Gender-Related Differential Item Functioning of Mathematics Computation Items among Non-native Speakers of English

S. Kanageswari Suppiah Shanmugam

Gender-Related Differential Item Functioning of Mathematics Computation Items among Non-native Speakers of English

S. Kanageswari Suppiah Shanmugam[1]
Universiti Utara Malaysia

Abstract: *This study aimed at determining the presence of gender Differential Item Functioning (DIF) for mathematics computation items among non-native speakers of English, and thus examining the relationship between gender DIF and characteristics of mathematics computation items. The research design is a comparative study, where the boys form the reference group and the girls form the focal group. The software WINSTEPS, which is based on the Rasch model was used. DIF analyses were conducted by using the Mantel-Haenszel chi-square method with boys forming the reference group and girls forming the focal group. A total of 988 boys and 1381 girls in form two were selected from 34 schools, with 17 schools located in the Penang island, 12 schools in Penang mainland and five schools in Perak. Some 20 items were selected from the grade eight TIMSS 1999 and TIMSS 2003 released mathematics items. Findings revealed that seven items were flagged as DIF, where two were of moderate DIF and one as large DIF. Two DIF items assessed combined operation from the topics of fraction and negative numbers in the Number domain and the cognitive domain of lower-order thinking skills of Knowing favoured girls. One moderate DIF which assessed higher order thinking skills of Applying from the Algebra domain favoured boys. The findings trigger a possibility that computation items with one step operation, which assess lower-order thinking skills favour girls, while items that assess higher-order thinking skills favour boys.*

Keywords: gender Differential Item Functioning, computation items, Mantel-Haenszel chi-square method

Introduction

Differential Item Functioning (DIF) is the result of producing different probability of answering an item correctly among examinees of equal proficiencies (Roussos & Stout, 2000), caused by differences unrelated to test proficiency. The versatility of DIF analysis to flag items that may function differently for different groups of students with the same ability enables it to be a

---

[1] kanageswari@uum.edu.my

commonly used method to analyse differences in performance by gender (Maranon, Garcia, & Costas, 1997).

In examining the issue of gender differences in mathematics achievement, the stereotyped belief is that boys are better than girls in mathematics (Davis, 2008). However, recent trends tend to debunk gender stereotypes as international studies such as Trends in International Mathematics and Science Study (TIMSS) reveal mixed results of favouring either or neither gender. In the most recent TIMSS 2015, there was little difference by gender for mathematics. Of the 39 countries from TIMSS 2015 Grade Eight, a higher number of 25 countries indicated no significant gender difference (Saudi, United Arab Emirates, Egypt, South Africa, Kuwait, Qatar, Turkey, Kazakhstan, Iran Islamic Rep. of, England, Malta, New Zealand, Japan, Morocco, Georgia, Korea Rep. of, Norway, United States, Australia, Israel, Slovenia, Lebanon, Lithuania, Ireland, Hong Kong SAR). Chinese Taipei was the only country that did not record any difference in the mathematics score between boys and girls. Seven countries recorded better performance among girls (Bahrain, Botswana, Jordan, Malaysia, Oman, Singapore, Thailand), while a slightly smaller number of six countries recorded higher performance among boys (Canada, Chile, Italy, Sweden, Hungary, Russian Federation) (Mullis, Martin, Foy, & Hooper, 2016).

In comparing the two most recent cycles of TIMSS 2011 and TIMSS 2015, data for 25 countries from 34 countries with comparable data revealed no change in the gender gaps. A total of 16 countries recorded no gender difference for both cycles and seven countries recorded better performance among girls than a slightly decreased number of two countries for boys (Mullis at al., 2016).

A similar analysis for the cycles since 1995 reveals that from 16 countries, boys obtained higher scores than girls in four countries with an average advantage score of 17 points, while 12 countries did not record any gender differences in the mathematics score. In 2015, the boys still performed better than girls but in three countries, with an average advantage score of 9. However, girls performed better than boys in Singapore, with an average advantage score of 10 points (Mullis et al., 2016).

In the local context, the mathematics results for Malaysia in TIMSS for the three consecutive cycles in 2007, 2011 and 2015 suggest otherwise. The average scale scores in TIMSS 2007 for Malaysian girls (479) was significantly higher than boys (468), as well as for TIMSS 2011 that recorded a significantly better performance for girls (449) when compared to boys (430) (Mullis, Martin, Foy, & Arora, 2012). Similarly in the TIMSS 2015, girls showed significantly better performance (470) than boys (461) (Mullis at al., 2016). In national assessments such as Primary School Achievement Test or Ujian Pencapaian Sekolah Rendah (UPSR), Secondary Examinations of Lower Secondary Assessment or Penilaian Menengah Rendah (PMR) and Malaysian Certificate of Education or Sijil Pelajaran Malaysia (SPM) and until the tertiary level, girls show superiority in mathematics (Malaysian Ministry of Education, 2013).

With the emerging trend of girls generally obtaining higher mathematics scores than boys in Malaysia, amidst the international results that have mixed results, it is interesting to examine the characteristics of mathematics computation items that function differently for boys and girls. Computation items are focussed on as they have relatively less 'language load' and with language removed as an extraneous variance, DIF analysis allows the exploration of other item characteristics, except linguistics features marked by language load. Within the context of this

study, computation items are defined as items that involve algorithm that most commonly involves manipulation of numbers and variables (Neidorf, Binkley, Gattirs & Nohara, 2006).

Accordingly, the main purpose of this study is to determine whether gender DIF exists in mathematics achievement, and thus, examine the relationship between gender DIF and characteristics of the mathematics computation items. By flagging DIF items it is possible to identify the computation mathematics items that function differently across gender groups, as an attempt to examine the characteristics of computation items by gender.

Statement of Problem

Gender differential performance in mathematics is a cause for alarm, especially lately with the uprising issue of women's underrepresentation in Science, Technology, Engineering and Mathematics (Hyde, Lindberg, Linn, Ellis, & Williams, 2008). However in studies that examine gender differential performance in mathematics, it is rather challenging to determine whether the statistical differences in the mathematics achievement between boys and girls is due to their true differences in mathematical ability or test-related factors such as item-type. Therefore, this study will address the gap in examining this issue of gender difference as a result of item-type and identify characteristics of mathematics computation items that cause the differential performance by gender. By detecting the characteristics of the mathematics computation DIF items, the issue of apparent widening gender gap will be explained from a new perspective of item characteristics.

Test validity 'refers to the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), p. 11), while fairness is the 'fundamental validity issue and requires attention

throughout all stages of test development and use" (AERA, APA, & NCME, 2014, p. 49). As stated in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), one of the empirical evidence to substantiate test validity is to conduct DIF analyses. The presence of DIF items violates test validity as they influences the prediction of getting the item correct by students of different subgroups (AERA, APA & NCME, 2-14). Accordingly, since developing tests that measure the intended construct minimises the harmful effects of the tests being affected by construct-irrelevant variance, DIF analyses also enhances fairness in testing.

Studies on DIF items suggest that complex multiple choice items (Liu & Wilson, 2009), real-life embedded items (Lane, Wang, & Magone, 1996) and computation items (Berberoglu, 1995) with higher skills that involve a combinations of at least three mathematical operations (Salubayba, 2014) favor boys. From the perspective of the content domain, certain content domains such as Arithmetic or Algebra (Hyde, Fennema & Lamon, 1990), numbers and operations, geometry, and data analysis and probability (Lindberg, Hyde, Petersen, & Linn, 2010) indicate no gender differences.

However, items involving figures (Lane, Wang & Magone, 1996) and from the topic Space and Shape (Liu & Wilson, 2009) favor boys. Simple items that require either single or two basic operations to solve do not favor boys if the contexts embodying the items are unfamiliar , such as cooking. On the other hand, word problem items (Berberoglu, 1995) requiring conceptual knowledge (Lane, Wang & Magone, 1996) favor girls. With this new dimension added by item-type, this study will bridge the gap between the characteristics of mathematics computation item in English with minimum language load and gender DIF in the context of non-native speakers of English that exist in countries such as Malaysia.

According to Pedrajita (2009), gender biased test items that contain materials that (dis)favor gender groups can be detected through DIF analysis, and the source of DIF can be further examined to improve the test. In addition, with the recent rising trend in Malaysia of girls achieving better in mathematics than boys, the findings will contribute to the growing body of knowledge on DIF items that aptly explains the Malaysian context. This is achieved by addressing the literature gap in examining the characteristics of mathematics computation items that behave differently for boys and girls.

Research Objective

The purpose of this study is to determine whether gender DIF exist in mathematics computation items and to examine the characteristics of the mathematics computations DIF items. Therefore, the research questions are

a) Does gender DIF exist for students' achievement in mathematics computation items?

b) What are the characteristics of mathematics DIF computation items?

Literature Review

Differential Item Functioning (DIF)

Measurement equivalence or measurement invariance is a statistical attribute of measurement that indicates that the same construct of the test is being measured across all subgroups of the student population (Desa, 2014). The lack of measurement invariance, indicated by the presence of DIF items threatens test validity. DIF results when items function differently when students with equal ability for the construct under measure provide different responses as a result of belonging to different sub-groups. When items composing a test behave differently for the reference group compared to the focal group, even after controlling for student proficiency, DIF is said to have occurred (Dodeen & Johanson, 2003). An item is tagged as non-

DIF when students with equal ability for the construct under measure have equal probability of getting that item correct, regardless of their sub-group (Holland & Thayer, 1988). When an item is flagged as DIF, it could be due to true differences in the students' ability or test inherent characteristics such as the linguistics characteristics or test content. The former results in impact while the latter results in item bias (Dorans & Holland, 1993). This means that when subgroups of students who have been matched to their ability produce different probabilities of getting a correct response to an item, then the item is tagged to be biased to the group that it disfavors. The former results in item bias while the latter results in impact, which reflects the differences in the overall ability distributions between the two groups (Dorans & Holland, 1993). The concern for test developers and educators is test bias.

Test bias can occur due to the (un)familiarity of the test content to particular subgroups of students, construct irrelevant variances that augment unnecessary difficulty to the test and flawed items (Pedrajita, 2009). Accordingly, Hyde, Fennema, and Lamon (1990) discovered that computation items favor girls, especially items that have no equations (Mendes-Barnett & Ercikan, 2006), unlike Berberoglu (1995), whose findings indicated that computation items favour boys while word problem items favour girls. Again, in opposing to Berberoglu's (1995) study, Lindberg et al., (2010) discovered that items involving complex problem solving strategies and measurement favour boys. They explained that items that assess higher cognitive levels in particular, items from the content domain of geometry will favour boys, as was discovered earlier by Geary (1996), who also highlighted the domain visualization as favouring boys. A likely explanation was provided by Engelhard (1990), who discovered that girls tend to do better on easy items, while boys tend to do better on more difficult items, probably because girls prefer items that involve memorization (Becker, 1990).

Straying from this pattern of explanation, a study conducted by Garner and Engelhard (1999) found that multiple choice items were found to be favouring the boys and not girls, while constructed response items favoured the girls (Garner & Engelhard, 1999). Their study also revealed another interesting finding from the perspective of mathematics content domain, similar to another study conducted years later by Mendes-Barnett and Ercikan (2006), which is there was no gender difference in the geometry content domain.

The content domains of Geometry and Algebra are interesting as there are mixed findings from studies conducted from 1990s to present. While Mendes-Barnett and Ercikan (2006) discovered that algebraic items favour boys, a later study by Lindberg et al. (2010) found them to favour girls. Interestingly, the study conducted Garner and Engelhard (1999) were even more uplifting as they revealed that girls not only favoured algebra but more abstract mathematics items, even though they did not favour geometry, measurement, and data analyses. Boys on the other hand, preferred the topics ratios, proportions, and percentage (Garner & Engelhard, 1999). They also favoured items involving real-world setting and unrehearsed algorithms (Harris & Carton, 1993), which explains the findings by Mendes-Barnett and Ercikan (2006), who pinpointed boys' preference for items on higher-order thinking skills (Frost, Hyde, & Fennema, 1994).

Gender differences were also noted in the strategy used to solve mathematics questions as discovered by Carr and Davis (2001), where the sampled elementary boys preferred abstract strategies, unlike girls who preferred concrete strategies. Similarly, the different approaches adopted by boys and girls to learn mathematics have also been used to explain gender Dif in mathematics (Gallagher (1992; Garner & Engelhard, 1999). Boys were found to be favouring items that require non-routine strategies that are challenge the rehearsed algorithms practised in

class, while girls preferred the standard strategies taught in the classroom (Gallagher, 1992). This

finding probably supports boys' preference for HOTs items.

In the Malaysian context, real world problem items, items from the domains of geometry

(Abedalaziz, 2010a) and items that assess spatial and deductive abilities (Abedalaziz, 2010b)

favor boys. Items that assess the content domain of Algebra (Abedalaziz, 2010a) and that assess

the lower order thinking of numerical ability (Abedalaziz, 2010b) favor girls.

Therefore, with multiple perspectives formed in addressing the sources of DIF, this study

attempts to isolate the computations items and examine their characteristics of the DIF items

among non-native speakers of English. The emphasis is placed on examining characteristics of

mathematics computations items that have comparatively less language load, which have been

flagged as DIF.

Theoretical Framework

Item Response Theory (IRT)

IRT describes a relationship between the probability of answering an item correctly to the

person's ability and originates from a family of mathematical models that predict examinees'

performance based on their ability (persons' ability) denoted by $\square$ and item characteristics such

as item difficulty denoted by the b parameter and is represented by the position of the Item

Characteristic Curve (ICC), item discrimination denoted by the a parameter and is represented by

the slope of the ICC and pseudo-guessing which is denoted by the c parameter and is represented

by the lower asymptote of the ICC (Stone & Zhang, 2003). The mathematical function of IRT

varies in accordance to the number of parameters used. The one parameter model (1-PL) is also

known as the Rasch Model with b parameter. Two parameter model (2-PL) has two parameters,

the a and b parameter; while the three parameter model (3-PL) has three parameters: the a

parameter, the b parameter and the c parameter. The mathematical expressions for each

parameter model are as exhibited (Yen & Fitzpatrick, 2006):

$$1\text{-PL} : P_i (\theta) = P_i : (X_i = 1/\theta) = \frac{1}{1 + \exp[-(\theta - b_i)]}$$

$$2\text{-PL} : P_i : (X_i = 1/\theta) = \frac{1}{1 + \exp[-Da_i(\theta - b_i)]}$$

$$3\text{-PL} : P_i (\theta) = c_i + \frac{1}{1 + \exp[-Da_i(\theta - b_i)]} \quad \text{where,}$$

$P_i (\theta)$ = the probability of a student answering item *i* correctly at ability θ,

b_i = item difficulty parameter,

a_i = item discrimination parameter,

D = 1.7 (scale factor)

IRT positions all the test items on a common scale alongside the examinees and allows

the measurement of any subset items to the person's ability on the latent trait. Cohen, Bottge, and

Wells (2001) clarified that a person's ability refers to the amount of latent trait and the test scores

represent the amount of latent trait that the examinees have. The latent trait is assessed by the

items composing the test. This is because the examinees' observed responses to the test items

indicate their position in a scale of unobservable latent trait, which the test content assesses

(Ellis, Becker, & Kimmel, 1993).

Ideally, in constructing the measurement model, the data need to fit the Rasch model,

which is challenged in practice. This is because the data will deviate from the model and to

examine the extent of the admissible departure, mean-square indices (infit mean-square and

outfit mean-square values) are used (Wright & Linacre, 1994). Infit mean-square is affected by

students' pattern of item responses. Outfit mean-square is affected by responses to very difficult

or very easy items. The mean-square indicates the size of randomness, which explains the

amount of distortion in the measurement system. Values less than one indicate too predictive

observations resulting in the data overfitting the model while values more than one indicate unpredictability or data underfitting the model (Linacre, 1994). Generally for MCQ, the acceptable values of infit mean-square and outfit mean-square should be in the range of 0.8 to 1.2 but for high stakes MCQ the range of 0.7 to 1.3 is used. If the infit-outfit mean-square values exceed 2.0, then the measurement may be degraded. If it is within the range of 1.5 to 2.0, it indicates that the items are unproductive for measurement but do not degrade it. Values in the range of 0.5 to 1.5 indicate that items are productive for measurement while values less than 0.5 indicate that items are not productive but do not degrade the measurement. Items with high mean-square values are recommended for removal only while developing new tests and not for pre-existing tests (Wright & Linacre, 1994).

Mantel-Haenszel Chi-square Method

Item Response Theory 1-Parameter model, or the Rasch model is a parametric model that forms the basis for the software Winsteps. There are two methods for evaluating DIF in WINSTEPs, which are the Mantel-Haenszel chi-square and the Welch t-test. Furthermore, the Mantel-Haenszel Chi-square method using WINSTEPS is not the same as the traditional method of computing Mantel-Haenszel Chi-square statistics. One reason is due to the conversion of test scores to an interval scale using WINSTEPS. Accordingly, it is imperative to make a distinction between

(a)   the WINSTEPS Mantel-Haenszel Chi-square from the traditional method of

computing Mantel-Haenszel Chi-square statistics, and

(b)   the Mantel-Haenszel Chi-square method from the Welch t-test in WINSTEPS.

In addressing (a), as Linacre (2017) clarified the Mantel-Haenszel Chi-square method using WINSTEPS does not use the test scores to match the reference and focal groups by ability. WINSTEPS converts students' raw test scores into person measure before stratifying the data and therefore, interval scores are obtained, which fulfils the assumption of a parametric test. The transformation of the data set into interval scale makes WINSTEPS Mantel-Haenszel Chi-square method (Winsteps M-H) different from the conventional Mantel-Haenszel computation (M-H computation). Linacre (2017) highlights their differences as

The usual M-H computation stratifies the sample by raw scores, so it works with case-wise deletion of cases with missing data. Winsteps stratifies cases by measure, so cases with missing data are stratified at their estimated measure. For complete data and thin-slicing, the conventional M-H computation and the Winsteps M-H computation produce the same numbers. With missing date or thick-slicing, the conventional M-H computations and the Winsteps M-H computations may differ (p.607).

As for (b), Welch's two-sided t-test is an indication of the statistical difference between the average difficulties of the two understudied sets of items (Linacre, 2011). The Welch t-test tests the null hypothesis of the DIF size is zero and rejects the null hypothesis of the obtained t-statistic as a part of the t-distribution if $p<.05$ (Linacre, 2011). Linacre (2012) explains that in theory the results obtained by using the Mantel-Haenszel Chi-square method and the Welch t-test in WINSTEPS to detect DIF items should be the same. However, as Linacre (2017) pinpointed the Mantel-Haenszel Chi-square method is highly preferred in comparison to the t-test in WINSTEPS since it is more accurate due to its robustness to missing data. Therefore, the data set is more complete, which explains the reason behind Educational Testing service, an established

and highly reputable test developer adopting Mantel-Haenszel Chi-square method in their DIF analysis (Linacre, 2012). In his words,

M-H and the t-tests in Winsteps should produce the same results, because they are based on the same logit-linear theory. But, in practice, M-H will be more accurate if the data are complete and there are large numbers of subjects at every score level, so called "thin" matching. Under other circumstances, M-H may not be estimable, or must use grouped-score "thick" matching, in which case the t-test method will probably be more accurate. (p. 607)

Using the Welsh t test, DIF items are identified when $p < 0.05$ and similarly, items are flagged as DIF when the Mantel-Haenszel probability value is at the most 0.05 and classified as displaying negligible, moderate or large DIF based on the criteria for the DIF size (Zwick, Thayer & Lewis, 1999).

C = moderate to large |DIF| $\geq 1.5 / 2.35 = 0.64$

B = slight to moderate |DIF| $\geq 1 / 2.35 = 0.43$

A = negligible |DIF| $\leq 1 / 2.35 = 0.43$

Positive Mantel-Haenszel size favors the focal group while negative Mantel-Haenszel size favors the reference group (Linacre, 2008b). In this study, girls formed the focal group while boys formed the reference group.

Methodology

This study is a comparative research study, where the boys form the reference group and the girls form the focal group since they form the interest of this study. A total of 34 schools were selected with 17 schools located in the Penang Island and another 12 schools in Penang mainland while only five schools in the Perak state were selected due to distance and mainly time constraints. In each school, six Form Two classes were selected and within each class, all

the students were selected regardless of their race, religion, gender, language background, academic achievement or language proficiency. Six classes were chosen to cover the high, intermediate and low abilities students for an even distribution of students. However, in schools with less than six classes, all the classes were selected. In this study, 12 schools had less than six classes. A total of 988 boys and 1381 girls answered the items and all of them are non-native speakers of English. For DIF analyses, the sample size required for each of the focal and reference groups need to exceed 100 (Fidalgo, Ferreres, & Muniz, 2004).

State		
Penang		Perak
Island	Mainland	
17	12	5

Instrument

The 20 mathematics test items for this study were selected from the TIMSS Grade 8 released items for the two cycles since 1999. Items from TIMSS 2007 and TIMSS 2011 were not used as they were the two immediate cycles to avoid the practice-effect. These released items were then mapped to the learning outcomes described in the Integrated Curriculum for Secondary Schools Curriculum Specifications Mathematics Form 1 (2003) and Integrated Curriculum for Secondary Schools Curriculum Specifications Mathematics Form 2 (2003) to establish content validity of the test. The items were then classified as computation items according to the definitions of Neidorf et al. (2006). The rationale for selecting only computation items is to conduct preliminary DIF analysis on items that have little 'language load' so that a secondary dimension (language) is not introduced. The mathematics test had two sections. Section A contains demographic information such as gender, class, name of the school and

section B had the 20 computation items written in English. The items were neatly arranged and each page had two items on it so that there was ample space for students' work.

Data Collection

The 20 computation mathematics items were administered to students with the help of the class teachers, in accordance to the routine practices of an examination. The class teacher distributed the test booklets and gave the students 5 minutes to fill in the particulars required in section A. They were given one hour to answer the test items. Calculators were not allowed based on the following considerations; the test objective is to assess student's mathematical proficiency and not their skills in using calculators, there are items that can be answered by the use of only calculator and as such, contradict the test objectives. As the students answered the test, the class teacher invigilated and the researcher monitored to ensure no malpractice occurred. At the end of one hour, the test papers were collected.

Data Analyses

The test booklets were scored dichotomously and DIF analysis was conducted to flag DIF items. The DIF items are flagged by using the Mantel-Haenszel chi-square method based on IRT 1-PL and the Welch t-test statistics. There are two methods for evaluating DIF in WINSTEPs, which are the Mantel-Haenszel Chi-square and the Welch t-test. Even though, this study is focussed on employing the former as DIF classification is derived from the well-established criteria developed from the Educational Testing Service (Longford, Holland & Thayer, 1993), as a comparison, the Welch t-test was also conducted. Both methods are available using WINSTEPS (Linacre, 2008a).

Results

Measurement Model

The data was first analyzed to examine the extent to which the data fit the 1-Parameter model, which is also known as Rasch model by analyzing the infit and outfit indices. Table 1 exhibits the infit and outfit indices for the computation items.

Table 1

Summary of 20 Measured (non-Extreme) Computation Items

	measure	model error	infit mnsq	infit zstd	outfit mnsq	outfit zstd
M	0.00	.06	0.98	-0.4	1.03	0.8
SD	1.14	.01	0.11	4.0	0.28	4.2
max.	2.06	.09	1.17	6.5	1.68	9.9
min.	-2.55	.05	0.76	-9.8	0.65	-5.6

real root-mean-square-error .05 adj.sd .77 separation 4.99   item reliability 1.00

model root-mean-square-error .05 adj.sd .96 separation 18.68   item reliability 1.00

As displayed, the average infit mean square is 0.98 and the outfit Mean Square (MNSQ) value is 1.03. These values are within the acceptable range of 0.5 to 1.5, and are acceptable for a good measurement. Thus, they fit the model. The MNSQ values for both infit and outfit are also within the range of 0.8 and 1.2, which is the recommended range for high stakes multiple-choice test (Wright & Linacre, 1994). In addition, the item reliability index is a perfect 1.0.

Distribution of Item Difficulty and Student Ability

Figure 1 illustrates the person to item map. From Figure 1, it can be deduced that all the 20 computation items are well distributed within students' ability. The most difficult item is Item 15 from the content domain of algebra that assesses the evaluating of expressions by substituting

numbers for letter. This is followed by Item 20 from the content domain of Geometry on determining the centre of rotation. Examining the clusters of items in Figure 1, it can concluded that the difficulty decreases as less able students tend to answer items from linear equations, decimal number, Cartesan plane (stating coordinates), lines and angles, fraction, ratio and combined operations for directed numbers. The easiest items tend to be assessing one step operations such as subtraction (Items 1, 7 and 4).
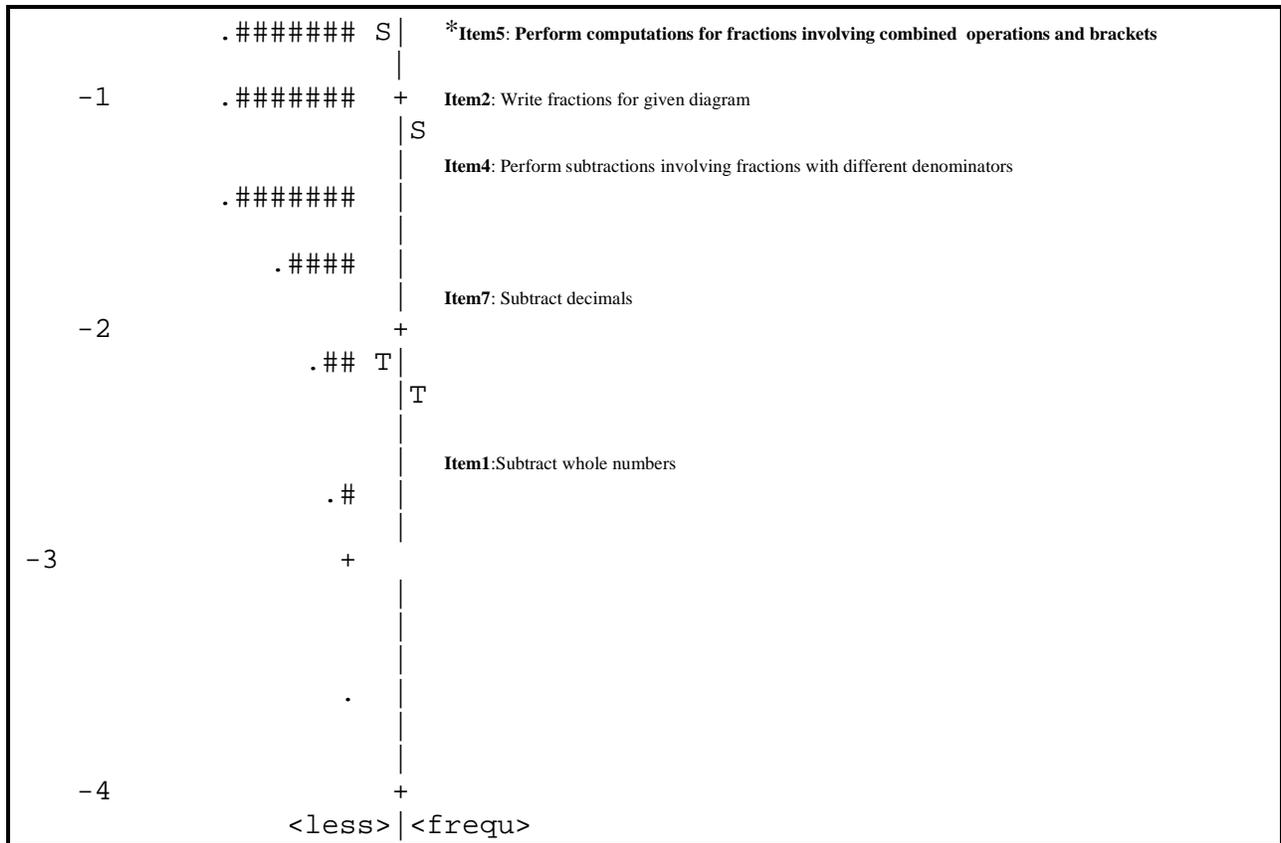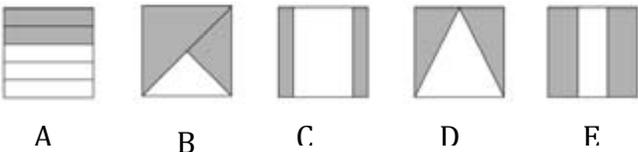
```
PERSONS - MAP - ITEMS
   4       #######   +
                      |
                    T |
                      |
       .##########    |
                      |
                      |
   3                  +
                      |
                      |
     ############     |
                      |
                    S | T
       .##########    |
   2                  +   **Item 15: Evaluating expressions by substituting numbers for letter
                      |
       .##########    |     Item20: Determine the centre of rotation
                      |
        .##########   |
                      |
       .##########    | S
   1                  +   Item17: Write linear equations in one unknown for given statement
        .#########    |   Item10:Solve problems    Item14: Simplify algebraic expressions by      Item16: Solve linear equations
                      |         quadrilaterals              involving collecting like terms
                    M |   Item6: Arrange decimals in order
     .###########     |
                      |   Item8: Divide a decimal by a decimal
       .##########    |   Item19:State coordinates of points on Cartesian plane
                      |   Item9:Solve problems involving angles formed by        Item 3:Find equivalent fraction for a given fraction
                      |        intersecting lines
   0   .#########   + M
                      |
         .######      |   Item18: Find the value of a quantity given ratio
         .#######     |   Item13: Identify in terms of two or more unknowns as the product of the unknowns with a number
                      |   *Item11:Perform computations for directed numbers     Item12: Estimate squares of numbers
                      |       involving  combined operation and brackets
```

```
              .#######  S|      *Item5: Perform computations for fractions involving combined  operations and brackets
                         |
   -1         .#######   +       Item2: Write fractions for given diagram
                         |S
                         |       Item4: Perform subtractions involving fractions with different denominators
             .#######    |
                         |
               .####     |
                         |       Item7: Subtract decimals
   -2                    +
               .##  T|
                         |T
                         |
                         |       Item1:Subtract whole numbers
                 .#      |
                         |
   -3                    +
                         |
                         |
                         |
                 .       |
                         |
                         |
   -4                    +
                  <less>|<frequ>
```

Each '#' is 14. each '.' is 10   *- favoring girls   **favoring boys

Figure 1. Person-Item map

Using the software WINSTEPS, DIF analyses was conducted. Both Welch t-test statistics and Mantel-Haenszel Chi-square probability recorded values of less than 0.05 for the same seven items flagged as displaying DIF. Table 2 exhibits the seven DIF items with item specifications related to their content and cognitive domains. From Table 2, seven items had the Mantel-Haenszel probability value of less than 0.05 and therefore, were flagged as DIF items. The items were from either from the content domains of Number or Algebra and from the cognitive dimensions of Knowing and Applying.

Table 2

Mathematics Computation Items Flagged as DIF

Item Number	Item	Content Domain
Item 2	Which shows $\frac{2}{3}$ of the square shaded? A B C D E	Number Knowing
Item 4	What is the value of $\frac{4}{5} - \frac{1}{3} - \frac{1}{15}$? A $\frac{1}{5}$ B $\frac{2}{5}$ C $\frac{7}{15}$ D $\frac{3}{4}$ E $\frac{4}{5}$	Number Knowing
Item 5	What is the value of $\frac{3}{5} + (\frac{3}{10} \times \frac{4}{15})$? A $\frac{3}{51}$ B $\frac{1}{6}$ C $\frac{6}{25}$ D $\frac{11}{25}$ E $\frac{17}{25}$	**Number Knowing**
Item 6	Which of these is the smallest number? A 0.625 B 0.25 C 0.375 D 0.5 E 0.125	Number Knowing
Item 11	What is the value of $1 - 5 \times (-2)$? A 11 B 8 C −8 D −9	**Number Knowing**
Item 15	If $\frac{a}{b} = 70$, then $\frac{a}{2b} =$ A 35 B 68 C 72 D 140	**Algebra Applying**
Item 16	If $x - y = 5$ and $\frac{x}{2} = 3$, what is the value of y? A 6 B 1 C −1 D −7	Number Applying

Based on the values of the DIF measure for the items, a DIF measure plot was plotted as displayed in Figure 2. DIF measure plot reports the item difficulty for students by gender classification. Items with bigger values for the DIF measure indicate higher difficulty for the group involved (Linacre, 2008b).
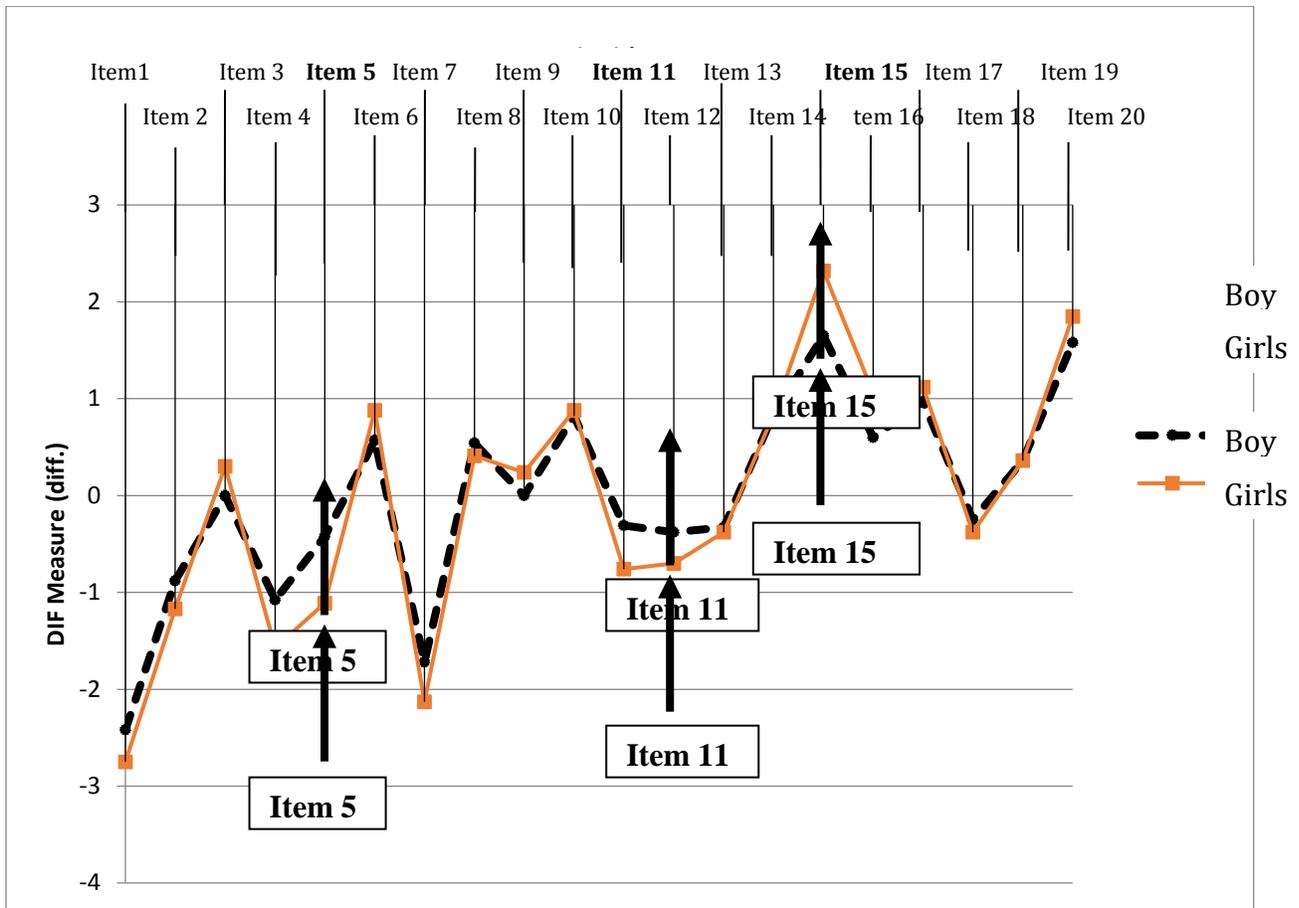
Figure 2. DIF measure plot for the mathematics computation items.

As illustrated in Figure 2, Items 5 and 11 are more difficult for boys as indicated by the bigger values of the DIF measure when compared to the girls. Only Item 15 is more difficult for the girls. Table 3 exhibits the statistical values that were obtained for the DIF classification as negligible, moderate or large based on the DIF size by Zwick, Thayer and Lewis (1999).

Table 3

DIF Analysis for the Mathematics Computation Items

Item	Welch *t*-test	Mantel-Haenszel prob	Mantel-Haenszel Size	DIF	Favours	Content/ Cognitive Domains
2	.0155	0.0360	0.27	Negligible		
4	.0001	0.0419	0.34	Negligible		
5	**.0000**	**0.0000**	**0.65**	**Large**	**Girls**	**Number Knowing**
6	.0031	0.0002	0.38	Negligible		
11	**.0001**	**0.0000**	**0.46**	**Moderate**	**Girls**	**Number Knowing**
15	**.0000**	**0.0000**	**-0.44**	**Moderate**	**Boys**	**Algebra Applying**
16	**.0000**	0.0011	0.32	Negligible		

From Table 3, both methods of using Welch t-test and Mantel-Haenszel Chi-square methods flagged the same seven items with probability values of less than 0.05. Detailed classification using the ETS DIF category reveals that only one item recorded a large DIF and two items recorded moderate DIF from the seven items with the Mantel-Haenszel probability value of less than 0.05. They are Item 5 (flagged as having large DIF) from the content domain of Number and the cognitive dimension of Knowing. Two other items that signaled moderate DIF are Item 11 and Item 15. Just like Item 5, Item 11 is also from the content domain of Number and the cognitive dimension of Knowing, but, Item 15 is from the content domain of Algebra and the cognitive dimension of Applying.

The Mantel-Haenszel size is positive for the two DIF items (Items 5 and 11). This suggests that these two DIF items from the content domain Number and assess the cognitive domain Knowing favor the focal group (girls). Only one DIF item (Item 15) has a negative

Mantel-Haenszel size, which suggests that Item 15 from the content domain of Algebra and the cognitive dimension of Applying favors the reference group (the boys).

Discussions and Implications

DIF analyses using the Welch t-test and the Mantel-Haenszel Chi-square methods identified the same seven items as having the probability values of less than 0.05, which is the first condition of identifying DIF items according to ETS. Using the DIF size, one large and two moderate DIF items were identified. There appear to be a pattern as both items that favor girls (Items 5 and 11) are from the content domain of Number, while the single item (Item 15) that favors boys is from the Algebra content domain. When examining further these DIF items from the perspective of the cognitive process that they assess, items that assess the cognitive domain of Knowing (Items 5 and 11) favor girls, while the item (Item 15) that assesses Applying favors boys. In other words, items assessing the Number domain and Knowing favor girls, while Algebraic items assessing Applying favor boys.

Item 5 indicates a large DIF favoring girls. This raises the question as to why an item that assesses combined operations of fractions is biased toward boys. Item 11 indicates a moderate DIF and just like Item 5 also favors girls. In addition, it also assesses combined operations but of negative numbers. The common denominator for these two items is that they assess combined operations involving addition, subtraction and the use of brackets. Could it be possible that these two items were biased for the boys as they faced difficulty in the mastery of simplifying numbers in brackets, followed by multiplication or subtraction before proceeding finally to addition or subtraction. A possible explanation could be that they are simplifying from left to right as seen in some of their working. These two computation items on numerical ability seem to favor the girls

just like the findings of Abedalaziz, (2010b), who found that items assessing numerical ability favor girls.

Item 15 indicates a moderate DIF favoring boys. It is the only item among the three DIF items that is biased towards girls. Just like Item 5, Item 15 also assesses a concept in fraction but it appears to be a novel question that is not found in Malaysian textbooks or revision books. In order to solve this question, students need to halve both the terms on the left-hand side and right-hand side. The algorithm used is 'new' or unrehearsed. A possible explanation is that Item 27 assesses higher order thinking skills (HOTs). According to Rajendran (2010), HOTs involves solving non-routine questions using non-rehearsed algorithm unlike lower-order thinking skills that involves solving familiar 'text book' problems using well-known algorithms. Since this item requires students to go beyond a simple recall of a learned fact or application of routine problem, it is a HOTs item (Zohar & Dori, 2003) that does not favor the girls. This finding possibly suggests that HOTs item may not be favoring girls. These findings indicate that two of the computation items favor girls and that HOTS computation item favors boys. It appears that items with simple one step operation such as Items 5 and 11 favor girls and that boys are more appreciative of more challenging items that assesses HOTs such as Item 15.

Apart from that, as indicated in Figure 1, Item 15 is the most difficult item and as exhibited by Figure 2, it is more difficult for girls than boys. Unlike the findings of Berberoglu (1995) who highlighted that computation items favor boys, the findings of this study reveal that not all the computation items favor boys. The findings of this study concur with the findings of Le (2006) suggesting that science items which are more difficult tend to favor boys and not girls. Is it possible that difficult items and items that assess higher order thinking skills regardless of the subject domain tend to favor boys, even though girls generally tend to do better than boys?

Would girls still consistently do better than boys if tests are composed of more HOTs items? In addition, items from the content domain of Number and the cognitive domain of Knowing favor girls, while items from the Algebra content domain and Applying cognitive domain favor boys. Contradictory to the findings of Abedalaziz (2010a) that revealed girls performing better in Algebra, the findings of this study indicate that items from the Algebra content domain do not favour girls. Perhaps because the cognitive process that Item 15 is assessing is Applying, which is HOTs.

As Abedalaziz (2010a) discovered, items that assess lower-order thinking skills (LOTs) favour girls and as revealed in this study, all the two items that assess LOTs seem to favour girls. The only one item that favour boys appear to be a HOTs items and as suggested by Lindberg et al. (2010), items that assess HOTs favour boys. Therefore, the findings of this study seem to be aligned to the findings of previous studies that suggest an apparent trend of LOTs items favouring girls and HOTs items favouring boys.

The findings have broader implication to curriculum specialists and examinations boards in reducing gender-biased practices. Increasing the number of HOTs items in mathematics text books at all levels, including at the early stages of learning mathematics is a promising step forward that will enhance students' exposure and increase the opportunity to practice these items. Especially in view of the 21st century skills that necessitate critical thinking and critical numeracy, the inclusion of HOTs items is no longer an option but a relevant requirement. Thus, a revamp on the content and format of assessment may reduce gender bias as increasing the number of test items in classroom tests and examinations that invoke HOTs is a worthy consideration for test developers. The finding also has crucial implications for teacher educators. During instructions, students need to be more exposed to HOTs items both in classroom

discussions and as take home exercise for students of both gender to develop their HOTs. Questioning techniques directed towards enhancing and increasing the number of oral questions in classrooms that tap on students' HOTs need to be considered as a compulsory training module at teacher training institutes for both pre-service and in-service teachers.

Conclusion

At this initial stage of studying these three DIF items, it appears that computation items that assess concepts related to combined operation from the topics of fraction and negative numbers are biased against towards boys, while HOTs computation items are biased against girls. The findings of this study concur with that of Salubayba (2014) as two of these computation items favor girls. Items that assess the Knowing cognitive domain and from the Number content domain favor girls and items that assess the HOTs cognitive domain of Applying and from the content domain of Algebra favour boys. Even though it is rather premature to make generalizations as only one DIF item was detected as behaving in this manner, the distinctive trend of certain characteristics of items favoring certain gender groups tend to emerge from this study. Could it be possible that inclusion of more HOTs items in national assessments may revert the trend of boys not performing as well as our girls among non-native speakers of English? More research needs to be done to examine why HOTS computation items, with minimum language load does not favor girls. Future studies can be directed towards examining gender differential performance for HOTs items, specifically among non-native speakers of English.

References

N. Abedalaziz, N. A gender-related differential item functioning of mathematics test items. International Journal of Educational and Psychological Assessment, 5, 101- 116, 2010a.

N. Abedalaziz, N. Detecting gender related DIF using logistic regression and Mantel-Haenszel approaches. Procedia-Social and Behavioural Sciences, *7*(C), 406-413.2010b

J. Abedi, Courtney, M., Leon, S., Kao, J. & Azzam, T. English language learners and Math achievement: A study of opportunity to learn and language accommodation. Technical Report 702. CRESST, 2006.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. Standards for educational and psychological testing. American Educational Research Association, Washington, DC, 2014.

B. J. Becker, Item characteristics and gender differences on the SAT-M for mathematically able youths. American Educational Research Journal, 27, 65-87. 1990.

G. Berberoglu, Differential item functioning (DIF) analysis of computation, word problem and geometry questions across gender and SES groups. Studies in Educational Evaluation, 21(4), 439-456, 1995.

M. Carr and D. L. Jessup, Gender differences in first grade mathematics strategy use: Social and metacognitive influences. Journal of Educational Psychology, 89, 318- 328, 1997.

K. Y. Chan and J. Mousley, Using word problems in Malaysian mathematics education: Looking beneath the surface. In Chick, H. L. & Vincent, J. L. (Eds.), Proceedings of the 29th Conference of the International Group for the Psychology of Mathematics Education, 2, 217-224. Melbourne: PME, 2005.

A. Cohen, B. Bottge and C. S. Wells, Using Item Response Theory to assess effects of Mathematics instruction in special population. Council for Exceptional Children, 68:1, 23 – 44, 2001.

H. Davis, Gender gaps in Math and Science education. Undergraduate Research Journal for the Human Sciences. 7, 2008. Retrieved from http://www.kon.org/urc/v7/davis.html

D. Desa, (2014). Evaluating measurement invariance of TALIS 2013 complex scales: Comparison between continuous and categorical multiple-group confirmatory factor analyses. OECD Education Working Papers, No. 103, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/5jz2kbbvlb7k-en

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

N. J. Dorans and P. W. Holland, DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland, and H. Wainer (Eds.), Differential Item Functioning (pp. 35-66). Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.

G. Eisenkopf, Z. Hessami, U. Fischbacher and H. Ursprung, Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland, University of Konstanz Working Paper, 2012.

B. B. Ellis, P. Becker and H. D. Kimmel, An Item Response theory evaluation of an English version of the Trier personality inventory (TPI). Journal of Cross-Cultural Psychology, 24 (2), 133-148, 1993.

G. Engelhard, Gender differences in performance on mathematics items: Evidence from the United States and Thailand. Contemporary Educational Psychology, 15, 13-26, 1990.

K. Ercikan, M. J. Gierl, T. McCreith, G. Puhan and K. Koh, Comparability of bilingual versions of assessments: sources of incomparability of English and French versions of Canada's achievement tests. Journal of Applied Measurement, 17(3), 301-321, 2004.

M. M. Ferrara, The single gender middle school classroom: A close up look at gender differences in learning. Paper presented at the AARE 2005 Conference, Parramatta, Australia, November 2005.

E. Fennema, T. P. Carpenter, R. V. Jacobs, M. L. Frank and L. W. Levi, A longitudinal study of gender differences in young children's mathematical thinking. Educational Researcher, 27:5, 6-11, 1998. Retrieved from   http://www.jstor.org/stable/pdf/1176733.pdf

A. M. Fidalgo, D. Ferreres and J. Muniz, Liberal and conservative differential item functioning using Mantel Haenszel and SIBTEST: Implications for Type I and Type II error. Journal of Experimental Education, 73, 23–29, 2004.

M. Gamer and G. Engelhard Jr, Gender differences in performance on 45 multiple-choice and constructed response mathematics items. Applied Measurement in Education, 12 (1), 29–51, 1999.

M. Garner, and G. Engelhard, Gender differences in performance on multiple-choice and constructed response mathematics items. Applied Measurement in Education, 12, 29-51, 1999.

M. J. Gierl, Using a multidimensionality-based framework to identify and interpret the construct-related dimensions that elicit group differences. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA, April 12-16, 2004.

D. C. Geary, Sexual selection and sex differences in mathematical abilities. Behavioral and Brain Science, 19, 229-284, 1996.

P. W. Holland and D. T. Thayer, Differential item performance and the mantel-haenszel procedure. Test Validity, 129–145, 1988.

J. S. Hyde, E. Fennema and S. J. Lamon, Gender differences in mathematics performance: a meta-analysis. Psychological Bulletin, 107:2, 139–155, 1990.

Kolstad, Rosemarie and L. D. Briggs, Incorporating language arts into the Mathematics curriculum: A literature survey. Education, 116:3, 423, 1996.

S. Lane, N. Wang and M. Magone, Gender-related differential item functioning on a middle-school mathematics performance assessment. Educational Measurement: Issues and Practice, 15:4, 21-27, 1996.

L. Y. Lie, L. Angelique and E. Cheong, How do male and girls approach learning at NUS? CDTL Brief 7, 1–3, 2004.

C. S. Lim, Cultural Differences and Mathematics Learning in Malaysia. The Mathematics Educator, 7:1, 110-122, 2003.

J. M. Linacre, What do infit and outfit, mean square and standardized mean, 1994. Retrieved from http://www. Rasch.org/rmt/rmt162f.htm

J. M. Linacre, *Winstep*s (Version 3.67.0) [Computer Software]. (2008a). Chicago: Winsteps.com

J. M. Linacre, (2008b). *A user's guide to WINSTEP MINISTEP Rasch model computer programme manual 3.67*. Chicago: Winsteps.com

J. M. Linacre, (2011). Tutorial 3– *Investigating test functioning*. Retrieved from https://www.winsteps.com/a/winsteps-tutorial-further-3.pdf

S. M. Lindberg, J. S. Hyde, J. L. Petersen and M. C. Linn . New trends in gender and mathematics performance: a meta-analysis. Psychological Bulletin, 136 (6), 1123- 1135. 2010

O. L. Liu and M. Wilson, Gender differences in large-scale math assessments: PISA trend 2000 and 2003. Applied Measurement in Education, 22, 164-184, 2009.

Le, L. (2006). Analysis of Differential Item Functioning. Paper prepared for the Annual Meetings of the American Educational Research Association, April 7-11, San Francisco, Retrieved from https://www.acer.org/files/analysis_of_dif.pdf

Malaysian Ministry of Education. Malaysia Education Blueprint 2013-2025 (preschool to Post-Secondary Education). Putrajaya, Malaysia: Ministry of Education Malaysia, 2013.

[39] N.T. Longford, P.W. Holland and D.T. Thayer, Stability of the MH D-DIF statistics across populations. In: Holland P.W., Wainer H., eds. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1993.

P. P. Maranon, M. I. Garcia and C. S. L. Costas, Identification of non-uniform Differential Item Functioning: a comparison of Mantel-Haenszel and Item Response Theory analysis procedures. Journal of Educational and Psychological Measurement, 57:4, 559 – 568, 1997.

M. O. Martin, I. V. S. Mullis and G. M. Stanco, TIMSS 2011 international results in science. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/isc/publications.html. 2012

S. Mendes-Barnett and K. Ercikan. Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. Applied Measurement in Education, 19, 289-304, 2006.

Mullis, I. V. S. TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth grade. Chestnut Hill, MA: International Study Center, Boston College, Lynch School of Education, 2003.

I. V. S. Mullis, M. O. Martin and P. Foy, TIMSS 2007 International Mathematics report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth

and Eighth Grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College, 2008.

I. V. S.Mullis, M. O. Martin, P. Foy and A. Arora, *TIMSS 2011 International Results in Mathematics*.TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA, 2012.

I. V. S. Mullis, M. O. Martin, P. Foy and M. Hooper, TIMSS 2015 International Results in Mathematics*, 2016*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: http://timssandpirls.bc.edu/timss2015/international-results/

T. S. Neidorf, M. Binkley, K. Gattirs and D. Nohara, Assessment technical report: Comparing mathematical content in the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science study (TIMSS) and the Programme for International Students Assessment (PISA) 2003, 2006. US Education Department of Education Statistics, US Department of Education, Institute of Education Sciences, NCES 2006-029. Retrieved from http://nces.ed.gov/pubs2006/2006029-2.pdf

S. M. Lindberg, J. S. Hyde, J. L. Petersen, and M. C. Linn, New trends in gender and mathematics performance: a meta-analysis. Psychological bulletin, 136:6, 1123–1135, 2010.

S. Mendes-Barnett, and K. Ercikan, Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. Applied Measurement in Education, 19: 4, 289–304, 2006.

C. Randall, Solving Word Problems: Developing Students' Quantitative Reasoning Abilities. Reach Into Practice Mathematics, 2011. Pearson. Retrieved from

https://assets.pearsonschool.com/asset_mgr/current/201034/Randy%20Charles%20Monog raph.pdf

N. S. Rajendran, Teaching & acquiring higher-order thinking skills, Penerbitan Universiti Pendidikan Sultan Idris, Perak, Malaysia, 2010.

S. K. Reed, Learning rules in word problems: Research and Curriculum Reform, 1999. Lawrence Erlbaum Associates. Retrieved from http://books.google.com.my/books?id=0FcfDR4PBHOC7dq=nctm= (Word Problem)

L. A. Roussos and W. F. Stout, A multidimensionality-based analysis of DIF paradigm. *Applied Psychological Measurement*, 20:4, 355-371, 1996.

T. M. Salubayba, Determining Differential Item Functioning in Mathematics Word Problems Using Item Response Theory, 2014. Retrieved from http://www.iaea.info/documents/paper_226dc2c441.pdf

C. A. Stone and B. Zhang, Assessing goodness of fit of Item Response Theory models: A comparison of traditional and alternative procedures. Journal of Educational and Psychological Measurement, 40:4, 331-352, 2003.

J. Pedrajita, Using logistic regression to detect biased test items. International Journal of Educational and Psychological Assessment, 2, 54-73, 2009.

W. M. Yen and A. R. Fitzpatrick, Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 111-153). American Council of Education, United States of America: Praeger Publishers, 2006.

E. Wehrwein, H. Lujan and S. DiCarlo, Gender difference in learning style preferences among undergraduate physiology student. Advances in Physiology Education, 31, 153-175, 2007.

B. Wright and Linacre (1994). A Rasch Unidimensionality coefficient [Electronic Version]. Retrieved from http://www.rasch.org/rmt/rmt83p.htm

A. Zohar and Y. J. Dori, Higher order thinking skills and low achieving students: Are they mutually exclusive? Journal of the Learning Sciences, 12:2, 145-182, 2003.

R. Zwick, D. T. Thayer and C. Lewis, An empirical Bayes approach to Mantel– Haenszel DIF analysis. Journal of Educational Measurement, 36, 1–28, 1999.