Caribbean Educational Research Journal
Vol. 3, No. 2, September 2015, 109-121

The University of the West Indies
Cave Hill Campus

Gender Bias in the Performance of Trinidad and Tobago Students on PISA 2009

Vivian Alexander* & Yukiko Maeda

Purdue University, Department of Educational Studies, Beering Hall of Liberal Arts & Education, West Lafayette, Indiana, USA

This study aimed to investigate the presence of gender bias in the Trinidad and Tobago sample of the Program for International Student Assessment (PISA) 2009 data (Organization for Economic Cooperation and Development [OECD], 2009a). We analysed the Mathematics, Science, and Reading Literacy items for gender differential item functioning (DIF). Our findings revealed three Mathematics items exhibiting DIF for Mathematics Literacy—two in favour of males and one in favour of females. We found no gender DIF in the Science, but we found one Reading Literacy item that showed DIF in favour of males. We also found that Trinidad and Tobago students did not respond to a substantial proportion of the Reading Literacy items. Further studies of gender differences in academic achievement in Trinidad and Tobago and the wider Caribbean should evaluate the extent to which assessment items show bias towards particular groups of students.

Keywords: Trinidad and Tobago, gender, Differential Item Functioning, PISA

Parents, students, teachers, governments and the general public need to know how well their education systems prepare students for real-life situations (Organization for Economic Cooperation and Development [OECD], 2010a). Educational assessments are one way through which stakeholders and experts measure the extent to which their education systems have met their curriculum goals(Wolff, 2004).Measuring students' learning is necessary to establish targets, assess the trade-offs of alternative resource allocation strategies, and reward good results (Wolff, 2004).Additionally, developing educational assessments require commitment, adequate financing, technical knowledge, managerial know how and political leadership (Wolff, 2004).

Many countries use educational assessments to evaluate the extent to which their educational systems are preparing students to meet the challenges of the workplace (OECD, 2010a).At the national level, assessments help measure the performance of individual students for the purposes of certification, school completion, determining access to the next level of education, and/or evaluating school/teachers accountability (Wolff, 2004). At the international level, assessments help extend and enrich the national picture by providing a larger context within which to interpret national performance (OECD, 2010a). Past decades have seen an increase in the number of international assessments administered globally such as the International Adult Literacy Survey (IALS), the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA) and the Progress in International Reading Literacy Study (PIRLS)(Brown, Micklewright, Schnepf, & Waldmann, 2007). The data generated from these studies are widely used by governments,

*Corresponding Author. Email: valexan@purdue.edu

ISSN 1727-5512

http://www.cavehill.uwi.edu/fhe/hum/publications/EducationCERJ.htm

international organizations, and researchers from various disciplines within and outside of education (Brown et al., 2007).

Unfortunately, the participation of Latin American and Caribbean (LAC) countries, particularly Caribbean countries in international tests of achievement has been both limited and poor. For example, in the Program for International Student Assessment (PISA) 2012 assessment of achievement of 15-year old students, the eight participating LAC countries all scored significantly below the average OECD score in Mathematics OECD, 2014). Similarly, in PISA 2006, the six participating LAC countries scored significantly below the average OECD score in Science (see OECD, 2007). Other international studies such as the Trends in International Mathematics and Science Study (TIMSS) corroborate the results found in PISA (Martin et al., 2012).Researchers attribute the limited participation of Caribbean countries in such assessments to factors such as funding, fear of poor performance, and limits to the institutional capacity needed to implement the test and analyze the data collected (Blome, 2013).

To date, Trinidad and Tobago has participated in three international standardized assessments of achievement: the International Association for the Evaluation of Educational Achievement (IEA) 1991 study of reading achievement of nine- and fourteen-year old students;the IEA's Progress in International Reading Literacy Study (PIRLS) 2006 of fourteen-year old students, andthe (OECD's), Program for International Student Assessment (PISA) 2009 study. Unfortunately, the country's performance on these assessmentswere were below the international benchmarks. For example, in the IEA (1991) study of reading, Trinidad and Tobago nine-year old students had a mean score of 451 (79) points which was well below the international benchmark of 500 points (Elley, 1992). In the same study, fourteen-year old students fared better in reading achievement with a mean reading score of 479 (87) points but still scored below the international benchmark. In the PIRLS 2006, another IEA study, students from Trinidad and Tobago again showed overall poor reading performance (Mullis, Martin, Kennedy, & Foy, 2007). The trend of underperformance of Trinidad and Tobago's students continued in PISA 2009 (OECD, 2010a) where they scored significantly below the international benchmarks for Reading, Mathematics, and Sciencerespectively.

Gender Achievement Gap in Trinidad and Tobago

In addition to academic underperformance, there were some notable gender differences in Trinidad and Tobago students' performance on these assessments. Trinidad and Tobago females had consistently higher reading achievement. For example, in the IEA 1991 study of reading achievement, females scored significantly higher than males at both ages nine and fourteen. Similarly, in the IEA's PIRLS 2006 study, fourth grade females scored significantly higher in reading achievement (Mullis et al., 2007). The trend continued in the PISA 2009 study, females again scored significantly higher than males in reading (OECD, 2010a,). The observed gender differences in reading achievement are consistent with those found from industrialized nations. Females also outperformed males in Mathematics and Science (OECD, 2010a) but the score point difference was smaller when compared to Reading.

Gender differences in academic achievement at the national level are also readily apparent. Reports suggest that Trinidad and Tobago females outscore males academically at all levels of the school system. For example, at the primary school level De Lisle, Smith, & Jules (2005) reported females scoring consistently higher in the Secondary Entrance Assessment (SEA) and the national assessment. Similarly, George, Quamina-Aiyejina, Cain, and Mohammed (2009) reported significantly lower mean scores for males in each subtest of the SEA. They also found greater variability in the score distributions for males on the areas assessed. At the secondary level, George et al. (2009) reported greater enrolment by females in all subject areas evaluated except Physics. They also reported higher achievement by females in all subject areas assessed except Mathematics. Together these studies show a trend of overall greater academic performance by females in Trinidad and Tobago.

Caribbean researchers have posited several explanations for the apparent greater academic achievement of females. Miller (1991) theorized that the marginalization of males, engendered by dramatic transformations within Caribbean society, was the cause of male underachievement (e.g., lower test scores). However, Barriteau (2003) argued against Miller's marginalization theory stating that it did not address the multifaceted nature of Caribbean masculinity and that researchers must give attention to factors such as changes in the political economy of the Caribbean, changes in the gender identities of Caribbean women, gender justice, and social inequities. Alternatively, Parry (2000) argued that anti-school masculinities, gender-roles socialization in-and-out of school, the existence of coeducational and same sex schools, the lack of male role models, and differences in the development of males' and females gender/sex identities all contribute to the observed differences in attainment (e.g., years schooling) and achievement (e.g., test scores) between males and females. Nonetheless, researchers acknowledge that not all boys identify with constructions of masculinity which place them at risk (Mills & Keddie, 2010) and that there are still issues concerning girls' education that need urgent attention(Warrington & Kiragu, 2012).

Researchers have also alluded to the low quality of Caribbean schooling as another possible explanation for the observed gender achievement patterns in the Caribbean (Bailey, 2000; Kutnick, 2000). All of these possible explanations for the pervasive gender gap in achievement have a common socio-economic thread that ties them together. However, the apparent lack of educational achievement and attainment in males is hardly a simple topic of debate. For example, Gosine (2007) found in his study that in spite of greater female course enrolment and educational attainment, there is still persistent gender subject stereotyping and males are over-represented in the Science, Technology, Engineering, and Mathematics (STEM) disciplines while females over-represented in the Humanities and Languages.

Fairness of Measuring Academic Performance

Despite the varied reasons for explaining gendered patterns of academic achievement in Trinidad and Tobago, researchers have given little attention to the nature of the test that may engender differing patterns of achievement between males and females. When a test itself or the manner in which it is used results in different meanings for scores earned by members of different identifiable subgroups, that test is said to be biased (American Educational Research Association (AERA), the American Psychological Association (APA), & the National Council on Measurement in Education (NCME), 1999, p. 74). The results obtained from a biased test are not valid and should not be used in the educational decision-making process. Within the Trinidad and Tobago context, bias in evaluation instruments may possibly explain the persistent trend of higher female academic achievement. As mentioned earlier, researchers have theorized varied reasons for the observed gendered patterns of academic achievement and attainment. Moreover, for much of the theories posited the researchers utilized either qualitative or aggregate quantitative data with little attention given to item-level data analyses. Item level analyses will allow test developers to design and implement targeted interventions to improve students' academic achievement.

Ensuring that an assessment is free of bias is necessary because stakeholders make important educational decisions concerning students' educational careers that can affect their life outcomes based on educational assessment data. Therefore, it is important that the assessments used to measure students' are unbiased. Given these points concerning the importance of measuring test fairness, there is need to evaluate the extent to which items on a test may biased towards a particular group of students.

Purpose of Study

Previous work by Brown and Kanyongo (2010) examined gender differential item functioning (DIF) in the Mathematics portion of the National Test which is administered annually to Standard 1 (primary school) students in Trinidad and Tobago. The authors found that 17% of the items on the Mathematics portion of the test showed gender-related DIF, but had negligible effect sizes. However, the authors failed to take into account the nestedness of data that may have serious implications on the estimates obtained. Nested (hierarchical) data structures are prevalent in many social and behavioral studies (Raudenbush & Bryk, 2002). The basic premise of nested data structures is that the social group or context, to which persons belong, influences persons in the group and the characteristics of the persons who comprise the group influence the group as a whole (Maas & Hox, 2005). Therefore, the analysis of nested data structures using conventional statistical procedures such as ANOVA and regression will result in biased estimates and can degrade DIF detection (French & Finch, 2010; Raudenbush & Bryk, 2002).Additionally, the Brown and Kanyongo (2010) study utilized primary school data as the focal point of the study. However, it is equally important to examine test bias at the secondary level as well since students' performance at the secondary level also has important implications for their life outcomes. Finally, there seems to have been very few studies evaluating DIF while accounting for the nestedness of the data.

For the purposes of this study, we focus only on the statistical hypothesis test results since previous studies show that analyzing multilevel (nested data) using standard statistical techniques has the greatest influence on the standard errors of the estimates (Raudenbush & Bryk, 2002). We would like to make it clear to readers that our focus on the hypothesis tests does not imply that we consider effect sizes as an unimportant tool in DIF detection. Indeed, effect sizes provide valuable information concerning the magnitude of the differences between groups. However, given the detrimental impact of ignoring multilevel data on standard errors (used in hypothesis testing), this should be the first aspect of DIF detection to be investigated with respect to multilevel data (French & Finch, 2010). Additionally, there is little, if any, agreement upon criteria or guidelines for interpreting a DIF effect size and its meaningful magnitude making it difficult to interpret any results obtained. This problem is particularly acute because prior research on DIF effect sizes using logistic regression have not examined the impact of correctly modelled multilevel structure or improper models that ignore such structure (French & Finch, 2010). Moreover, in DIF analyses, the statistical test for DIF occurs before or at least in combination with effect sizes. Therefore, for these reasons we focus only on detecting DIF rather than the effect sizes.

Research Questions

Given the points highlighted above this study investigates the following research questions:

1. To what extent do the items on PISA 2009 show differing patterns of achievement for males and females?
2. To what extent do the items on PISA show differential item functioning (DIF)?

Methods

Data and participants

For the current investigation, we derive the item response data on the Reading, Mathematics, and Science literacy items from the OECD's PISA 2009 study (www.pisa.oecd.org). PISA assesses the extent to which fifteen-year old students have acquired the knowledge and skills essential for their full participation in modern societies, and focuses on Reading, Mathematics, and Science literacy (OECD, 2010a).PISA does not seek to assess if students can reproduce knowledge, but rather how well they can extrapolate from what they have learned and apply it in unfamiliar settings, both in and outside of school (OECD, 2010a).

PISA describes *Mathematics literacy* as an individual's capacity to understand the role mathematics in the world, to make well-founded judgments, and to use mathematics to meet the needs of the individual's life as a constructive, concerned, and reflective citizen (OECD, 2010b, p.14). PISA's 2009 assessment used 32 dichotomous and three graded response items to measure Mathematics literacy. *Science literacy* describes an individual's use of scientific knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues (OECD, 2010b, p.14).PISA used 53 dichotomous items with no graded response items to measure science literacy. *Reading literacy* is an individual's capacity to understand, use, reflect on, and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society (OECD, 2010b, p.14). PISA used 131 reading items: 31 items had completely missing data on all items i.e., none of the students responded to the items—these items were not used in the analyses, 8 items were graded response with three categories (0, 1, 2), and the 92 remaining items were dichotomously coded. For the current investigation, we only focused on dichotomous items to explore the trend. The Trinidad and Tobago sample of students who participated in 2009 data collection consists of 4,778 students (age M = 15.68, SD = 0.29) from 158 schools. Of the 4,778 students, 2,495 (52.0%) were female and 2,283 (48.0%) were male. Data obtained these 4,778 cases were used for the current investigation.

Analysis

Two methodological decisions guided our data analysis for this study. First, PISA utilised a two-stage cluster sampling design to gather data for the Trinidad and Tobago sample. In the first stage of sampling, schools were selected from a school sampling frame using probability proportional to size (PPS) sampling in which larger schools had a greater probability of being selected. In the second stage of sampling, a random sample of eligible fifteen-year old students were selected from each of the schools chosen in the first stage of sampling. Thus, the sampling design used by the OECD for the study naturally gave rise to a nested data structure and thus a multilevel approach was required to account for the nestedness of the data.

The second major methodological decision guiding our analysis was the use of the IRT framework instead of the classical test theory (CTT) framework for evaluating test bias. Item response theory is a general statistical theory about students' performance on an item and a test and how their performance relates to the abilities that are measured by the items in the test. Researchers use IRT models to analyze items and scales, create and administer psychological measures, and measure individuals on a psychological construct (Reise et al., 2005). In the IRT framework, a latent trait (usually ability) is a construct underlying and directly influencing a person's responses to items on a scale designed to measure that trait (Reise, et al., 2005). Overall, the basic goal of IRT modelling is to evaluate item quality and to derive other important psychometric properties of the test or scale (Reise, et al., 2005).

We chose the IRT method over the classical method for several reasons. First, PISA generates student scores using IRT methods. Therefore, we saw it prudent to work within the IRT framework as a matter of consistency. Second, IRT models are considered strong models because the statistical assumptions of the model (unidimensionality and invariance) are more difficult to meet as compared to CTT models. Third, IRT links students' item responses to their ability, and places the item and the ability parameter on the same scale, which is not the case in CTT. Fourth, the item and person parameters generated in IRT are sample independent and this gives IRT more flexibility in the use of the person and item statistics for practical test development and makes analyses less complicated.

Differential item functioning (DIF) occurs when, accounting for the proficiency measured by the test, the probability of a correct response for one group of interest is not equivalent to the probability of a correct response for a comparison group (DeMars, 2011). Differential item functioning in a test may occur for a number of reasons. The wording of the items on the test may be confusing to some respondents and may introduce a level of difficulty not associated with the item, or the presentation of the test content may prevent respondents from demonstrating their true knowledge of a topic and may introduce a source difficulty (Brown & Kanyongo, 2010). The items may be in a language that is not native to the respondents or the test may be measuring more than one ability.

IRT provides a theoretically useful tool for detecting DIF by estimating item parameters and ability.IRT assumes unidimensionality i.e., that the test measures only one latent trait (ability). Therefore, no other trait should influence the item response, and if DIF is present the item may be measuring more than one trait (i.e. the test is multidimensional). IRT modelling assumes local independence, which means that a person's responses for any pair of items are independent conditioned on ability. This assumption allows one to study DIF item-by-item. Third, IRT modelling assumes that the item parameters do not vary across sample, i.e., sample characteristics do not have an influence on the item parameters. Therefore, for our study we combined multilevel modelling with IRT in order to account for the nestedness of the data while examining the items for potential DIF.

We analyzed the data using a 2-parameter multilevel IRT model using the flexMIRT statistical software (Cai, 2013). The 2-parameter IRT model assesses the item discrimination and the item difficulty. Within the IRT framework, the discrimination parameter denoted by a represents the extent to which the items distinguishes between students of high and low ability levels (Hambelton & Jones, 1993). Positive values indicate higher discrimination while negative values indicate lower discrimination. Discrimination values greater than 1 are good, and values above 1.75 are excellent. The difficulty parameter denoted by b represents the ability level at which students will have a 50% chance of obtaining the correct response (Hambelton & Jones, 1993). A positive value indicates that a student needs to be at a higher ability level in order to have a 50% chance of getting the item correct the i.e., the item is difficult. A negative value indicates that students need to be of a lower ability level to have a 50% chance of getting the item correct i.e., the item is easy (Hambelton & Jones, 1993). A b-parameter less than -2 is for the bottom 5% of the students (very easy) in the sample, while a *b*-parameter greater than 2 is for the 5% of students in the sample (very hard).

Results

Table 1provides the results of the variance component analysis. The results of the analysis show that schools account for more than 50% of the variation is students' responses to the literacy items in all three content areas. This means that school context is important for explaining the variations in students' responses on the test.

Table 1

Variance Components for the IRT Analyses for Mathematics, Science, and Reading

Fixed Effect	Between School Mean (SE)	Within School Mean (SE)
Mathematics	-0.59 (0.03)	0.02 (0.01)
Science	-0.34 (0.03)	-0.01 (0.01)
Reading	-0.11 (0.03)	0.00 (0.01)
Random Effect	Between School Variance (SE)	ICC
Mathematics	1.51 (0.06)	0.60
Science	1.24 (0.05)	0.55
Reading	1.26 (0.05)	0.56

Note. Within school variance fixed to 1 for model identification purposes

Table 2 shows the results of the IRT analysis for Mathematics literacy items. The table shows that 74% of the between school items and 63% of the within school items had good or excellent discrimination values. Similarly, the vast major of items (95%) had good or excellent discrimination for both males and females. The majority of items were difficult for students between schools (21% easy, 79% difficult) and within schools (16% easy, 84% difficult), with 21% and 11% of the items being very difficult for students between and within schools respectively. Additionally, the same proportion of items was easy and difficult for females and males (11% easy, 89% difficult), with 58% and 47% of the items proving to be very difficult for males and females respectively ($b > 2$). We found that Item 5, Wald's χ^2 (2) =7.70, p = .021 and Item 11, Wald's χ^2 (2) = 8.20, p = .016 had DIF in favor of males. This means that for males and females matched on the ability trait, males had a greater probability of getting items 5 and 11 right. On the contrary, Item 17 Wald's χ^2 (2) = 15.10, p <.001 had DIF in favor of females.

Table 2

The 2-parameter IRT Model for Mathematics Literacy Items

	Discrimination (a)		Difficulty (b)		Discrimination (a)		Difficulty (b)		DIF
Item	Between School	Within School	Between School	Within School	Female	Male	Female	Male	χ^2
Item 1	1.47	1.14	-0.64	-0.50	1.90	2.06	-0.02	-0.03	0.80
Item 2	1.51	1.36	-1.88	1.70	2.54	3.29	1.39	1.24	3.60
Item 3	1.80	2.14	2.10	2.50	3.67	4.48	1.65	1.49	3.40
Item 4	1.60	1.20	0.74	0.56	2.62	2.21	0.74	0.71	2.30
Item 5	1.55	1.28	1.46	1.21	2.28	3.03	1.21	1.00	7.70^{m}
Item 6	1.27	0.80	-0.39	-0.24	1.59	1.62	0.11	0.18	0.80
Item 7	2.20	2.11	2.87	2.75	3.57	5.08	1.91	1.84	2.20
Item 8	0.67	0.68	-0.51	-0.52	1.09	0.92	-0.02	-0.11	2.20
Item 9	1.17	1.40	0.09	0.11	1.59	1.91	0.36	0.39	3.00
Item 10	1.67	1.16	1.35	0.94	2.35	2.83	1.00	0.96	1.90
Item 11	0.95	0.90	0.62	0.59	1.25	1.79	0.73	0.64	8.20^{m}
Item 12	1.78	1.37	1.69	1.30	2.92	2.91	1.14	1.20	0.80
Item 13	0.98	0.89	1.29	1.17	1.57	1.45	0.99	1.16	2.20
Item 14	0.78	0.68	0.25	0.22	1.34	1.07	0.37	0.53	4.40
Item 15	1.10	0.95	1.88	1.63	2.08	1.62	1.29	1.39	3.00
Item 16	0.70	0.66	0.56	0.53	0.94	1.14	0.63	0.67	2.60
Item 17	1.57	1.02	3.01	1.96	3.54	1.74	1.53	1.83	15.10^{f}
Item 18	1.10	1.72	0.45	0.70	2.14	1.86	0.67	0.63	2.00
Item 19	2.28	1.26	2.77	1.53	3.07	3.26	1.48	1.40	0.70

Note. χ^2 with superscript *m* indicates DIF for males, χ^2 with superscript *f* indicates DIF for females

Table 3 shows the results of the IRT analysis for the Science literacy items. The table shows that 65% of the between-school items and 50% of the within-school items had good or excellent discrimination values. The majority of items had good or excellent discrimination for females (96%) and males (88%). The majority of items were difficult for students between schools (31% easy, 69% difficult) and within schools (35% easy, 65% difficult), with 8% of items between schools and 12% of the items within schools being very difficult for students. There was one very easy item for students within schools. Additionally, the items were mostly difficult for both females (27% easy, 73% difficult) and males (23% easy, 77% difficult) and none of the items were very easy or difficult for students. We found no gender DIF in the science literacy items.

Table 3

The 2-parameter IRT Model for Science Literacy Items

	Discrimination (a)		Difficulty (b)		Discrimination (a)		Difficulty(b)		DIF
Item	Between School	Within School	Between School	Within School	Female	Male	Female	Male	χ^2
Item 1	1.16	1.02	1.09	1.24	1.79	1.52	0.88	1.05	1.90
Item 2	1.11	0.95	2.00	-2.34	1.15	1.53	1.75	1.55	1.90
Item 3	1.47	1.21	0.59	0.72	2.07	1.98	0.57	0.64	0.30
Item 4	1.20	1.08	0.25	0.28	1.49	1.92	0.47	0.31	4.20
Item 5	0.68	0.72	-0.15	-0.14	1.09	0.99	0.08	0.06	0.30
Item 6	0.83	0.78	1.33	1.41	1.17	1.21	1.05	1.15	0.80
Item 7	0.65	0.80	1.68	1.36	1.14	0.94	1.06	1.37	1.70
Item 8	1.36	1.19	1.30	1.49	1.98	1.87	1.22	1.05	3.50
Item 9	0.83	0.90	-0.86	-0.79	1.26	1.29	-0.29	-0.43	1.20
Item 10	0.83	0.81	0.54	0.56	1.11	1.31	0.64	0.52	1.30
Item 11	1.11	0.90	0.59	0.72	1.52	1.58	0.66	0.59	0.30
Item 12	1.23	1.22	-0.25	-0.25	1.93	1.88	-0.03	0.07	0.80
Item 13	1.55	1.48	1.05	1.09	2.72	1.98	0.87	0.93	2.50
Item 14	1.15	1.37	-0.50	-0.42	1.76	1.95	-0.17	-0.07	0.60
Item 15	1.37	0.87	1.36	2.14	0.97	0.93	0.61	0.86	2.50
Item 16	1.67	1.76	-0.84	-0.80	2.73	2.43	-0.35	-0.37	0.50
Item 17	0.87	0.94	1.59	1.47	1.35	1.31	1.33	1.14	2.40
Item 18	1.86	1.57	0.47	0.56	2.39	2.78	0.53	0.55	0.60
Item 19	1.45	1.18	0.85	1.04	2.16	1.80	0.74	0.86	1.50
Item 20	1.58	1.10	-1.20	-1.73	2.03	2.13	-0.78	-0.68	0.70
Item 21	1.33	0.90	1.71	2.53	1.56	1.66	1.68	1.53	0.80
Item 22	0.98	0.89	0.72	0.80	1.28	1.53	0.73	0.70	1.10
Item 23	1.01	1.27	-1.71	-1.36	1.94	1.36	-0.81	-0.94	4.60
Item 24	0.64	0.69	-1.25	-1.16	1.07	0.88	-0.59	-0.70	1.10
Item 25	0.95	1.25	0.83	0.63	1.62	1.51	0.69	0.66	0.50
Item 26	1.05	0.76	2.11	2.92	1.40	1.30	1.80	1.85	0.10

Note. χ^2 with superscript *m* indicates DIF for males, χ^2 with superscript *f* indicates DIF for females

Table 4 shows the results of the IRT analysis for the Reading literacy items. The table shows that 78% of the between-school items and 61% of the within-school items had good or excellent discrimination values. For females, all of the items had either good or excellent discrimination values, while for males only one item had a discrimination value less than 1. Unlike Mathematics and Science, the Reading literacy items were mostly easy for students between schools (61% easy, 39% difficult) and within schools (61% easy, 39% difficult), with approximately 10% of the items being very easy and 7% of the items were very hard. Additionally, the items were mostly easy for both females (78% easy, 22% difficult) and males (76% easy, 24% difficult), with none of the reading items begin very easy or difficult. We found gender DIF for Item 39 in favor of males Wald's χ^2 (2) = 9.70, p = .008.This means that even when matched on ability, males were more likely to get the item right.

Table 4

The 2-parameter IRT Model for Reading Literacy Items

	Discrimination (a)		Difficulty (b)		Discrimination (a)		Difficulty (b)		DIF
Item	Between School	Within School	Between School	Within School	Female	Male	Female	Male	χ^2
Item 1	1.19	1.21	-1.18	-1.17	2.17	1.76	-0.74	-0.81	1.70
Item 2	1.30	1.08	0.89	1.07	1.71	1.97	0.57	0.64	2.1
Item 3	1.33	1.29	0.41	0.42	2.27	1.83	0.23	0.26	1.3
Item 4	1.60	1.44	0.02	0.02	2.29	2.26	-0.03	-0.01	0.00
Item 5	1.07	1.01	-1.94	-2.06	1.49	1.65	-1.52	-1.23	3.80
Item 6	1.35	1.46	-1.66	-1.53	1.78	1.77	-0.03	-0.18	1.80
Item 7	1.33	1.53	-1.77	-1.54	1.48	1.69	-0.74	-0.70	0.60
Item 8	1.62	1.54	-1.90	-1.99	1.66	1.40	1.12	1.08	2.20
Item 9	1.00	1.03	-0.81	-0.79	1.26	1.28	-0.55	-0.60	0.10
Item 10	0.81	0.87	1.73	1.61	1.84	2.06	0.64	0.63	0.70
Item 11	1.08	1.10	-0.61	-0.60	1.87	2.51	-0.25	-0.32	1.70
Item 12	1.42	1.18	1.42	1.71	2.19	2.14	-1.17	-1.00	2.70
Item 13	0.97	0.79	1.02	1.25	2.26	2.22	-1.20	-1.02	2.90
Item 14	0.89	0.82	-0.84	-0.91	2.17	2.93	-1.43	-1.17	3.60
Item 15	0.88	0.95	-0.99	-0.92	1.49	1.55	-0.47	-0.43	0.20
Item 16	1.19	0.96	-0.22	-0.27	1.25	1.25	-1.75	-1.51	2.50
Item 17	1.11	0.87	-0.68	-0.86	2.05	1.43	0.94	1.19	2.00
Item 18	1.45	1.09	0.70	0.93	1.37	1.27	0.63	0.77	0.40
Item 19	1.21	1.55	-1.84	-1.44	1.35	1.31	-0.62	0.60	0.00
Item 20	0.78	0.74	-1.76	-1.85	1.02	1.22	-0.80	-0.91	1.40
Item 21	1.03	0.89	2.43	2.81	1.76	1.55	-0.69	-0.68	4.60
Item 22	1.04	1.38	-2.74	-2.07	1.76	2.03	-0.62	-0.65	0.40
Item 23	0.85	1.22	-3.06	-2.13	1.63	1.54	-0.52	-0.54	0.10
Item 24	1.03	1.16	-0.97	-0.86	2.80	2.49	-0.38	-0.39	0.30
Item 25	1.14	1.07	-2.14	-2.28	1.83	2.07	0.12	0.05	0.40
Item 26	0.78	0.76	-0.94	-0.96	1.83	1.86	-0.77	-0.64	1.30
Item 27	1.17	0.93	-1.00	-1.26	1.92	1.81	-0.36	-0.34	0.10
Item 28	1.03	0.98	-0.07	-0.07	1.91	2.25	0.47	0.44	1.00
Item 29	1.63	1.16	-0.15	-0.21	2.23	2.22	-1.18	-0.98	3.60
Item 30	1.14	1.11	-0.38	-0.39	1.50	1.77	-1.66	-1.43	1.20
Item 31	0.81	0.72	0.00	0.00	1.33	1.33	-0.35	-035	0.00
Item 32	1.55	1.24	-0.94	-1.18	1.36	1.26	1.66	1.95	1.70
Item 33	1.15	1.11	-2.28	-2.36	1.84	1.97	-1.59	-1.54	0.10
Item 34	1.33	1.02	1.22	1.59	1.75	1.39	-1.69	-1.72	2.10
Item 35	1.33	0.97	0.25	0.34	1.40	1.67	-0.76	-0.60	1.90
Item 36	1.45	1.03	-0.19	-0.26	2.78	1.68	-1.21	-1.55	2.80
Item 37	1.37	0.98	0.72	1.01	1.16	1.11	-0.57	-0.84	3.40
Item 38	1.13	0.81	2.34	3.26	1.54	1.36	-0.81	-0.91	0.70
Item 39	1.06	0.82	2.78	3.60	2.22	1.59	-0.10	-0.33	9.70^m
Item 40	0.74	0.66	0.15	0.17	1.57	1.59	-0.22	-0.37	1.90
Item 41	1.39	1.00	1.35	1.88	1.15	0.93	-0.05	-0.03	1.10

Note. χ^2 with superscript *m* indicates DIF for males, χ^2 with superscript *f* indicates DIF for females

Overall, we found significant variation in students' responses to the items across schools. We also found the mathematics and science items to be mostly difficult for Trinidad and Tobago students. In contrast, the PISA reading literacy items were relatively easy for Trinidad and Tobago students. We found DIF in three of the mathematics items (two in favor of males and one in favor of females), and DIF in one reading literacy item in favor of males.

Discussion

In this study, we sought investigate to what extent do the items on the PISA 2009 assessment show gender differential item functioning (DIF). The results of our study indicate that the PISA items generally functioned similarly for male and female students except for Mathematics literacy and Reading literacy. We found DIF on one item in favor of males in Reading, which was surprising given the dominance of females on Reading. We found DIF in three of the Mathematics items: one in favor of females and two in favor of males. This means that the items on the scales are not free of bias and may affect the scaling of scores used to derive the plausible values for the analyses. This also means that researchers should use caution when making generalizations about students' performance based on tests/assessments not evaluated for DIF (or item bias). Additionally, the results from our study provide some evidence that when investigating gender differences in achievement and/or attainment researchers need to account for the nature of the test along with the other factors posited by researchers.

Also, it is very important for researchers undertaking large-scale quantitative studies using education data to control for the nestedness of the data, in order to prevent the mis-estimation of parameter estimates. This is vitally important because interpreting information or scores using biased estimates can have serious consequences since policy makers, teachers, administrators and parents use such information when making key educational decisions. Corollary to this, the government and educational administrators must invest in training for teachers concerning test construction. This is particularly important in the classroom setting where teachers may unintentionally develop biased tests that can result, for example, in student retention in the same grade level for an additional year, or placed in remedial classes unnecessarily. Now, while we were able to demonstrate DIF, we were unable to identify the specific domains to which the items exhibiting DIF belonged to due to data restrictions implemented by PISA. As a result, we were unable to go into detail about the nature of the items that exhibited DIF. Without such restrictions, we would have been able to identify precisely to which domains the items that exhibited DIF belonged. This information would have been helpful to researchers, policy makers, and school administrators in identifying academic areas of weaknesses and strengths of students for developing targeted interventions.

Another point of interest is the large proportion of Reading items that Trinidad and Tobago students neglected to answer. Again, due to the restrictions on access to the data we were unable to determine the nature of the items that the students neglected to answer. Another major point to note is that in PISA, the Reading, Mathematics, and Science literacy items were designed to be unidimensional (OECD, 2010b). One question that arises is: are the items actually part of a multi-dimensional construct? An implication of this unidimensionality assumption is that if a second dimension expressed by a small number of items is present, subsequent analyses will likely remove these items resulting in the loss of potentially important information (Blum, Goldstein, & Guérin-Pace, 2001; Goldstein, 2004).

Overall, this study shows the importance of evaluating the extent to which items on assessments function similarly or differently for students in different subgroups. Evaluating bias in test instruments is a necessary step in ensuring test-fairness. This is especially important in the Trinidad and Tobago context where high-stakes testing is the norm. The need for such evaluations holds for not only international assessments but national assessment as well. Such evaluations will help stakeholders develop assessments that are free of bias.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Bailey, B. (2000). School failure and success: A gender analysis of the 1997 General Proficiency Caribbean Examinations Council (CXC) examinations for Jamaica. *Journal of Education and Development in the Caribbean, 4*(1), 1-18.

Barriteau, E. (2003). "Requiem for the Male Marginalization Theory in the Caribbean. Death of a Non theory," in *Confronting Power Theorizing Gender: Interdisciplinary Perspectives in the Caribbean*. ed. Eudine Barriteau. Kingston: University of the West Indies Press, 2003.

Bloem, S. (2013), "PISA in Low and Middle Income Countries", OECD Education Working Papers, No. 93, OECD Publishing. http://dx.doi.org/10.1787/5k41tm2gx2vd-en.

Blum, A., Goldstein, H., & Guérin-Pace, F. (2001). International Adult Literacy Survey (IALS): An analysis of international comparisons of adult literacy. *Assessment in Education: Principles, Policy & Practice*, *8*(2), 225-246. doi:10.1080/09695940123977

Brown, L. I., & Kanyongo, G. Y. (2010). Gender Differences in Mathematics Performance in Trinidad and Tobago: Examining Affective Factors. *International Electronic Journal of Mathematics Education, 5*(3), 113-130.

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2007). International Surveys of Educational Achievement: How Robust Are the Findings? *Journal of the Royal Statistical Society. Series A (Statistics in Society), 170*(3), 623-646. doi: 10.2307/4623193

Cai, L. (2013). flexMIRT version 2: *Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

De Lisle, J., Smith, P., & Jules, V. (2005). Which males or females are most at risk and on what? An analysis of gender differentials within the primary school system of Trinidad and Tobago. *Educational Studies, 31*(4), 393-418.

Demars, C. E. (2011). An Analytic Comparison of Effect Sizes for Differential Item Functioning. *Applied Measurement in Education, 24*(3), 189-209. doi: 10.1080/08957347.2011.580255

Elley, W. B. (1992). *How in the world do students read?IEA study of reading literacy*. The Hague: The International Association for the Evaluation of Educational Achievement.

French, B. F., & Finch, W. H. (2010). Hierarchical Logistic Regression: Accounting for Multilevel Data in DIF Detection. *Journal of Educational Measurement, 47*(3), 299-317. doi: 10.1111/j.1745-3984.2010.00115.x

George, J., Quamina-Aiyejina, L., Cain, M., & Mohammed, C. (2009). *Gender Issues in Education and Intervention Strategies to Increase Participation of Boys.* (LO2138/OC-TT). Ministry of Education Retrieved from http://www.moe.gov.tt/national_consultation_primaryschool/GEORGE%20J%202009%20Gender%20Issues%20in%20Education%20and%20Intervention%20Strategies%20to%20Increase%20participation%20of%20Boys.pdf.

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: issues and practice, 12*(3), 38-47.

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. doi: 10.1016/s0738-0593(99)00045-0.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86-92.

Miller, E. (1991). *Men at Risk*. Kingston, Jamaica: Jamaican Publishing House, Kingston, Jamaica.

Mills, M., & Keddie, A. (2010). Gender justice and education: constructions of boys within discourses of resentment, neo-liberalism and security. *Educational Review, 62*(4), 407-420.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary School in 40 Countries.*

Organization for Economic Cooperation and Development (OECD). (2010a). *PISA 2009 Results: What Students Know and Can Do-Student Performance in Reading, Mathematics and Science (Volume I)*. Paris: OECD.

Organization for Economic Cooperation and Development (OECD). (2010b). *PISA 2009 Technical Report*. Paris: OECD.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2 ed.): Sage Publications.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item Response Theory: Fundamentals, Applications, and Promise in Psychological Research. *Current Directions in Psychological Science, 14*(2), 95-101.

Warrington, M., & Kiragu, S. (2012). “It makes more sense to educate a boy”: Girls ‘against the odds’ in Kajiado, Kenya. *International Journal of Educational Development, 32*(2), 301-309.

Wolff, L. (2004). Educational Assessments in Latin America: The State of the Art. [Article]. *Applied Psychology: An International Review, 53*(2), 192-214. doi: 10.1111/j.1464-0597.2004.00168.x