessment in Trinidad and Tobago?
2. How prevalent is gender-related DIF on a Standard 1 large-scale mathematics assessment in Trinidad and Tobago?

Method

Data

The data used in this study came from the Standard 1 national mathematics test of 2004. The test consisted of 20 items, which are a mixture of objective and structured-type items. Each of the items fell into one of the following categories: *Number*, *Measurement and Money*, *Geometry*, and *Statistics*. Each item tested either knowledge computation (KC), algorithmic thinking (AT), or problem solving (PS). Some items had multiple parts, with each part testing a different skill. For example, item 10 had three parts and it measured all three skills. Overall, students responded to 30 questions. Table 1 shows the structure of the examination, and the number of items under each category and what skill each item measured.

Most items in the examination were dichotomously scored as either 1 for a correct response or 0 for an incorrect response. Only items 14-KC and 17-AT were polytomously scored. A correct response was scored as 1 or 2 (depending on whether a complete answer was provided or not) and an incorrect response as 0. To make responses of these two items suitable for analysis in this study, they were recoded as follows: Response $k = 0$ as 0; response $k = 1, 2$ as 1. Since this is a national examination, it was conducted under government regulations by which all Standard 1 students the country take the examination at the same time and date.

Table 1. Examination Questions (Items) Under Each Category

Number	Measurement and Money	Geometry	Statistics
1-KC	8-KC	13-KC	17-AT
2-KC	9-KC	14-KC	18-KC
3-AT	10-KC	15-AT	18-PS
4-KC	10-AT	16-KC	19-KC
4-AT	10-PS		20-KC
5-KC	11-KC		20-PS
5-AT	11-PS		
6-KC	12-AT		
6-AT	12-PS		
7-KC			
7-PS			

Sample

The original data contained 16,210 examinees, of whom 1,143 were absent on the day of the examination, or their papers were not scored, or gender was not recorded. These were excluded and this resulted in 15,067 (male = 7,720; female = 7,347) valid cases being considered for this study. DIF requires that groups be comparable with respect to the attribute the items supposedly measure. Because the National Test is curriculum based, we expect all students to be similarly familiar with the math content and, as a result, differences in performance on the exam should be random. Therefore, the total test score was used as a measure of ability, and was the criterion on which boys and girls were matched.

For this study, logistic regression was chosen over the MH and SIB Test procedures because of its superior ability to detect both uniform (the item favours one group over another at every ability level) and non-uniform or crossing DIF (there is an Ability * Group Membership interaction). With this approach, the logistic regression model is used for predicting the probabilities of correct and incorrect responses to each dichotomously scored item, given an observed total test score and its associated group membership.

Analysis

To investigate gender differences, descriptive statistics were obtained. The descriptive statistics for the test items and the four categories of the examination for male and female examinees are shown in Tables 1 and 2. The independent t-test was performed to find whether the differences are statistically significant on each of the four categories of the test and on the entire test. Since the t-test is sample size dependent and could yield significant results even when the differences are small, we calculated the effect sizes (*d*) to determine the practical significance or meaningfulness of the findings. The interpretation of d is as follows: Trivial: $d < .20$; small: $d = .20$; medium: $d = .50$; large: $d = .80$. To investigate DIF, logistic regression analysis was performed.

Table 2. Descriptive Statistics of the Test Sections for Male and Female Examinees

	Male Examinees *(n = 7,720)*		**Female Examinees** *(n = 7,347)*		**Total** *(n = 15,067)*	
Category	*Mean*	*SD*	*Mean*	*SD*	*t**	*d*
Number (11 items)	.64	.45	.69	.43	-6.98	.11
Measurement and Money (9 items)	.49	.44	.53	.44	-5.58	.09
Geometry (4items)	.60	.47	.68	.45	-10.68	.17
Statistics (6 items)	.45	.47	.53	.45	-10.68	.17
Entire Exam	.55	.46	.61	.44	-8.19	.13

* All t-values significant at $p < .001$

Results

Results are presented according to the research questions investigated in the study. The first research question was: Do male examinees perform significantly different from female examinees on a Standard 1 large-scale mathematics assessment in Trinidad and Tobago? To answer this question, descriptive statistics and an independent t-test were computed. The descriptive statistics are presented in Tables 2 and 3. To interpret the

results in Tables 2 and 3, the means represent the proportion of examinees that get an item correct. For example, Table 2 shows that for the Number category, the mean score for male students was .64 while the mean score for female students was .69 out of a possible score of 1.00. Alternatively, this can be interpreted as 64% of male students getting this category correct while 69% of female students got it correct. A similar interpretation can be made to Table 3. It is clear that the mean scores for female students are higher than the mean scores for male students for each category and item. To investigate whether these differences were statistically significant, an independent t-test was performed and the results are also presented in Table 2.

Table 3. Descriptive Statistics of the Test Items for Male and Female Examinees

	Male Examinees *(n = 7,720)*		**Female Examinees** *(n = 7,347)*		**Total** *(n = 15,067)*	
Item	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Number						
1-KC	.85	.36	.91	.28	.88	.32
2-KC	.44	.50	.47	.50	.46	.50
3-AT	.71	.45	.73	.44	.72	.45
4-KC	.64	.48	.68	.47	.66	.47
4-AT	.66	.47	.71	.45	.68	.47
5-KC	.34	.47	.39	.49	.36	.48
5-AT	.43	.50	.48	.50	.45	.50
6-KC	.75	.43	.82	.38	.79	.41
6-AT	.81	.39	.87	.33	.84	.36
7-KC	.68	.47	.72	.45	.70	.46
7-PS	.74	.44	.79	.41	.76	.43
Measurement and Money						
8-KC	.86	.34	.90	.30	.88	.33
9-KC	.90	.30	.93	.26	.91	.28
10-KC	.28	.45	.32	.47	.30	.46
10-AT	.51	.50	.55	.50	.53	.50

	Male Examinees *(n = 7,720)*		**Female Examinees** *(n = 7,347)*		**Total** *(n = 15,067)*	
Item	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
10-PS	.33	.47	.39	.49	.36	.48
11-KC	.21	.41	.24	.43	.22	.42
11-PS	.37	.48	.44	.50	.40	.49
12-AT	.50	.50	.52	.50	.51	.50
12-PS	.49	.50	.51	.50	.50	.50
Geometry						
13-KC	.59	.49	.67	.47	.63	.48
14-KC	.79	.41	.85	.36	.82	.39
15-AT	.59	.49	.69	.46	.64	.48
16-KC	.44	.50	.51	.50	.47	.50
Statistics						
17-AT	.77	.42	.87	.34	.82	.39
18-KC	.41	.49	.53	.50	.47	.50
18-PS	.42	.49	.53	.50	.47	.50
19-KC	.58	.49	.67	.47	.63	.48
20-KC	.27	.45	.29	.46	.28	.45
20-PS	.27	.45	.29	.46	.28	.45

The differences between male and female students were all significant for the test and for the four categories on the test. For the Number category, t (15065) = -6.98, p < .001; the Measurement category, t (15065) = -5.58, p < .001; Geometry category, t (15065) = -10.68, p < .001; and the Statistics, t (15065) = -10.68, p < .001. For the entire test, t (15065) = -8.19, p < .001. The effect sizes, *d*, were trivial indicating that significant differences were more due to the sample size than true differences in the population.

The second question was: How prevalent is gender-related DIF on a Standard 1 large-scale mathematics assessment in Trinidad and Tobago? To answer this question, logistic regression was employed. The analytic strategy in logistic regression is model comparison by successively adding the terms into the model (θ, $G, \theta G$). DIF evaluation is carried

out by statistically evaluating the incremental contribution of each successive model term (θ, G, θG). If the group effect G is statistically significant ($\tau_1 \neq 0$) and the effect of the interaction θG is not ($\tau_3 = 0$), then the item has uniform DIF. When $\tau_{2□} > 0$ DIF favours the reference group, and when $\tau_{2□} < 0$ DIF favours the focal group. If the interaction is statistically significant, that is, an item favours the higher-ability members of the reference group and the lower ability members of the focal group or vice versa, the item has non-uniform DIF. While non-uniform DIF does occur, this is not commonly found (Jodoin & Gierl, 2001). Therefore in this study, our focus is on uniform DIF.

The statistic used is the Wald test statistic, which follows a chi-squared (χ^2) distribution with degrees of freedom (*df)* = 1. The results in Table 4 show that item 1-KC displayed non-uniform DIF because the Total Score by Gender (θG) interaction term was statistically significant, χ^2 = 6.55, p < .011. For item 13-KC, the group effect (Gender) was statistically significant, χ^2 = 6.74, p < .009; but the interaction term was not statistically significant, χ^2 = 2.15, p < .142, which means that this item displayed uniform DIF. Similarly, items 15-AT, 20-KC, and 20-PS displayed uniform DIF because the interaction terms were not significant, but gender was.

Effect size measures

As previously stated, we compute and interpret ΔR^2 using the Jodoin and Gierl (2001) guidelines. The statistics summary of the effect sizes for the DIF test is the differences in Nagelkerke's R^2 between the model with absence of DIF and the full model; that is, the additional proportion of variability accounted for as each term is added to the model. These are provided in Table 5. Based on Jodoin and Gierl's criteria for using ΔR^2, a weighted least squares effect size measure to quantify the magnitude of uniform or non-uniform DIF, all these items are Type A items—negligible DIF, because $\Delta R^2 <$ .035. Item 1KC is the only item that approached moderate DIF. Thus, although these items displayed DIF, the magnitude was negligible as measured by the effect size.

Table 4. Items Displaying Uniform and Non-Uniform DIF

		Wald	*df*	*Sig.*
Item 1-KC				
	Total score	265.876	1	.000
	Gender	20.277	1	.000
	Total score by Gender	6.545	1	.011
	Constant	96.025	1	.000
Item 13-KC				
	Total score	341.381	1	.000
	Gender	6.742	1	.009
	Total score by Gender	2.151	1	.142
	Constant	261.961	1	.000
Item 15-AT				
	Total score	339.694	1	.000
	Gender	7.568	1	.006
	Total score by Gender	.640	1	.424
	Constant	279.711	1	.000
Item 20-KC				
	Total score	293.502	1	.000
	Gender	7.956	1	.005
	Total score by Gender	3.024	1	.082
	Constant	292.728	1	.000
Item 20-PS				
	Total score	301.692	1	.000
	Gender	6.048	1	.014
	Total score by Gender	2.037	1	.154
	Constant	302.711	1	.000

Table 5. Effect Size Measures for Differential Item Functioning Magnitude

Item	ΔR^2
Item 1-KC	.03
Item 13-KC	.01
Item 15-AT	.02
Item 20-KC	.02
Item 20- PS	.02

Figures 1 and 2 show two item characteristic curves for items that displayed DIF (item 1-KC) in Figure 1 and an item that did not display DIF (item 6-KC) in Figure 2. The fact that these two figures look similar even when one item has DIF and the other does not, supports results in Table 5 that the magnitude of the effect size is negligible.

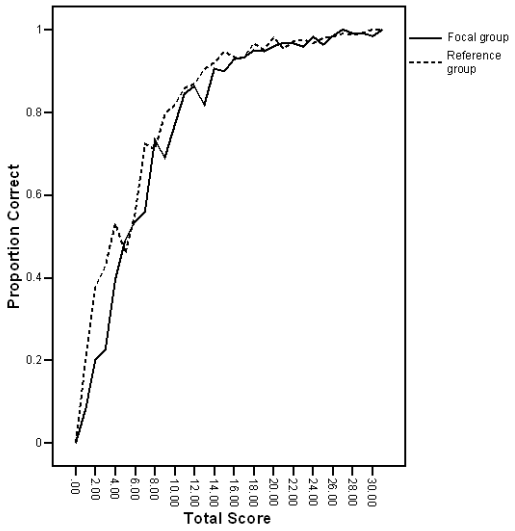

Figure 1. Example of item exhibiting differential item functioning.

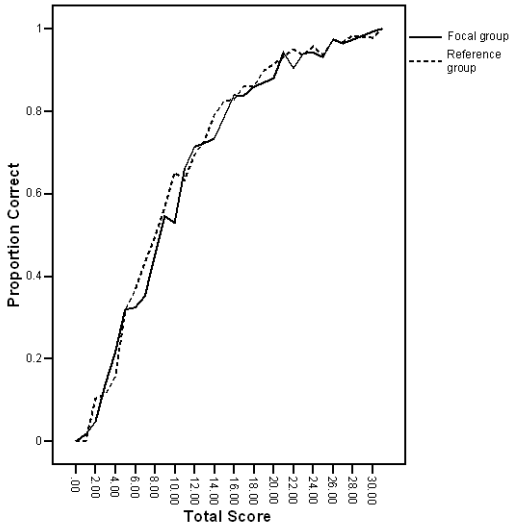

Figure 2. Example of item not exhibiting differential item functioning.

Discussion and Conclusion

Almost all researchers like their findings to be significant and of practical importance; however, sometimes an alternative finding, that is, one that is non-significant, or trivial in its effect, yields information that is just as important. It is tempting to say that this study discounts DIF as a possible source for the gender differences in performance on a test, because the magnitude of the male-female differences on particular items after controlling for group differences in mathematics ability was negligible. Yet, it is important to note Hidalgo and Lopez-Pina's (2004) observation that the statistic ΔR^2 appeared to be insensitive in classifying variables with moderate DIF and, therefore, one needed to be aware of the limitation of the interpretation criteria.

Additionally, it is necessary to recognize the probability of DIF amplification (Nandakumar, 1993). As Nandakumar explains, while items individually might display negligible DIF, collectively these items may have a significant effect on total test scores. Therefore, 5 of 30 items on a test exhibiting statistically significant DIF in favour of girls could be a cause for concern, because one group is being given an undue

advantage over another. It was also noted that even where items did not display significant DIF, on many items the item characteristic curve (ICC) showed the reference group (girls) to be above the focal group (boys).

Whether these results are interpreted to mean that DIF is not a problem, or emphasis is placed on possible DIF amplification, in the context of Trinidad and Tobago and the wider English-speaking Caribbean, this finding is important. We have not been able to locate any record of test items ever being examined for DIF. Therefore, this study adds a new dimension to the ongoing debate in the Caribbean about differences in performance between boys and girls. Best practice suggests that items that exhibit DIF should be identified before a test is published. When items show DIF, test developers must decide whether these items should be retained by determining if gender is a relevant dimension in the measurement of mathematics competency.

Ignoring the indication of DIF, the consistency of the mean differences in favour of girls seems to confirm findings of other research on mathematics achievement in the Caribbean that have shown female students doing better than male students, not only in mathematics but in other subjects as well. This evidence strongly suggests that while the examination board addresses the construction of the tests, the research on differences in the academic performance of male and female students also has to focus on sociocultural and socio-psychological factors as explanatory variables. In addition, the school has to become a focus of the research. Caribbean researchers will need to extend the work of Jules and Kutnick (1990) to better understand school and classroom dynamics that impact the sexes differently, and potentially limit one group or enhance another group's performance.

Walstad and Robson, (1997) posit that possible factors which may explain gender differences in achievement are socialization, differential reasoning, instructional practices, or the format of testing. Also, and as already noted (Jussim & Eccles, 1992; Leedy, Lalonde, & Runk,, 2003; Siegle & Reis, 1998; Tiedemann, 2000), teacher attitude and beliefs also more than likely reflect the prevailing attitude and beliefs in the wider society. While some of these factors are beyond the control of the school, others are not. It is expected that as schools are made aware of their role

in contributing to the disproportion in academic achievement between the sexes, they are more likely to initiate solutions to the problem. This is critical because despite the social and cultural factors that may explain the differential in performance, most issues involving education still have to be addressed in the school and classroom.

Although the gender differences on mathematics achievement were small, they are worth noting and need to be addressed from a policy standpoint. What was interesting in this study was that the female-male differences were consistent across the different sections of the test, with female students consistently outperforming male students in all content areas.

Our findings also indicate that, overall, students performed the worst on statistics, followed by measurement and money. Also, our findings indicate that, generally, the problem-solving skills for most students are not well-developed. These findings have important implications for classroom practice. They are important because they provide evidence of the need to develop effective strategies to teach these concepts. Teaching mathematics is not just about the focus on numbers; rather, it requires teachers to tailor their strategies to specific areas of mathematics. For example, teachers cannot teach statistics the same way they teach measurement or money.

Also, teachers need to be innovative in developing effective teaching methods. For example, the use of real-life examples is one very effective way of teaching statistics or money concepts. Students always want to see how concepts being taught in class relate to their everyday life situations. However, improving classroom practice is not the sole responsibility of the teachers, but of all stakeholders involved. Ultimately, teachers and school administrators are responsible for the implementation of the strategies in the classroom, but policymakers should be willing to provide funding and incentives for teachers to be creative in developing effective teaching strategies.

References

Bielinski, J., Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*(1), 51–77.

Blevins-Knabe, B., Austin, A., Musun, L., Eddy, A., & Jones, R. M. (2000). Family home care providers' and parents' beliefs and practices concerning mathematics with young children. *Early Child Development and Care, 165*, 41–58.

Bolt, D. M. (2000). A SIBTEST approach to testing DIF hypotheses using experimentally designed test items. *Journal of Educational Measurement, 37*(4), 307–328.

Boughton, K. A., Gierl, M. J., & Khaliq, S. N. (2000, May). *Differential bundle functioning on mathematics and science achievement tests: A small step toward understanding differential performance*. Paper presented at the Annual Meeting of the Canadian Society for Studies in Education (CSSE), Edmonton, Alberta, Canada.

Brown, L. (2005). Gender and academic achievement in math: An examination of the math performance data on seven to nine year olds in Trinidad and Tobago. *Caribbean Curriculum, 12*, 37–56.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Casey, M. B., Nuttall, R., Pezaris, E., & Benbow, C. P. (1995). The influence of spatial ability on gender differences in mathematics college entrance test scores across diverse samples. *Developmental Psychology, 31*(4), 697–705.

Clauser, B. E., & Mazor, K. M. (1998) Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice, 17*(1), 31–44.

DeMars**,** C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11(3),* 279–299.

Dodeen, H., & Johanson, A. G. (2003). An analysis of sex-related differential item functioning in attitude assessment. *Assessment and Evaluation in Higher Education, 28*(2), 129–134.

Doolittle, A .E., & Cleary, A. T. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*(2), 157–166.

Duffy, J., Gunther, G., & Walters, L. (1997). *Gender and mathematical problem solving. Sex Roles, 37*(7–8), 477–494.

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27*(5), 6–11.

Fennema, E., & Sherman, J. A. (1977). Sex-related differences in mathematics achievement, spatial visualization and sociocultural factors. *American Educational Research Journal, 14*(1), 51–71.

Forgasz, H. J., & Leder, G. C. (1996). Mathematics classrooms, gender and affect. *Mathematics Education Research Journal*, *8*(2), 153-173.

Friedman, L. (1989). Mathematics and gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research, 59*, 185–213.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items**.** *Journal of Educational Measurement*, *33(3)*, 315–332.

Gallagher, A. M., De Lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely M., & Cahalan C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, *75*(3), 165–190.

Garner, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education*, *12*(1), 29–51.

Gierl, M. J., Bisanz, J., Bisanz, G., & Boughton, K. (2003). Identifying content and cognitive skills that produce gender differences in mathematics: A demonstration of the multidimensionality-based DIF analysis paradigm. *Journal of Educational Measurement*, *40*(4), 281–306.

Halpern, D. F. (1997). Sex differences in intelligence: Implications for education. *American Psychologist*, *52*(10), 1091–1102.

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

Hanna, G. (2003). Reaching gender equity in mathematics education. *The Educational Forum*, *67*(3), 204–214.

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education*, *6*(2), 137–151.

Hidalgo, D. M., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903–915.

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

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type 1 error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349.

Johanson, G. A. (1997). Differential item functioning in attitude assessment. *Evaluation Practice*, *18*(2), 127–135.

Jules, V., & Kutnick, P. (1990). Determinants of academic success within classrooms in Trinidad and Tobago: Some personal and systemic variables. *Educational Studies*, *16*(3), 217–235.

Jussim, L., & Eccles, J. S. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, *63*(6), 947–961.

Kutnick, P. (2000). Girls, boys and school achievement: Critical comments on who achieves in schools and under what economic and social conditions achievement takes place – A Caribbean perspective. *International Journal of Educational Development, 20*(1), 65–84.

Kutnick, P., Jules, V., & Layne, A. (1997). *Gender and school achievement in the Caribbean* (Education Research Serial, No. 21). London: Department for International Development.

Leedy, G. M., Lalonde, D., & Runk, K. (2003). Gender equity in mathematics: Beliefs of students, parents, and teachers. *School Science and Mathematics*, *103(6)*, 285–292.

Maccoby, E. E., & Jacklin, C. N. (1974). T*he psychology of sex differences*. Stanford, CA: Stanford University Press.

Meyer, J. P., Huynh, H., & Seaman, M. A. (2004). Exact small-sample differential item functioning methods for polytomous items with illustration based on an attitude survey. *Journal of Educational Measurement, 41*(4), 331–344.

Meyer, M. R., & Koehler, M. S. (1990). Internal influences on gender differences in mathematics. In E. Fennema & G. Leder (Eds.), Mathematics and *gender* (pp. 60–95). New York: Teachers College Press.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy–Stout's Test for DIF. *Journal of Educational Measurement, 30*, 293–311.

Parry, O. (2000). *Male underachievement in high school education in Jamaica, Barbados, and St. Vincent and the Grenadines*. Kingston, Jamaica: Canoe Press.

Parsons, J. E., Adler, T. F., & Kaczala, C. M. (1982). Socialization of achievement attitudes and beliefs: Parental influences. *Child Development*, *53*(2), 310–321.

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*(1) 73–90.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. *Applied Measurement in Education*, *10*(4), 299–319.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performances: A meta-analysis. *Psychological Bulletin, 107*(2), 139–155.

Siegle, D., & Reis, S. M. (1998). Gender differences in teacher and student perceptions of gifted students' ability and effort. *Gifted Child Quarterly, 42*(1), 39–47.

Stipek, D. J., & Gralinski, J. H. (1991). Gender differences in children's achievement-related beliefs and emotional responses to success and failure in mathematics. *Journal of Educational Psychology, 83*(3), 361–371.

Swaminathan, H., & Rogers, J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

Tapia, M., & Marsh G. E. (2004). The relationship of math anxiety and gender. *Academic Exchange Quarterly*, *8*(2), 130–134.

Tiedemann, J. (2000). Parents' gender stereotypes and teachers' beliefs as predictors of children's concept of their mathematical ability in elementary school. *Journal of Educational Psychology, 92*(1), 144–151.

Tocci, C. M., & Engelhard, G. (1991). Achievement, parental support, and gender differences in attitudes toward mathematics. *Journal of Educational Research, 84*(5), 280–286.

Walstad, W. B., & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics, *Journal of Economic Education*, *28*(2), 155–171.

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education*, *9*(2), 175–199.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah. NJ: Erlbaum.

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

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Prince George, Canada: University of Northern British Columbia.