DO BOYS AND GIRLS INTERPRET MATHEMATICS TEST ITEMS SIMILARLY? INSIGHTS FROM RASCH MODEL ANALYSIS

Hasni Shamsuddin*, Nordin Abd. Razak, Lei Mee Thien and Ahmad Zamri Khairani

School of Educational Studies, Universiti Sains Malaysia, 11800 USM Pulau Pinang, Malaysia

*Corresponding author: emelhasni@gmail.com

Published date: 30 June 2020

To cite this article: Hasni Shamsuddin, Nordin Abd. Razak, Lei Mee Thien, & Ahmad Zamri Khairani. (2020). Do boys and girls interpret mathematics test items similarly? Insights from Rasch model analysis. *Asia Pacific Journal of Educators and Education, 35*(1), 17–36. https://doi.org/10.21315/apjee.35.1.2

To link to this article: https://doi.org/10.21315/apjee.35.1.2

Abstract: The difference between boys and girls in mathematics achievement have long been a concern among educators. Various factors are associated with this issue, and the present study will discuss one of the factors – namely, the extent to which both groups of boys and girls can interpret mathematics item similarly. Data was collected from 158 Form 1 students (boy = 76, girl = 82) in a secondary school in Perak. A 43-item partial credit test was developed to gather information from the students. Rasch model analysis was employed for the following purposes: (1) to examine the quality of the items, and (2) to provide empirical evidence whether both groups interpret the test items similarly. Results showed that all items satisfactorily met the model's expectations. Meanwhile, based on the differential item functioning (DIF) statistics, we found that five items differed significantly, with three items favoured the boys while two items favoured the girls. We also discussed the implication of the study, particularly towards the teaching and learning of mathematics in the classroom. The findings provide empirical evidence about whether boys and girls interpret the test items similarly. This finding suggests that teachers should seriously consider the difference between boys and girls in mathematics achievement when designing their teaching and learning strategies in the classroom.

Keywords: Rasch model analysis, boys and girls, differential item functioning (DIF), quality of the items, mathematics achievement

INTRODUCTION

Differences in gender performance on mathematics tests have long been the focus of research. In general, extensive studies have found boys perform better in mathematics achievement compared to girls (Mendes-Barnett & Ercikan, 2006). In Malaysia, the issue has been examined at national (e.g., Numeracy: Ameer & Singh, 2013) and at international levels (e.g., Malaysia-Singapore: Ismail & Awang, 2009; PISA: Husin, 2014; TIMSS: Ministry of Education, 2016a). These studies report empirical findings of the differences, strengths, and weaknesses of each group, as well as possible reasons why the differences exist, especially among boys and girls. Surprisingly, the results from these studies show that girls perform better than their boy counterparts. For instance, the report from Trend in Mathematics and Science (TIMSS) 2015, girls outperform boys in both cognitive as well as content domains in Malaysia. Table 1 shows a comparison between their scores in the respective domains in Malaysia.

Table 1. Comparison mathematics scores between boys and girls in TIMSS 2015

Domain	Score	
	Boy	Girl
Cognitive		
Knowledge	462	482
Application	461	465
Reasoning	452	454
Content		
Number	469	474
Algebra	458	476
Geometry	453	457
Data and probability	447	456

Source: Ministry of Education (2016a)

Gender differences in mathematics is essential since it has implications, particularly in teaching and learning. Gender differences may shape the teachers' belief about the performance of their students. For example, studies such as Legewie and DiPrete (2012) shows that some teachers have been found to hold gender-stereotyped expectations of boys and girls. Meanwhile, Cornwell, Mustard and Van Parys (2013) report that teachers may rate the students' work differently based on gender. Moreover, teachers' belief about gender differences may result in different interaction in the class (Spilt, Koomen, & Jak, 2012). Due to different beliefs, female students may not have the same opportunities in the classroom that in turns, may affect their future participation in mathematics. As such, it is essential to understand gender differences, particularly in terms of how much is the difference and where the difference exists.

Gender differences in mathematics scores have been studied from various perspectives. The differences have been identified in multiple aspects. For example, from a social standpoint, gender differences are associated with stereotypes (Cvencek, Meltzoff, & Greenwald, 2011). Meanwhile, from the cognitive perspective, the difference can be associated with the difference in spatial reasoning between boys and girls (Klein, Adi-Japha, & Hakak-Benizri, 2010). Also, gender differences were related to the affective domain of mathematics, such as mathematics ability attribution (Dickhauser & Wulf-Uwe, 2006). Nevertheless, despite intensive studies, there is no comprehensive explanation of the source of gender differences in mathematics. Several reasons have been attributed to these differences, from biological factor (Spelke, 2005), learning strategies (Bezzina, 2010) to culture (Fryer & Levitt, 2010). Even city location provides influence on the difference between boy and girl (Isiksal & Cakiroghi, 2010).

However, we would like to suggest that the difference in achievement between boys and girls is due to differences in how they understand the mathematics items. Based from the studies by Zhang and French (2010) as well as Taylor and Lee (2012), we argue that gender differences can be identified at the item level (by item format and item content) rather than at test level (by test scores). In other words, studies at item level involve examining every item in the test to identify whether there are items that show gender differences. Item level analysis is crucial since it is highly likely that each item in a test is of different difficulty. As such, each item may be perceived differently by the boys and the girls.

In contrast, in test level analysis, each item is treated as having a similar difficulty. Therefore, the score of each item is added to calculate the overall test score. Gender differences at this level is examined using statistics such as t-tests or Mann-Whitney U test. At the item level, a gender differences in mathematics can be investigated using the differential item functioning (DIF) analysis (Lyons-Thomas, Sandilands, & Ercikan, 2014). DIF in mathematics refers to the situation where examinees from two groups such as boy and girl with equal mathematics ability have different probabilities of answering correctly on a particular item (Zumbo, 1999). DIF can be found in the item difficulty parameter; that is, a specific item is perceived as having significantly different difficulty level by a different group. DIF analysis is vital since it may affect the performance of the groups (Pae & Park, 2006).

OBJECTIVE OF THE STUDY

Objectives of the present study are given as follows:

1. Examine the quality of mathematics test items based on Rasch model analysis.
2. Identifies items that exhibit gender DIF.

LITERATURE REVIEW

Hanizah (2007) examines gender DIF in the high-stakes mathematics test with regards to the item types. The sample consists of 1,213 Form 5 students, and she employs a 46 multiple choice mathematics achievement test. It was documented that items with texts and number favour the girl students, while the boy students have the advantages when it comes to spatial and graphics test items. Meanwhile, Abedalaziz (2010) investigates gender DIF in mathematics item among 3,390 students. Using the transformed difficulty, Mantel-Haenszel, and item-characteristic curve methods, the results show that difference in the difficulty of the items exists between boys and girls. Most of the items are identified as favouring the girls. Kanageswary and Ong (2014) examine the gender DIF between 4,768 Form 2 students using two different DIF procedure, namely, the Mantel-Haenszel and the multidimensional procedures. Results show that gender DIF items were identified using both procedures.

DIF analysis also heavily used in other subjects or competencies. For example, Izazol (2012) examines gender DIF for science process skills among 570 Form 4 students using the Rasch model framework. Results show that four of the items in the 36-item Test for Integrated Science Process Skills II (TIPS II) used in the study demonstrates significant gender differences. She suggests that demographic factors might be influential in the outcome of the study. Meanwhile, Lim and Subashini (2015) investigate gender DIF among 1,000 Form 4 students in the subject of Principle Accounting. Using the same measurement framework, they find that eight of the 40-item instrument are seen differently (with regards to difficulty) between boy and girl students.

In short, various studies provide empirical evidence on how boy and girl groups see a particular item in terms of their difficulty. However, these mentioned studies focus heavily on the identification of gender DIF items but lack the explanation on the source of the difference as well as its impacts on the teaching and learning in a classroom.

THEORETICAL FRAMEWORK: THE RASCH MODEL ANALYSIS

The Rasch model is a modern theory that relates two important parameters in a measurement of any construct, namely, (1) student ability and (2) items difficulty. The used of polytomous data required multiple steps to assign partial credit for completing several steps in the solution process. The Partial Credit Model (PCM; Masters, 1982) which can be considered as an extension of the Rasch model was designed to be used when partial credit can be awarded for degrees of success. The equation for the partial credit model is shown in Equation (1) (Masters, 1982).

$$P(\theta)_{xni} = \frac{e\sum_{j=0}^{x}(\beta_n - \delta_{ij})}{\sum_{k=0}^{m}\left[e\sum_{j=0}^{k}(\beta_n - \delta_{ij})\right]} \qquad x = 0, 1, ...\mathrm{m} \qquad (1)$$

where

$P(\theta)_{xni}$ = probability of student n completing x steps on item i

β_n = ability of student n

δ_{ij} = difficulty of item i on step j

The PCM has all the standard Rasch model features. Students ability, β, and items difficulty, δ, parameters were estimated based on the ratio of a number of correct items to a number of incorrect ones. This score was then transformed into equal interval score (call *measure*) using natural log (ln) or 'logits' unit in a procedure called calibration. Measures from Rasch model calibration is essential in measurement since it possesses the equal-interval property as in a thermometer or a ruler. With this property, Rasch model measurement not only able to identify the difference between items but also specify how much the differences are. Overall, items with higher measures were more difficult to answer, that in turns, represent more difficult-to-master knowledge, skills and abilities (KSAs).

Even though the Rasch model provides essential information on the measured construct, its analysis came with two strong assumptions that must be met for the data to have the property of equal interval. Firstly, the data needs to fit the model's expectation. This quality-control assumption is to ensure that the data collected did not contain too much unintended construct or 'noise'. According to Bond and Fox (2015), the infit and outfit mean-square (MNSQ) values of 0.6–1.4 are considered reasonable to ensure this assumption is being fulfilled. The statistics indicate how accurately the data fit the model. Infit is an inlier-sensitive and highly influenced by the pattern of responses to items targeted on the students, whereas the outfit is more sensitive to responses to items with difficulty far from a person (Linacre, 2002). The MNSQ fit statistics show the amount of distortion (or noise) of the measurement. The expected values of MNSQ are 1.00, where this indicates little distortion in the measurement. Linacre (2002) also cautions that infit statistic is important indicator than the outfit for evaluating the threat to the measurement.

Secondly, the data collected must pose unidimensionality property. This assumption states that the test should measures only a single construct (Wright & Masters, 1982). In Rasch analysis, this assumption is assessed using the principal component analysis (PCA) of residuals procedure. This procedure seeks to identify the presence of the second construct when the primary construct was extracted out. The procedure is looking to gather information if groups of items share the same unexpected pattern apart from the intended measured construct. If the pattern exists, then the items may share a substantive attribute in common, that is, the so-called secondary dimension. Eakman (2012) quotes that the issue of unidimensionality is compromised when the unexplained variance explained by the first contrast (eigenvalue size) was less than 3.0.

Within the framework of the Rasch model, an item is considered as exhibiting gender DIF when the difference in item difficulty between two groups is significant ($|t| > 1.96$) in the Winsteps 3.63 output. For every item with substantial gender differences, percentages of boy and girl students that achieve the full and partial scoring are presented. This additional information is provided so that readers will have a better picture of where the differences lies with regards to every item.

METHODOLOGY

Research Design

The present study employed a cross-sectional study research design. Data were collected in a single time period, and the study involved exploring descriptive characteristics of mathematics items that show gender differences between the boys and the girls.

Respondent

A total of 158 Form 1 (13 years-old) students (boy = 76, girl = 82) in a secondary school in Perak provided responses for this study. Data from the respondent is collected during the mid-year examination conducted in May 2018.

Instrumentation

The instrument for this study is a self-developed paper and pencil test. It contains 43 partial credit test items. We develop the test item based on Standard Based Curriculum for Secondary School (Ministry of Education, 2016b). Each item measures one specific learning outcome from the curriculum. The test was validated in terms of its contents by two expert teachers with 40 years of experience between them. Scoring is done based on the completion of steps. Examples of scoring are given as follows:

Example 1: Item 12

The volume of a cube is $3\frac{3}{8}\text{cm}^3$. Find the length of each side of the cube. [2 marks]

Scoring:

$\sqrt[3]{\frac{27}{8}}$ ..... 1 mark

$\frac{3}{2}$ @ 1.5 cm ...... 1 mark

Example 2: Item 20

8 apples were sold at the price of RM4.80, while 9 oranges were sold at RM6.30. Jamal bought 25 apples and 13 oranges. How much money should he pay? [3 marks]

Scoring:

$\frac{4.80}{8}$ @ $\frac{6.30}{9}$ ..... 1 mark

$25 \times \frac{4.80}{8} + 13 \times \frac{6.30}{9}$ ..... 1 mark

24.10 ..... 1 mark

Data Collection

Data collection was conducted with the help of the schoolteachers. The test was administered in a mid-year examination to ensure good returns. The students were given two hours to complete the test. The data were then transferred to the IBM SPSS 23.0 statistics software package.

Data Analysis

In this study, a Rasch model analysis was employed for data analysis using the software named Winsteps 3.63. The primary statistics for this study was the DIF analysis. However, firstly, we examine the issue model-data fit issue using the infit and outfit MNSQ statistics for every item using the guideline of acceptable values of between 0.6 and 1.4 logits (Bond & Fox, 2015). In this study, the dimensionality assumption was investigated using the guidelines provided by Eakman (2012), in which the unexplained variance from the second construct extracted from the procedure needs to be less than 10%, whereas the eigenvalue extracted from the PCA of residual should be less than 3.0. With regards to the DIF analysis, we flagged items that demonstrated the significant difference in DIF measures ($t > 1.96$, or $t < -1.96$) between the boys and the girls for further discussions.

RESULTS

Table 2 shows statistics for all 43 items. The raw score provides information on the item score. For example, for Item 1, the raw score is 246. This score might come from 100 students who score two marks, 46 students who score one mark and 12 students who score 0 marks. The measure statistics is the item difficulty parameter estimated from the Equation (1) discussed in the previous section. It should be noted that both the raw score and item difficulty measure is inversely correlated. That is, higher the raw score indicates that more students can answer the item that in turns produces lower item difficulty measure. The standard error (SE) statistics indicates the precision of the Rasch estimation. The smaller the value of SE, the more precise the estimation is.

Meanwhile, the infit and outfit MNSQ statistics report the amount of the unintended 'noise' measured by the item (see the Data Analysis section). It shows that the values of the infit and outfit MNSQ statistics did not exceed 1.4, thus indicates that the data fulfilled the model-data fit assumption. Note that some items such as Items 2, 5 and 10 demonstrate infit and outfit values of less than 0.6 and considered less productive, the items did not degrade the measurement (Wright & Linacre, 1994). As such, these items are kept for further analysis.

Item 33 is the most difficult item for the sample of students based on the highest value of item difficulty measure (3.22 logits) followed by Item 30 (2.06 logits).

Meanwhile, Item 20 is the easiest item since it has the lowest item difficulty measure (–3.56 logits) followed by Item 43 (–2.84 logits).

Table 2. Item statistics

Item	Topic	Raw Score	Measure (logits)	SE (logits)	Infit MNSQ	Outfit MNSQ
1	Rational numbers	246	–0.54	0.13	0.92	0.98
2	Rational numbers	185	0.27	0.10	0.58	0.66
3	Rational numbers	321	–1.99	0.07	1.05	1.02
4	Rational numbers	250	–0.79	0.15	1.26	1.38
5	Factors and multiples	278	–1.31	0.18	0.56	0.61
6	Factors and multiples	118	0.97	0.10	1.04	1.05
7	Factors and multiples	160	0.54	0.10	1.13	1.13
8	Factors and multiples	129	0.86	0.10	1.30	1.34
9	Factors and multiples	146	0.68	0.10	1.31	1.33
10	Squares, square roots, cubes, cube roots	234	–0.34	0.12	0.56	0.51
11	Squares, square roots, cubes, cube roots	189	–1.68	0.08	1.02	0.99
12	Squares, square roots, cubes, cube roots	91	1.28	0.11	0.67	0.65
13	Squares, square roots, cubes, cube roots	89	0.00	0.17	0.88	0.87
14	Squares, square roots, cubes, cube roots	126	0.89	0.10	0.94	0.94
15	Squares, square roots, cubes, cube roots	355	–2.16	0.07	1.14	1.15
16	Squares, square roots, cubes, cube roots	65	1.62	0.12	1.14	0.94
17	Squares, square roots, cubes, cube roots	109	–0.83	0.09	1.08	0.99
18	Ratios, rates and proportions	108	1.08	0.11	1.03	1.08
19	Ratios, rates and proportions	251	–2.06	0.08	0.64	0.70
20	Ratios, rates and proportions	385	–3.56	0.16	1.01	0.88
21	Ratios, rates and proportions	135	0.80	0.10	1.08	1.09
22	Ratios, rates and proportions	198	0.13	0.11	0.95	0.88
23	Ratios, rates and proportions	333	–2.61	0.09	1.01	1.17

(*continued on next page*)

Table 2. (*continued*)

Item	Topic	Raw Score	Measure (logits)	SE (logits)	Infit MNSQ	Outfit MNSQ
24	Ratios, rates and proportions	176	0.37	0.10	1.11	1.12
25	Ratios, rates and proportions	88	–0.96	0.10	1.02	0.85
26	Ratios, rates and proportions	269	–2.17	0.08	0.98	1.00
27	Algebraic expressions	79	1.43	0.11	0.99	0.98
28	Algebraic expressions	78	1.44	0.11	0.66	0.66
29	Algebraic expressions	203	–1.77	0.08	1.25	1.25
30	Algebraic expressions	40	2.06	0.15	1.05	0.89
31	Algebraic expressions	98	1.20	0.11	1.12	1.13
32	Algebraic expressions	144	0.70	0.10	0.77	0.77
33	Algebraic expressions	10	3.22	0.29	1.39	0.82
34	Algebraic expressions	94	1.24	0.11	1.22	1.19
35	Algebraic expressions	39	1.50	0.19	1.02	1.06
36	Algebraic expressions	102	1.15	0.11	0.98	0.95
37	Squares, square roots, cubes, cube roots	135	0.80	0.10	0.78	0.74
38	Squares, square roots, cubes, cube roots	79	–0.87	0.10	1.31	1.27
39	Linear equations	65	1.62	0.12	1.13	1.08
40	Factors and multiples	100	1.17	0.11	1.13	1.15
41	Factors and multiples	213	–1.83	0.08	1.32	1.34
42	Ratios, rates, and proportions	91	1.28	0.11	0.75	0.73
43	Ratios, rates, and proportions	312	–2.84	0.10	1.25	1.22
	Mean	160.80	0.00	0.11	1.01	0.99
	SD	92.20	2.53	0.04	0.22	0.22

Note: SE = standard error, SD = standard deviation

Meanwhile, the result from the PCA of residuals showed that the unexplained variance explained by the first contrast (eigenvalue size) was less than 3.0 (Table 3). As such, the unidimensionality assumption was also fulfilled (Eakman, 2012). That is, the test items were only measuring students' mathematical ability and no other unintended constructs.

DIF analysis was conducted to assess construct equivalence between the boy and girl students. As depicted in Table 4, 41 items (95.35%) show differences in item difficulty between boys and girls. For example, for Item 1, item difficulty measure for boy is –0.63 logits, while for girls perceive the item as slightly harder, that is –0.45 logits. Nevertheless, since the *t*-values for this difference is –0.69, which is lower than the cutoff value of |*t*| > 1.96, this difference is considered as not significant. In contrast, Item 3, there is a considerable difference between the item difficulty measures between the two groups. Based on their responses, the girls acknowledge that the item is significantly easier (–2.19 logits) compared to the boy students (–1.78 logits) based on the significant *t*-values of 2.9 which is larger than 1.96. Thus, Item 3 is considered as favouring the girls. Table 5 depicts the percentages of scoring for both groups. For example, 44.74% of the boy students score 0 for Item 3, while only 21.95% of girls have the same score. It can be seen that there are more girl students score 2 for the item compared to the boys, that in turns, contributes to lower item difficulty measure for the girls (–2.19 logits) compared to the boy (–1.78 logits). Apart from Item 3, Item 7 also demonstrated significant gender DIF (*t* = 2.14) and also favouring the girls (Table 6). Meanwhile, the item difficulty measure statistics showed that another three items that demonstrated significant gender DIF (Items 8, 9 and 26), were favouring the boys. Note that Items 3, 7, 8 and 9 are related to the domain of Number and Operations, while Item 26 is associated with the domain of Relationship and Algebra (Curriculum Development Division, 2016). These items were subjected to further scrutiny, especially concerning their impact on the teaching and learning mathematics in the classroom.

Table 3. Results from the PCA of residuals

		Empirical (%)		Modelled (%)
Total raw variance in observations	=	84.2	100.0	100.0
Raw variance explained by measures	=	41.2	48.9	50.1
Raw variance explained by persons	=	9.1	10.8	11.1
Raw variance explained by items	=	32.1	38.1	39.0
Raw unexplained variance (total)	=	43.0	51.1	49.9
Unexplained variance in 1st contrast	=	2.5	3.0	5.9
Unexplained variance in 2nd contrast	=	2.3	2.7	5.3

Table 4. Item difficulty and *t*-value statistics

Item	Learning standards	Item difficulty (Boy)	Item difficulty (Girl)	*t*
1	Rational numbers	–0.63	–0.45	–0.69
2	Rational numbers	0.27	0.25	0.11
3	Rational numbers	–1.78	–2.19	2.9
4	Rational numbers	–0.76	–0.84	0.26
5	Factors and multiples	–1.37	–1.26	–0.28
6	Factors and multiples	0.84	1.09	–1.21
7	Factors and multiples	0.76	0.31	2.14
8	Factors and multiples	0.63	1.07	–2.15
9	Factors and multiples	0.44	0.91	–2.29
10	Squares, square roots, cubes, cube roots	–0.23	–0.48	1.01
11	Squares, square roots, cubes, cube roots	–1.62	–1.74	0.75
12	Squares, square roots, cubes, cube roots	1.16	1.38	–1.02
13	Squares, square roots, cubes, cube roots	–0.07	0.08	–0.43
14	Squares, square roots, cubes, cube roots	0.86	0.89	–0.12
15	Squares, square roots, cubes, cube roots	–2.05	–2.27	1.53
16	Squares, square roots, cubes, cube roots	1.47	1.76	–1.2
17	Squares, square roots, cubes, cube roots	–0.93	–0.75	–1.04
18	Ratios, rates and proportions	1.08	1.08	0
19	Ratios, rates and proportions	–2.06	–2.06	0
20	Ratios, rates and proportions	–3.53	–3.59	0.19
21	Ratios, rates and proportions	0.74	0.85	–0.54
22	Ratios, rates and proportions	0.07	0.18	–0.53
23	Ratios, rates and proportions	–2.59	–2.64	0.28
24	Ratios, rates and proportions	0.29	0.44	–0.73
25	Ratios, rates and proportions	–0.96	–0.96	0
26	Ratios, rates and proportions	–2.34	–2.01	–2.11
27	Algebraic expressions	1.59	1.29	1.29
28	Algebraic expressions	1.42	1.44	–0.11
29	Algebraic expressions	–1.74	–1.80	0.38
30	Algebraic expressions	1.99	2.13	–0.47
31	Algebraic expressions	1.2	1.20	0

(*continued on next page*)

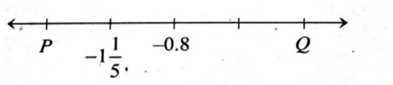
Table 4. (*continued*)

Item	Learning standards	Item difficulty (Boy)	Item difficulty (Girl)	*t*
32	Algebraic expressions	0.65	0.75	–0.46
33	Algebraic expressions	3.33	3.13	0.33
34	Algebraic expressions	1.24	1.24	0
35	Algebraic expressions	1.53	1.47	0.14
36	Algebraic expressions	1.34	0.99	1.61
37	Squares, square roots, cubes, cube roots	0.69	0.89	–0.95
38	Squares, square roots, cubes, cube roots	–0.82	–0.92	0.46
39	Linear equations	1.65	1.59	0.26
40	Factors and multiples	1.14	1.20	–0.3
41	Factors and multiples	–1.73	–1.92	1.27
42	Ratios, rates and proportions	1.31	1.25	0.29
43	Ratios, rates and proportions	–2.87	–2.81	–0.25

Descriptions of items as well as the topic that demonstrated significant DIF are given as follows:

Item 3 (Rational number, *t* = 2.90, favouring girls)

Find the values of *P* and *Q* in the following line number.



P : ____________________ *Q*: ______________________ [2 marks]

Item 7 (Factors and multiples, *t* = 2.14, favouring girls)

Write the first three common multiples of numbers 4 and 6. [2 marks]

Item 8 (Factors and multiples, *t* = –2.15, favouring boys)

3 is a factor of *x*
x is a factor of 12

Based on the above information, find all possible values of *x*. [2 marks]

Item 9 (Factors and multiples, *t* = –2.29, favouring boys)

There are three traffic lights at an intersection. The first will turn red after 18 seconds, and the second will turn red after 16 seconds while the third will turn red after 24 seconds. When will all the traffic lights turn red simultaneously again?
[2 marks]

Item 26 (Ratios, rates and proportions, *t* = –2.11, favouring boys)

A shirt costs RM *y*, and a pair of shoe costs RM *z* was sold at M. Nisa shop-bought three shirts and two pairs of shoes.

Construct an algebraic expression that describes the situation above. [3 marks]

DISCUSSION

Item 3 requires the students to find the values on a number line. It assessed students' understanding of the concept of negative numbers on a number line as well as their ability to see the difference between mixed numbers (and decimals (–0.8). The number line can be considered as a graphic stimulus that helps the students to find the values of the unknown (*P*). The item evaluates the skill of computation in numbers, including negative rational number such as fraction and decimal before using the number line to solve it. The result concurs well with findings from previous studies showed that girl always has an advantage in number sense compared to their boy counterparts (Hanizah, 2007). That is, the girls have a higher ability to deal with items that involve numbers. Similarly, Item 7 requires the student to determine the values of multiples and common multiples of the number given. The item was also favouring girls because the item assessed students in understanding the concept of numbers. Crombie, Sinclair, Silverthorn, Byrne, DuBois and Trinneer (2005) suggest that the advantage may be explained by the fact that their superior mathematical self-concept more influences the girls compared to boys. That is, the girls are more confident to solve items related to the number sense compared to the boys.

Meanwhile, Item 8 requires the student to find the value of an unknown, namely, *x*. To solve the item, students need to find the multiples of 3 and the factors of 12 before comparing the equivalent values to find the answers. The indirect statement from the stimulus, which mentioned that 3 is the factor of *x* is unfamiliar compared to the second statement that required students to find the factor of 12. We believe the nature of how the item was written may influence the result. This is because as documented by the study of Innabi and Dodeen (2018), the boys are better at understanding unfamiliar statement compared to the girls who are better at gauging straightforward information from the item. The same reason might be able to explain why Item 9 favours the boys. Item 9 is a problem-solving item which is related to a real-life situation. It assessed students' understanding of the information. More specifically, to solve Item 9, the students need to understand the problem, change the information into a mathematical equation, and find rules or formula to solve the equation. This is considered a daunting task for the girls who are more able at a straightforward item. The result is also very much similar to the study by Innabi and Doddeen (2018) on Jordanian students.

Item 26 requires students to form an algebraic expression from the statement given. The item assessed students' understanding of the meaning of the unknown and how to differentiate between the variable and the non-variable. To solve this question, the students need to show the relationship of the unknown to represent the situation in mathematics statement. Students also needed to understand the meaning of the statement and changing them to the mathematics equation. The fact that this item is favouring the boys requires further study, especially regarding the mental image of the students. In another words, there is a possibility that the boys process the information differently (and more accurately) compared to the girls. Assessing how both groups process the information might shed some light regarding Item 26.

The present study documents that gender DIF is detected, particularly in items related to the domain of Number. Nevertheless, the findings of this study showed mixed results, where some of the items associated with the domain of Number favour boy, whereas some other items favour the girls. This contradicts the findings of previous studies that showed girl always have an advantage in Number sense compared to their boy counterparts (Hanizah, 2007). In this study, even though the number of items that exhibit gender DIF is considered small, which is 5 out of 43 (11.63%), the fact that the difference exists needs to be taken seriously especially by the teachers. As such, this discussion section will focus on two important issues. Firstly, the fact that most of the items are in the domain of Number requires some immediate attention. This is because number sense is found to be related to mathematical ability (Jordan, Kaplan, Oláh, & Locuniak, 2006). That is, mastery in number sense may have resulted in higher achievement in mathematics. Students with poor Number sense can be identified as having fewer strategies in completing skills related to calculations and rely heavily on calculators to complete the tasks. To address this issue, teachers may try to improve the students' estimation skills as well as exposing them to various strategies to solve problems. Also, the students need to see connections between numbers since it is the bases for successful completion of fundamental skills in mathematics such as addition, subtraction, multiplication, and division.

Secondly, in the domain of Number and Operations, the findings of this study showed mixed results, where some of the items favour the boys, whereas some other items favour the girls. This contradicts the findings of previous studies that showed girl always have an advantage in number sense compared to their boy counterparts (Hanizah, 2007). Crombie et al. (2005), in their study in mathematics competence, suggest that the advantage might because of their math self-concept more influences girls. In this study, most of the items that favour girl evaluate the skill of computation in numbers, including positive and negative number, fraction, decimal, and an integer. Computational skills are defined as abilities to calculate basic addition, subtraction, multiplication, and division. The computational skills are the basic concept in mathematics which includes numbers in the domain of Number and Operations. However, the items which favour boys are not only evaluating the mathematical knowledge in computational skill including procedure and concept in numbers but also assess the relationship of the unknown value to solve the problem in the domain of Number and Operations. The results are familiar with the research from Jordan by Innabi and Dodeen (2018), which showed that boys answered correctly to the more complicated, unfamiliar and life-related mathematical problems. Furthermore, this study also showed that the learning standard of the last item which favour boy is to derive algebraic expressions to show the relationship of the unknown to represent a situation. Based on the study, the students need to interpret the relation between numbers since it is the bases for successful completion of computational skills in mathematics.

Thirdly, teachers should acknowledge that boys and girls see some of the mathematics items differently. As such, the best teaching practice should differ as well, where teachers should modify their teaching by taking consideration of various factors that affect the boys' and girls' performances in mathematics. For example, the teachers should always observe emotion towards mathematics teaching and learning, especially among the girls. This is because they were reported to enjoy less pride, excitement, and hope but higher anxiety compared to the boys during class (Frenzel, Pekrun, & Goetz, 2007). Teachers, thus, should encourage activities that would able to reduce these negative emotions, such as active learning, among his or her students in the classroom. It is also interesting to know that girls are more anxious and less confident about the examination compared to the boys (Roger, 2003). Girls require more time for problem-solving items, always feel the need to verify their answers, less likely to take risks and have the tendency leave more problems blank than boys (Forgasz, Becker, Lee, &

Steinthorsdottir, 2010). As such, it is perhaps useful for teachers to provide prior examination setting such as a mock test or trial examination to minimise the effect of test anxiety among the girls. Meanwhile, since the boys are less able in terms of items related to number sense, we recommend that the teachers can provide more exercises so that the boys can improve on their computational skills. More exercises can encourage students to become more fluent in recalling mathematics facts and formulae, which are essential parts for computational skills.

Finally, it is worth to mention that apart from the students themselves, teachers are considered as the most important factors that contribute to students' achievement. As observed by Hattie (2009), it is estimated that teachers account for about 30% of the variance in students' achievement. As such, there is no surprise that teachers also contribute to gender differences in students' performance by engaging and having counter-productive activities and belief such as holding stereotyped expectations of the boys and girls (Cvencek et al., 2011), practicing double-standard, i.e. assessing the boys and girls work differently (Cornwell et al., 2013) as well as interact differently with different gender (Spilt et al., 2012).

CONCLUSION

The purpose of this study is to investigate whether the boy and girl students see mathematics items similarly with regards to item difficulty. Using gender DIF analysis provided by the Rasch model framework, the difference in item difficulty parameter is detected for every item. However, only five items exhibit significant gender DIF. Most of these items are in the domain of Numbers. Nevertheless, the present study provides mixed results, in which some results concur well with previous studies, while others did not. We also offer discussion on improving students' learning in Number sense as well as teachers' role in addressing the issue of gender differences in mathematics.

This research area still has a lot of unanswered questions, however, we are still trying to understand the best things to do in the classroom to help close gender gaps. Nevertheless, gender differences in mathematics have been and will be one of the important topics because it is an important identification of item bias. An item is considered bias if it functions differently for a specified subgroup of test-takers (such as boy and girl). Unlike in DIF, item bias may be resulting in a situation where equally able students do not have an equal chance of success in a particular item. Bias test item may contain sources of difficulty that are not relevant to the construct being measured that impact test-takers' performance. As such, DIF studies need to be carried out for each testing made to ensure that the information obtained from the test gives meaning to the group being tested.

REFERENCES

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

Ameer, I. S., & Singh, P. (2013). Exploring grade levels and gender differences in numeracy thinking among secondary school students. *Procedia-Social and Behavioral Sciences*, *90*, 187–195.

Bezzina, F. H. (2010). Investigating gender differences in mathematics performance and in self-regulated learning: An empirical study from Malta. *Equality, Diversity, and Inclusion: An International Journal*, *29*(7), 669–693. https://doi.org/10.1108/02610151011074407

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: L. Erlbaum. https://doi.org/10.4324/9781315814698

Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Non-cognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human Resources*, *48*(1), 237–264. https://doi.org/10.1353/jhr.2013.0002

Curriculum Development Division. (2016). *Dokumen standard kurikulum dan pentaksiran matematik tingkatan 1.* Putrajaya: Kementerian Pendidikan Malaysia.

Crombie, G., Sinclair, N., Silverthorn, N., Byrne, B., DuBois, D., & Trinneer, A. (2005). Predictors of young adolescents' math grades and course enrollment intentions: Gender similarities and differences. *Sex Roles, 52*, 351–367. https://doi.org/10.1007/s11199-005-2678-1

Cvencek, D., Meltzoff, A. N., & Greenwald, A. G. (2011). Math-gender stereotypes in elementary school children. *Child Development*, *82*, 766–779. https://doi.org/10.1111/j.1467-8624.2010.01529.x

Dickhauser, O., & Wulf-Uwe, M. (2006). Gender differences in young children's math ability attributions. *Psychology Science*, *48*(1), 3–16.

Eakman, A. M. (2012). Measurement characteristics of the engagement in meaningful activities survey in an age-diverse sample. *American Journal of Occupational Therapy, 66*(2), e20–e29. https://doi.org/10.5014/ajot.2012.001867

Forgasz, H. J., Becker, J. R., Lee, K.-H., & Steinthorsdottir, O. B. (Eds.). (2010). *International perspectives on gender and mathematics education.* Charlotte, NC: Information Age Publishing.

Frenzel, A. C., Pekrun, R., & Goetz, T. (2007). Girls and mathematics – A "hopeless" issue? A control value approach to gender differences in emotions towards mathematics. *European Journal of Psychology of Education*, *22*, 497–514. https://doi.org/10.1007/BF03173468

Fryer, R., & Levitt, S. (2010). An empirical analysis of the gender gap in mathematics. *American Economic Journal: Applied Economics*, *2*(2), 210–240. https://doi.org/10.1257/app.2.2.210

Hanizah Hamzah. (2007). Kebezaan kefungsian item berkaitan gender (gender related differential item functioning) dalam peperiksaan matematik peringkat kebangsaan: kewujudan dan perkaitan dengan jenis Item. Unpublished doctoral dissertation, Universiti Kebangsaan Malaysia.

Hattie, J. (2009). Visible learning. A synthesis of over 800 meta-analyses relating to achievement. In J. A. Banks (Ed.), *Encyclopedia of diversity in education* (pp. 1455–1460). Thousand Oaks, CA: Sage. https://doi.org/10.4324/9780203887332

Husin, M. (2014). *Assessing mathematical competence in second language: Exploring DIF evidences from PISA Malaysian data*. Unpublished Master's thesis, The University of Wisconsin – Milwaukee.

Innabi, H., & Dodeen, H. (2018). Gender differences in mathematics achievement in Jordan: A differential item functioning analysis of the 2015 TIMSS. *School Science and Mathematics, 118*(3–4), 127–137. https://doi.org/10.1111/ssm.12269

Isiksal, M., & Cakiroghi, E. (2010). Gender differences regarding mathematics achievement: The case of Turkish middle school students. *School Science and Mathematics*, *108*(3), 113–120.

Ismail, N. A., & Awang, H. (2009). Mathematics achievement among Malaysian students: What can they learn from Singapore? *International Education Studies*, *2*(1), 8–17. https://doi.org/10.5539/ies.v2n1p8

Izazol Idris. (2012). *Kebezaan kefungsian item berkaitan jantina (GDIF) dalam instrumen TIPS II*. Unpublished Master's thesis, Universiti Kebangsaan Malaysia.

Jordan, N. C., Kaplan, D., Oláh, L., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77*(1), 153–175. https://doi.org/10.1111/j.1467-8624.2006.00862.x

Kanageswary, S. S. S., & Ong, S. L. (2014). The validity of administering bilingual mathematics test among Malaysian bilingual students using differential item functioning (DIF). *Asia Pacific Journal of Educators and Education*, *29*, 1–18.

Klein, P. S., Adi-Japha, E., & Hakak-Benizri, S. (2010). Mathematical thinking of kindergarten boys and girls: Similar achievement, different contributing processes. *Education Studies in Mathematics*, *73*, 233–246. https://doi.org/10.1007/s10649-009-9216-y

Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review, 77*(3), 463–485. https://doi.org/10.1177/0003122412440802

Lim, H. L., & Subashini, S. (2015). Mengesan keberbezaan fungsi item bagi item ujian pencapaian Prinsip Perakaunan tingkatan empat di Negeri Perak. *Jurnal Pemikir Pendidikan, 6*, 49–66.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Lyons-Thomas, J., Sandilands, D., & Ercikan, K. (2014). Gender differential item functioning in Mathematics in four international jurisdictions. *Educational and Science*, *39*(172), 20–32.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education 19*(4), 289–304. https://doi.org/10.1207/s15324818ame1904_4

Ministry of Education. (2016a). *Laporan TIMSS 2015*. Putrajaya: Ministry of Education.

Ministry of Education. (2016b). *Standard based curriculum for secondary school*. Putrajaya: Ministry of Education.

Pae, T., & Park, G. P. (2006). Examining the relationship between differential item functioning and differential test functioning. *Language Testing, 23*(4), 475–496.

Rogers, L. (2003). Gender differences in approaches to studying for GCSE among Year 10 pupils. *The Psychology of Education Review*, *27*(1), 18–27.

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist*, *60*(9), 950–958. https://doi.org/10.1037/0003-066X.60.9.950

Spilt, J. L., Koomen, M. Y., & Jak, S. (2012). Are boys better off with boy and girls with girl teachers? *Journal of School Psychology*, *50*, 363–378. https://doi.org/10.1016/j.jsp.2011.12.002

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*(3), 246–280. https://doi.org/10.1080/08957347.2012.687650

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Zhang, M., & French, B. F. (2010). Gender related differential item functioning in mathematics tests: A meta-analysis. Paper presented at the National Council on Measurement in Education Conference, May 2010, Denver, CO.

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