



# Content Analysis of Gender Differential Item Functioning of Mathematics Items among Secondary Students in an Eastern Chinese Culture

S.Kanageswari Suppiah Shanmugam<sup>1</sup>, Siti Noor Ismail<sup>2</sup>  
and Arsaythamby Veloo<sup>3</sup>

<sup>1</sup>School of Education (SoE) & Centre of Testing, Measurement and Appraisal (CeTMA),  
Universiti Utara Malaysia, Sintok, Malaysia

<sup>2,3</sup>School of Education (SoE), Universiti Utara Malaysia, Sintok, Malaysia  
kanageswari@uum.edu.my<sup>1</sup>siti.noor@uum.edu.my<sup>2</sup>arsay@uum.edu.  
my<sup>3</sup>

**Abstract.** This study reports on preliminary findings of gender differential item functioning (DIF) in an Eastern school culture, which is known for ‘special’ teaching approaches and successful mathematics learning that produces impressive mathematics results. This study aimed to identify mathematics items that function differently across gender groups in coeducational schools and examine the characteristics of the DIF mathematics items. A total of 63 boys and 55 girls in Grade Eight were selected from an Eastern school culture for the preliminary study. The software WINSTEPS was used to conduct DIF analysis. Items were flagged for DIF by using Mantel-Haenszel chi-square method and Welch t statistics with boys forming the reference group and girls forming the focal group. Some 12 computation and 12-word problem items from the grade eight. Word problem items were distinguished as items that are set in real-world context. Findings revealed that both methods flagged three common DIF items, of which one was a computation and two were word problem items. The Welch t-test flagged an additional computation item, which was undetected by the Mantel-Haenszel chi-square method. The computation item exhibited moderate DIF and favoured girls, while the word problem items exhibited large DIF, of which one item each favoured each gender group. These DIF items were from the Number domain from the topics involving decimal number, measurement, and percentage. While the computation item assessed the lower order thinking skills in the cognitive domain of Knowing, the word problem items assessed the cognitive domain of Applying. This initial exploration suggests that items that have more language load is more likely to function differently between the gender groups and that items that assess higher-order thinking skills favour boys, while items that assess lower-order thinking skills such as knowing and solving routine questions favour girls.

**Keywords:** Differential Item Functioning, Gender, Mathematics, Mantel-Haenszel chi-square method, Welch t-test

© The Author(s) 2024

Q. Zhang (ed.), *Proceedings of the Pacific-Rim Objective Measurement Symposium (PROMS 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 23,  
[https://doi.org/10.2991/978-94-6463-494-5\\_10](https://doi.org/10.2991/978-94-6463-494-5_10)

## 1 Introduction

Differential Item Functioning (DIF) occurs “when an item’s properties in one group are different from the item’s properties in another group” [1, p. 331]. In other words, DIF is the result of when items behave differently for different groups of students with similar abilities after controlling for their proficiency and is detected when the item displays different statistical properties [2]. Items are tagged as functioning differently when the probability of answering correctly on those items is unequal for different groups of students with equal ability due to group membership unrelated to the construct under measure such as age, race, or locality. In contrast, non-DIF items are items that different groups of students of the same ability have equal probability of answering correctly, regardless of their group membership. When differences exist in group settings of individuals, which are unrelated to the test proficiency, those DIF items warrant further investigation to determine the source of the items behaving differently for the reference and focal groups [3]. Some of the sources of DIF could be linguistics characteristics, the test item, or the true differences in the individuals’ ability. The former results in bias while the latter is due to impact. An item is tagged to be biased to the group that it disfavours based on the judgement made from the DIF analyses.

Word problem mathematics items have been found to be more challenging when compared to computation items, even among students who are highly proficient in mathematics [4], due to the language load of the word problem items. As [5] reiterated, students who are competent arithmetically, demonstrate different unequal competencies in solving computation items when compared to solving word problem items. They elaborated that the difference is not an indication of mathematical proficiency but rather due to the linguistics of the word problem items that is absent in the computation items. The linguistically denser word problem items augment the challenge students face in understanding the textual information and the contextual information, which elevate their linguistics challenge in solving the word problem items [4]. Accordingly, [6] reiterated that one of students’ difficulty is in order to filter the extraneous textual information to understand the mathematics items before selecting the correct problem-solving strategies and algorithm.

An interesting find from mathematics-related international assessments such as the Trends in International Mathematics and Science Study (TIMSS) is that since TIMSS 2003 East Asian countries have consistently recorded top performance in grades four and eight mathematics [7]. Although there are a host of variables that influence student mathematical achievement, culture is of particular interest and forms the focus of this paper. This is because there appears to be a demarcation between the culture of teaching and learning mathematics in the Eastern countries, marked by the “Chinese or Confusion tradition” [8, p. 21], and the Western mathematics adopted in the curriculum of most countries in the world. Even more profound is that within the context of multiple cultures co-existing in a country, there is the prevalent issue of gender differential performance among the different cultures adopting the same national curriculum. The striking element appears to be the procedural teaching that emphasises repeated ‘drill’ practices as a learning approach among students [8]. In

addition, results from the four most recent past cycles of TIMSS grade eight mathematics suggest girls' better performance in some countries [7]. Therefore, this paper examines the characteristics of the DIF mathematics items by item-type that indicate gender differential performance by using the Welch t-test and the Mantel-Haenszel chi-square method.

### **1.1 Problem Statement**

The international daily, the Telegraph, highlighted the fact that TIMSS 2015 Mathematics reported on the high standard of East Asian countries maintaining their 20 years lead for pupils aged 10 and 14 [10]. One of the many factors that have been studied to explain the superior performance among students is school culture [9]. [9] defines school culture as 'a system of shared knowledge, practices, beliefs, and values about mathematics learning' (p. 111), which is observed by the school community comprising students, teachers, and administrators. She further highlighted that school culture is significant in shaping the mathematics teaching and learning process as the whole school community stay connected through a common goal-sharing that promotes successful mathematics learning. In particular, the school culture of Chinese schools emphasises "lively learning atmosphere in class", "plenty of drills and practices", "more homework", "more tuition" as well as "more competition and quizzes" [9, p. 119]. Accordingly, [8] too shares similar sentiments about the Chinese or Confusion tradition school that encourage procedural teaching since procedural learning involves repeated practices, and therefore students are more able to understand better. However, he is quick to caution on the misconstrued understanding of procedural teaching as being rote learning, which is not the same as repeated exercises.

In some countries such as Malaysia, addressing gender differential performance has become a national agenda since the issue of "lost boys who either leave school early or with low attainment levels" have emerged and is a cause of concern [14, p. 7]. Of interest is to examine language as a source of DIF across gender groups for computation and word problem mathematics items. When mathematics test items introduce construct-irrelevant variance due to linguistic complexity, it is vital to examine the items characteristics in an Eastern school culture that promotes successful mathematics learning as evidenced by commendable results. However, to date, gender DIF mathematics items by item-type in an Eastern school culture known to be excelling in mathematics is limited. Thus, this paper attempts to identify gender DIF items and examine the characteristics of the flagged DIF mathematics items.

### **1.2 Research Aim**

The research aimed at examining the characteristics of DIF items between boys and girls. Accordingly, the main purpose of this study is to identify mathematics items that function differently across gender groups in coeducational schools and examine the characteristics of the DIF items that favour certain gender groups. In doing so, the

initial analyses were directed towards examining the extent the data fit the Rasch model. The research objectives are:

1. To identify DIF items that function differently between boys and girls.
2. To examine the characteristics of the flagged DIF items.

The research questions are:

1. Which items signal negligible, moderate, and adverse DIF that favour boys and girls?
2. What are the characteristics of the items that are flagged as DIF?

## **2 Literature Review**

### **2.1 Gender DIF Mathematics Items**

DIF studies have long been conducted to understand items that favour boys and girls. In examining DIF items by item-type, several perspectives can be explored such as multiple choice versus constructed response items [15], test format [16], and computation versus word problem items [17]. Real-world problem items, items from the domain Geometry and, assess spatial and deductive abilities favour boys [19]. On the other hand, items from the domain of Algebra and assess numerical ability [19] tended to favour girls. Computation and lower-order thinking questions tend to favour girls, while unconventional problem-solving strategies involving higher-order thinking skills favour boys.

As [11] clarified, girls tend to more likely to replicate the problem-solving procedures learnt in the classroom when compared to boys and therefore, they are better in conventional, computation and lower order thinking questions. Boys, on the other hand, tend to experiment with non-routine strategies and therefore, they favour unconventional, problem solving and higher order thinking skills items. In another later study, [12] further discovered that routine questions using mathematical algorithm favour girls while non-routine problem solving that uses logical thinking favour boys. Apart from that, [20] who studied the serial position of the mathematics items in a test found that when items have been arranged hierarchically according to their difficulty, items placed at the beginning of the test favour girls as they perform better on easy items located at the beginning when compared to more difficult items positioned at the test end.

### **2.2 Computation and Word Problem Mathematics Items**

[21] differentiated computation items from word problems by the absence of real-world setting. Since computation items can be in the form of a direct question involving a combination of either addition, subtraction, multiplication, or division or with the use of minimal language, numbers and operational symbols, a non-contextualised setting is necessary [21]. Therefore, computation items involve direct

algorithm, manipulate numbers and variables, and may have simple language but they do not carry real-world setting.

On the contrary, word problem items are regarded to be more difficult for students than computation items [22], due to the challenges posed in firstly understanding the textual information and translating it into the mathematical algorithm, before selecting a suitable problem-solving strategy [6]. In comparison to computation items, word problem items contain heavier language load and involve multiple problem-solving steps, in addition to non-linear process of converting the textual information into the abstract problem [23]. Word problem items resemble ‘real world like’ problems set in a context, which is discern in computation items.

### 3 Theoretical Framework

#### 3.1 The Rasch Model

The Rasch Model places all test items on a common scale alongside the person’s ability on that latent trait and is based on the relationship between the probability of answering an item correctly and the person’s ability. It has only the  $b$  parameter and the mathematical function is as exhibited below [24]:

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

In this model, for any item  $i$ ,  $P_i(\theta)$  is the probability of an examinee answering correctly on item  $i$  at  $\theta$  ability and  $b_i$  refers to the difficulty parameter for item  $i$ . In this study, DIF analysis will be conducted based on the Welch t-test and the Mantel-Haenszel (M-H) chi-square method obtained through DIF effect size. Using IRT, DIF is detected through t-test between-group differences in item parameters [25]. WINSTEPS flags DIF items using the M-H approach and the Welsch t-test. Mantel-Haenszel Chi-square method and Welch t-test produce similar results if the data fits the Rasch model [26]. Mantel-Haenszel Chi-square method is more preferred than Welch t-test in WINSTEPS since it is more accurate due to its robustness to missing data [27], and it is advocated by Education Testing Service. Therefore, this study has attempted to use Mantel-Haenszel chi-square, in comparison with Welch t-test.

#### 3.2 Word Problem Model

The Word Problem Model was proposed by [23] whereby word problem items are seen as having two distinct sets of structures in the form of the textual information, which develops into the abstract mathematics model. The textual information of the word problem comprises two levels. First is the text from the textual input and the second is the problem model that contains the relevant information necessary to successfully solve the abstract problem model. Accordingly, the mathematical

problem model is embodied by the textual information, which is complemented by the problem-solving strategies. Therefore, the Word Problem Model [23] distinguishes three different types of knowledge for the successful solution of word problem items. They are the knowledge involving the:

1. text model that translates the mathematical text into propositions.
2. situation model that complements the text with the inferences derived from the students' real world.
3. problem model that strategizes the relevant mathematical skills and operations for successful problem solving.

Therefore, this model postulates that solving word problems involve the input of students' knowledge of the mathematical language, building relations among the quantities involved and selecting suitable mathematical algorithm to solve the problems.

## **4 Methodology**

### **4.1 Sample**

This study is a comparative research study, where boys form the reference group, and girls form the focal group. A total of 118 grade eight students were selected, with 63 boys and 55 girls. Even though the sample appears to be small, [28] highlighted that regardless of sample size, there will be acceptable accurate measurement of student performance.

### **4.2 Instrument**

A total of 24 multiple choice mathematics items, with equal number of computation and word problem items were selected from TIMSS grade 8 released items for the cycles since 1999. Word problem items (W) were distinguished from computation items (C) as items that were set in a real-life [21]. The items were arranged according to the sub-concepts in the mathematics curriculum. The layout of the test booklet contained student particulars (gender and race) in Section A and the test items in Section B.

### **4.3 Procedure**

The test booklets were administered to the students according to the routine practices of a school examination with the help of the class teachers. In addressing the validity of the test scores, certain measures were observed, which include providing the teacher a set of standard instructions to ensure a uniform test administration. The students' table were arranged spaciouly adequate to discourage any form of unethical behaviour. The students were given five minutes to fill in the particulars required in section A and were reminded to show their working as a measure to address unethical

exam practices, in addition to obtaining a better insight of their mathematical understanding. They were given one hour to answer the test items during which the class teacher invigilated, and the researcher monitored to ensure no malpractice occurred during the test administration. Calculators were not allowed as the purpose of the test is to assess student's mathematical proficiency and not their skills in using calculators.

#### 4.4 Data Analyses

The options selected by the students for each item were keyed into the Excel Worksheet before analysing using WINSTEPS version 3.67.0 [29]. The analyses conducted include determining the item mean-square (MNSQ) infit and outfit indices and person measures to determine data fit and predictability respectively. The infit and outfit mean-square indices allow a check on the extent the data fit the Rasch model by examining the magnitude of the departure. The infit mean-square is affected by the pattern of examinees' responses to the test items, while the outfit mean-square is influenced by the examinees' responses to items that are very difficult or very easy [30].

After examining the fit of the model, DIF analyses was conducted to flag DIF items by using the Welch t-test and Mantel-Haenszel Chi-square method. Although [31] highlighted that a minimum sample size of 100 is admissible for focal and reference groups for DIF analysis, [28] clarified that there is still an acceptable accurate measurement of student performance regardless of sample size.

To examine for DIF using the Welch t-test, the probability value needs to be less than 0.05. As for the Mantel-Haenszel chi-square method, [29] states the Mantel-Haenszel chi-square probability value needs to be small enough to eliminate the possibility of DIF occurring due to chances and yet, the DIF size needs to be large enough to conclude its substantive impact on the test scores. Thus, items are flagged as DIF if the Mantel-Haenszel probability value is less than 0.05 before it is classified as negligible, moderate, or large DIF based on criteria proposed by [32].

- 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 favours the focal group (girls) while negative Mantel-Haenszel size favours the reference group (boys) [29]. Non-DIF items will function similarly for both the reference and focal groups.

## 5 Results

The data were first examined to determine their fit to the Rasch model by determining the infit and outfit mean square (MNSQ) values for the item measure and person measure. The findings are arranged according to the two research questions posed at the beginning of this paper.

**5.1 Item Measure**

The 24 mathematics items were analysed to examine the extent to which the data fit the Rasch model using the mean-square values of the infit and outfit for the non-extreme items.

**Table 1.**Summary of non-Extreme Mathematics Items for Pilot Test.

	Model		Infit		Outfit		
	Measure	Error	mnsq	zstd	mnsq	Zstd	
M	0.00	0.27	0.99	0.00	1.04	0.20	
SD	1.17	0.08	0.09	0.70	0.21	0.90	
max.	2.22	0.60	1.19	1.10	1.44	2.10	
min.	-2.95	0.22	0.84	-1.70	0.69	-1.60	
real rmse	0.28	adj.sd	1.14	separation	3.99	item reliability	0.94
model rmse	0.28	adj.sd	1.14	separation	4.07	item reliability	0.94
S.E. of item mean = 0.24							
item raw score-to-measure correlation = -0.97							

Table 1 exhibits that the infit and outfit mean square values were within the range of 0.8 and 1.2, suggesting that they are acceptable for high stakes multiple choice items [30]. The standardized z-score of 0.2 did not indicate over predictability or under predictability as it was neither below -2 nor above 2 and within the acceptable range of -1.9 to 1.9, suggesting reasonable predictability for the data. The raw score-to-measure correlation of -0.97 approximated to the recommended value of -1 [29], while the item reliability of 0.94 suggests a high reliability [33]. These indices suggest that the items were productive, did not degrade the measurement, fit the model, and demonstrated reasonable prediction. In addition, the test was also found to be unidimensional and thus, further Rasch model analyses could be carried out. The detailed item statistics for each of the 24 items is displayed in Table 2.

**Table 2.**Item Statistics for Mathematics Test Items.

Item	Infit		Outfit		PT-Measure	
	Measure	S.E.	Measure	S.E.	Correlation	Expected
C1	1.01	0.2	1.31	0.6	0.13	0.16
C2	1.04	0.5	0.50	0.5	0.48	0.50
W3	1.10	0.5	1.14	0.4	0.22	0.28
W4	0.86	-0.4	0.75	-0.3	0.32	0.25
C5	1.19	1.1	1.44	1.3	0.23	0.34
C6	0.89	-1.0	0.83	-1.0	0.50	0.43
C7	1.11	0.7	1.12	0.5	0.31	0.37
C8	0.94	-0.6	0.94	-0.3	0.47	0.43



W9	0.89	-0.8	0.84	-0.6	0.45	0.39
W10	0.88	-0.7	0.69	-1.1	0.44	0.36
W11	0.91	-0.8	0.83	-0.8	0.47	0.41
W12	0.87	-1.1	0.89	-0.5	0.48	0.41
C13	1.06	0.4	1.20	0.7	0.30	0.34
C14	1.09	0.5	1.31	0.9	0.27	0.33
C15	1.07	0.8	1.33	2.1	0.47	0.52
W16	1.10	1.0	1.17	1.1	0.39	0.44
W17	0.93	-0.5	0.82	-0.7	0.44	0.39
W18	0.84	-1.7	0.77	-1.6	0.54	0.45
C19	0.99	-0.1	0.95	-0.4	0.51	0.49
W20	0.95	-0.4	0.97	-0.1	0.47	0.44
C21	1.08	0.8	1.27	1.6	0.38	0.43
C22	1.01	0.1	1.27	1.3	0.54	0.53
W23	1.01	0.1	1.27	1.3	0.54	0.53
W24	0.98	-0.2	0.97	-0.2	0.53	0.51
Mean	0.99	0.0	1.04	0.2		
S.D.	0.09	0.7	0.21	0.9		

As exhibited in Table 2, the point-biserial correlation indices (represented as PT-Measure) were all positive values ranging from 0.13 to 0.54 for all the items. The positive value indicates that a higher proportion of high ability students answered correctly when compared to the lower ability group, which is desired and expected if the item is not flawed. Therefore, these indices indicate that the items were not flawed.

## 5.2 Person Measure

Analysis for person measure was also performed for the non-extreme scores among 107 students. This is because from the total 118 students, 17 students answered correctly for all the 24 items and were considered as having extreme scores. Table 3 exhibits the analysis for the non-extreme person measure.

**Table 3.** Summary of non-Extreme Person Measure for Pilot Test.

	Model		Infit		Outfit	
	Measur e	Error	mnsq	zstd	mnsq	Zstd
M	1.16	0.55	0.99	0.00	1.04	0.10
SD	1.09	0.11	0.20	0.90	0.65	0.90

max.	3.68	1.05	1.60	2.80	5.08	3.60
min.	-1.66	0.46	0.63	-2.30	0.31	-1.80
real rmse	0.58	adj.sd	0.92	separation	1.59	person reliability 0.72
model rmse	0.56	adj.sd	0.94	separation	1.68	person reliability 0.74
S.E. of person mean = 0.11						

The person reliability was a high 0.74 for the non-extreme scores. The fit statistics for person measure was not detailed out as the evidence of item measure is more critical to determine whether data fits the model. The responses from the student sample agreed that the data was reliable with reasonable predictability as desired. The WINSTEPS output for the extreme and non-extreme items and person measures is provided as Appendix A.

### 5.3 Person to Item Map

In addition to determining the mean-square values for the items, the item distribution was also examined by using the person to item map as illustrated in Figure 1. The items have been arranged according to its difficulty. The easiest item is at the bottom, while the most difficult item is at the top.

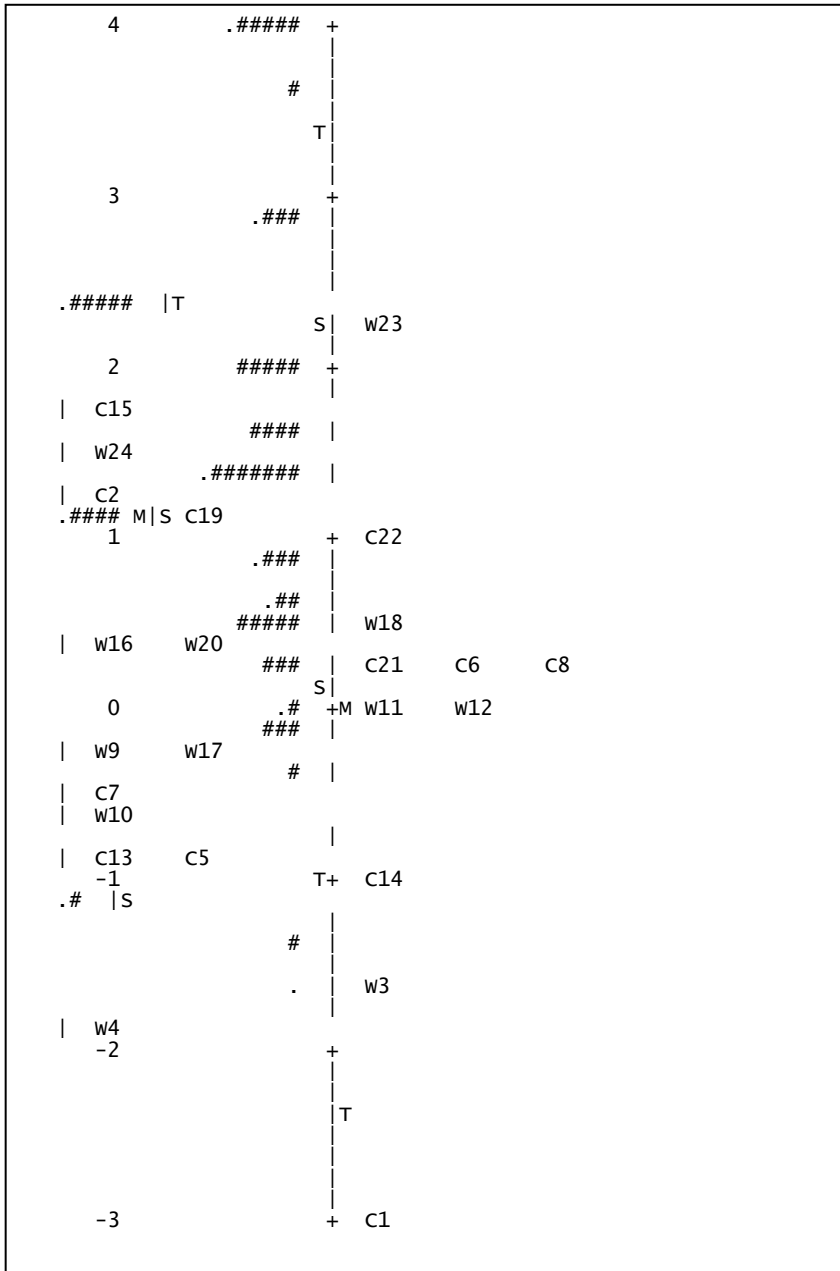


Fig. 1. Person to item map.

As illustrated in Figure 1, Item C1 was the easiest item while Item W23 was the most difficult. As anticipated, the easiest item was a computation item, and the most difficult item was a word problem item that has more language load. Another

interesting finding is Item W4, which was the second easiest item for this student sample is categorised in the TIMSS 1999 Grade Eight Mathematics report as an item that can only be answered by students at the top 10% of international benchmark [34]. The item distribution can be further improved by adding more challenging items. This is because there is a gap among the items for students at the higher ability group, with the absence of items to measure students at the ability  $\theta$  range of  $3 \leq \theta \leq 4$ . These results further strengthen the possibility of the sampled students habiting a school culture that exhibits successful mathematical learning.

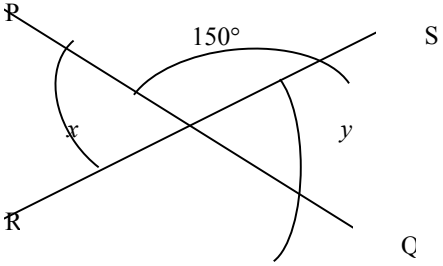
**5.4 Research Question 1: Which Items Signal Negligible, Moderate and Adverse DIF that Favour Boys or Girls?**

There were two computation items, items C13 and C21 and, two-word problem items, items W18 and W20 that have the Mantel-Haenszel probability of less than 0.05. All the items except one, Item C21 have Mantel-Haenszel chi-square value of less than 0.43, revealing that this one item displayed negligible DIF. Thus, three common items were flagged as exhibiting DIF by the Mantel-Haenszel chi-square method and the Welch t-test, of which one is a computation item (C13) and two are word problem items (W18 and W20) as exhibited in Table 4. However, the Welch t-test flagged an additional computation item (C7) as DIF, which was left undetected by the Mantel-Haenszel chi-square method.

For the three common DIF items, the computation item signalled moderate DIF, while the two-word problem items recorded large DIF. The positive value for the Mantel-Haenszel size indicates that the item favoured the focal group, which is the girls while the negative value for the Mantel-Haenszel size indicates that the item favoured the reference group, which is the boys. Therefore, the one moderate DIF computation item (C13) and one large word problem item (W20) favoured girls. Another large DIF word problem item (W18) favoured the boys.

**Table 4.**DIF Items.

Item	Item Description	Mantel-Haenszel Prob	DIF Size	DIF Type	Welch t statistics	Favour
C7	What is the value of $\frac{4}{5} - \frac{1}{3} - \frac{1}{15}$ ?	0.1704	0.32	-	0.0386	Girls
	A $\frac{1}{5}$ B $\frac{2}{5}$					
	C $\frac{7}{15}$ D $\frac{3}{4}$					
	E $\frac{4}{5}$					

C13	The total weight of a pile of 500 salt crystals is 6.5 g. What is the average weight of a salt crystal? A 0.0078 g C 0.0325 g	B 0.013 D 0.078 g	0.0429	0.61	Moderate	0.0422	Girls
W18	A shop increased its prices by 20%. What is the new price of an item which previously sold for RM 800? A RM 640 C RM 960	B RM9 00 D RM 1,000	0.0011	-1.10	Large	0.0076	Boys
W20	The total weight of a pile of 500 salt crystals is 6.5 g. What is the average weight of a salt crystal? A 0.0078 g C 0.0325 g	B 0.013 g D 0.078 g	0.0395	1.48	Large	0.0347	Girls
C21			0.0335	0.04	Negligible	0.0225	Girls
<p>In the figure, PO and RS are two intersecting straight lines. Which is the value of <math>x+y</math>?</p> <p>A <math>15^\circ</math>      B <math>30^\circ</math>      C <math>60^\circ</math></p> <p>D <math>180^\circ</math>      E <math>300^\circ</math></p>							

Another interesting highlight of this study is that when studying the position of the items in the person to item map (Figure 1), items that are easier, which appear at the bottom tend to favour girls, while difficult item at the topmost favour boys. In summarising, the one moderate DIF computation item and one large DIF word problem item favoured the girls and, one large DIF word problem item favoured the boys. The DIF person plot for all the 24 items is illustrated in Figure 2 and the detailed output is as shown in Appendix B.

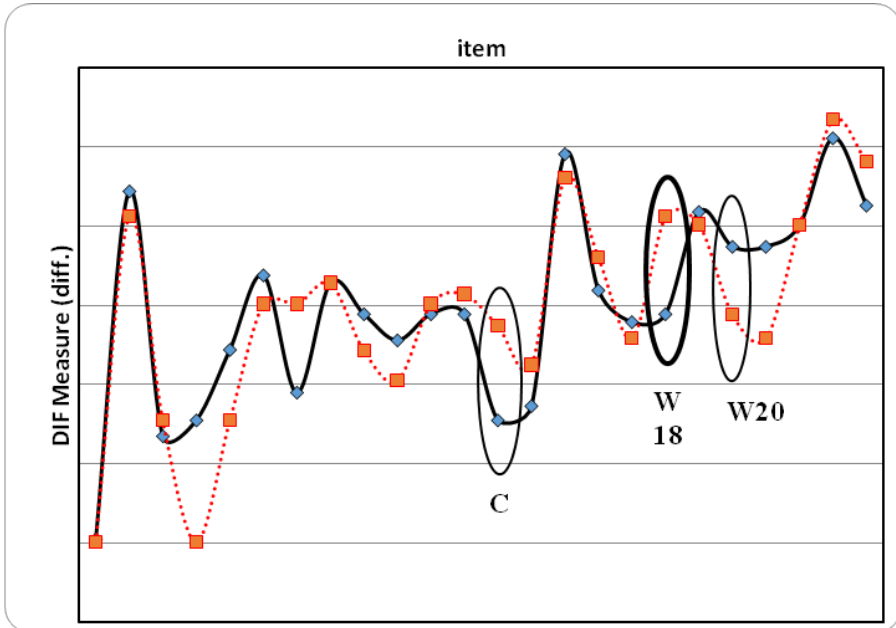


Fig. 2. Person DIF plot

**5.5 Research Question 2: What are the Characteristics of the DIF Items?**

Three items were flagged as exhibiting DIF. One moderate DIF, while another two-word problem items signalled large DIF. As can be seen from Table 5, the moderate DIF item is a computation item from the topic Decimals and assesses students’ ability to arrange the decimal numbers in an ascending order to determine the smallest decimal number. From the perspective of TIMSS, this item is from the content domain of Fractions and Number Sense and assesses the lower order thinking skills from the cognitive domain of Knowing.

Table 5. Moderate DIF Items.

Item	Topic	Learning Outcome	TIMSS Content Domain	TIMSS Cognitive Domain
C13	Decimals	Arrange decimals in order	Fractions and Number Sense	Knowing

As can be seen from Table 6, both large DIF items were word problem items and from the broad content domain of Number. Specifically, the items were from the topic Measurement and Percentage, and assessed the cognitive skills of using complex procedures and solving routine problems. The limited number of three DIF items seem to suggest that the item that assessed lower-order thinking skills of Knowing and

Solving Routine Problem favoured girls, while items that assessed the higher-order thinking skills of Using Complex Procedures favoured boys.

**Table 6.** Large DIF Items.

Item	Topic	Learning Outcome	TIMSS Content Domain	TIMSS Cognitive Domain
W18	Percentage	Solve problems involving percentage	Ratio, Proportions, and Percent	Solving Routine Problem
W20	Basic Measurement	Use four operations to solve problems involving mass	Fractions and Number Sense	Using Complex Procedures

## 6 Discussion

The initial findings of this study indicate that the data fits the Rasch model and, the Welch *t*-test detected an additional computation DIF item that was not flagged by the Mantel-Haenszel chi-square method. This item is from the content domain of Number and assesses subtraction of three proper fractional numbers of different denominators. Both methods of Mantel-Haenszel chi-square method and the Welch *t*-test flagged three common DIF items, which are one moderate-DIF computation item and two large-DIF word problem items. The flagged DIF items are from the TIMSS content domain of *Number* and specifically from the topic *Decimals*, *Percentage* and *Measurement*. However, the one computation item and one word problem item that favour girls assess the cognitive domain of lower-order thinking skills of *Knowing* and *Applying Routine Questions*, while the one-word problem item that assesses the higher-order thinking skills of using *Complex Procedures* favours the boys. This finding supports that boys are more able to solve higher order thinking skills items than girls [35]. The item distribution in the person to item map seems to suggest that items positioned at the bottom that are easier favour girls, while difficult item at the top favour boys.

The large DIF for the word problem item signals that language load tends to elevate the likelihood of an item functioning differently between the two gender groups, even though non-uniform DIF is detected. As [4] clarified word problem items are more demanding on students when compared to computation items, due to the heavier language load introduced by the linguistic of the word problem items. The language structure of the word problem items escalates the challenge faced by the students in comprehending the textual information as posited in the Word Problem Model [23]. When compared to computation items, students must unravel a deeper layer of filtering the extraneous information in the textual model before they can proceed to the situational model for the successful solution in the abstract model. In examining further the linguistics structure of the items, Item C13 (*Which of these is the smallest number?*), which favours girls is a WH question that has a simple basic structure of *wh-* + *an auxiliary verb (be, do or have)* + *subject* + *main verb*, with an

auxiliary verb (is) preceded by a subject and a verb [36]. It is also a one-step problem that requires only one step of arranging the numbers in an ascending order.

Similarly, Item W20 (*The total weight of a pile of 500 salt crystals is 6.5 g. What is the average weight of a salt crystal?*), which favours girls also has a basic linguistic structure. The item is phrased with one simple sentence that has a subject, verb, and object, and one WH question that has an auxiliary verb, subject and a main verb. This word problem is also a one-step problem that requires students to divide a decimal number (6.5) with a whole number (500). In contrast, Item W18 (*A shop increased its prices by 20%. What is the new price of an item which previously sold for RM 800?*), which favours boys does not have a straightforward sentence structure. The item uses multiple WH questions that has been found to create acute learning difficulty among students. In addition, this word problem item is also not a one-step problem as students need to determine the amount money for the 20% before determining the new price. Within the limitation of detecting three DIF items, the findings of this study to a certain extent suggest that items with simple linguistic structure and that assess lower order thinking skills favour girls over boys. Perhaps rewording and rearranging the sentences to “*An item was priced at RM 800. The shop increased the prices of all the items by 20%. What is the new price of that item after the increase?*” or to “*A shop increased its prices by 20%. If the old price is RM 800, determine the new price of that item.*” could reduce the linguistic challenge. This perspective of linguistic simplifications of mathematics items involving non-native English speakers in a mathematically superior school culture can be an avenue worth exploring in future studies.

As [20] elaborated, less complex mathematics items favour girls over boys as was also discovered in this study. Since girls have been found to more likely to follow strictly the problem-solving steps learnt in the classroom when compared to their counterpart [11], probably the items (C13 and W18) that assess lower-order thinking skills favoured the girls in this study and item that assessed higher-order thinking skills (W20) favoured the boys. A possible explanation is that girls prefer to adopt the learnt problem-solving strategies in their classrooms and less inclined to experiment with short cuts, which likely explains girls performing better in routine, computation and lower order thinking questions. On the contrary, non-routine questions that require problem solving and, higher order thinking skills favour boys [11], which possibly explains why items that require complex procedures favour boys. In another study, [12] elaborated the sources of DIF from the perspective of item-type and problem-solving strategy. Just like the findings of this study, they discovered that girls favour routine questions when compared to non-routine items favouring boys.

This initial exploration suggests that items that have more language load are more likely to function differently between the gender groups, of which items with simplified linguistics structure favour girls. Items that assess higher-order thinking skills tend to favour boys, while items that assess lower-order thinking skills such as *knowing* and *solving routine questions* favour girls. Complex sentence structures such as multiple WH questions are biased against girls, and they find these items more difficult when compared to boys. The findings of this study also shed some light on the possibility that the item difficulty based on the Rasch measurement and not



according to the serial position of test items should be considered as a source of bias. Future research should explore on this as a source of bias. With only a limited three DIF items detected in a pool of 24 items, more exploration is needed to conclude that mathematics items functioned differently between boys and girls in a school culture renowned for successful mathematics learning, even though the linguistics properties of test language as a source of bias cannot be ignored.

**Disclosure of Interests.**The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Furr, M. R., Bacharach, V. R.: *Psychometrics: An introduction*. SAGE, Thousand Oaks, CA (2007)
2. Angoff, W.H.: Perspectives on differential item functioning methodology. In: Holland P. W., Wainer, H. (eds.), *Differential item functioning*, pp. 3-23. Lawrence Erlbaum, Hillsdale, NJ (1993)
3. Dodeen, H., Johanson, G. A.: An analysis of sex-related differential item functioning in attitude assessment. *Assessment & Evaluation in Higher Education* **28**(2), 129-134 (2003)
4. Oviedo, G. C. B.: Comprehending algebra word problems in the first and second languages. In: Cohen J., McAlister K. T., Rolstad K., Macswan, J. (eds.), *Proceedings of the 4th International Symposium on Bilingualism*, pp. 267-295. Cascadia Press, Arizona State University (2005)
5. Valentin, J.D., Lim, C.S.: Roles of semantic structure of arithmetic word problems on pupil's ability to identify the correct operation, <http://www.cimt.plymouth.ac.uk/journal/valentin.pdf>, last accessed 2023/08/28
6. Reed, S.K.: *Learning rules in word problems: Research and Curriculum Reform*. Lawrence Erlbaum Associates (1999), <http://books.google.com.my/books?id=0FcfdR4PBHOC7dq=nctm=> (Word Problem), last accessed 2023/06/24
7. Mullis, I. V. S., Martin, M. O., Foy, P., Hooper, M.: *TIMSS 2015 International Results in Mathematics*. TIMSS & PIRLS International Study Center, Boston College, <http://timssandpirls.bc.edu/timss2015/international-results/>, last accessed 2023/08/10
8. Leung, F. K. S.: Mathematics education in East Asia and the West: Does culture matter?. In: Leung, F. K. S., Graf, K. D., Lopez-Real, F. J. (eds.), *Mathematics education in different cultural traditions— A comparative study of the East Asia and the West*, pp. 21–46. Springer, USA (2006)
9. Lim, C. S.: Cultural differences and mathematics learning in Malaysia. *The Mathematics Educator* **7**(1), 110-122 (2003)
10. Gurney, J.: Revealed: World pupil rankings in science and maths - TIMSS results in, *THE TELEGRAPH*, 29 NOVEMBER 2016, <http://www.telegraph.co.uk/education/2016/11/29/revealed-world-pupil-rankings-science-maths-timss-results/>, last accessed 2023/07/17
11. Gallagher, A. M.: Gender and antecedents of performance in mathematics testing. *Teachers College Record* **100**(2), 297-314 (1998)
12. Gallagher, A. M., DeLisi, R., Holst, P. C., McGillicuddy-DeLisi, A. V., Morely, M., Cahalan, C.: Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology* **75**(3), 165-190 (2000)

13. Eisenkopf, G., Hessami Z., Fischbacher, U., Ursprung, H.: Academic performance and single-sex schooling: Evidence from a natural experiment in Switzerland. *Journal of Economic Behavior and Organization* **115**, 123-143 (2012)
14. Ministry of Education Malaysia: Malaysian Education Blueprint 2013-2025 (preschool to Post-Secondary Education), <https://www.pmo.gov.my/wp-content/uploads/2019/07/Malaysia-Education-Blueprint-2013-2025.pdf>, last accessed 2023/08/23
15. Lyons-Thomas, J., Sandilands, D., Ercikan, K.: Gender differential item functioning in mathematics in four international jurisdictions. *Large-Scale Assessment Special Issue* **39**(172), 20-32 (2014)
16. DeMars, C.E.: Test stakes and item format interactions. *Journal of Applied Measurement in Education* **13**(1), 55- 57 (2000)
17. Hyde, J. S., Fennema, E., Lamon, S. J.: Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin* **107**(2), 139-155 (1990)
18. Abedalaziz, N.: A gender-related differential item functioning of mathematics test items. *International Journal of Educational and Psychological Assessment* **5**, 101-116 (2010a)
19. Abedalaziz, N.: Detecting gender related DIF using logistic regression and Mantel-Haenszel approaches. *Procedia-Social and Behavioural Sciences* **7**, 406-413 (2010b)
20. Bielski, J., Davison, M. L.: A sex difference by item difficulty interaction in multiple choice mathematics items administered to national probability samples. *Journal of Educational Measurement* **38**(1), 51-77 (2001)
21. Neidorf, T.S., Binkley, M., Gattis, K., Nohara, D.: 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. US Education Department of Education Statistics, US Department of Education, Institute of Education Sciences, NCES 2006-029, <http://nces.ed.gov/pubs2006/2006029-2.pdf>, last accessed 2023/09/12
22. Kolstad, R., Briggs, L. D.: Incorporating language arts into the mathematics curriculum: A literature survey. *Education* **116**(3), 423-449 (1996)
23. Kintsch, W., Greeno, J. G.: Understanding and solving word arithmetic problems. *Psychological Review* **92**(1), 109-129 (1985)
24. Yen W. M., Fitzpatrick A. R.: Item response theory. In: Brennan R. (ed.), *Educational measurement*. 4th ed, pp. 111-153. Praeger, Westport, CT (2006)
25. Thissen, D., Steinberg, L.: Item response theory. In: Maydeu-Olivares, A., Millsap, R. E. (eds.), *The Sage handbook of quantitative methods in psychology*, pp.148-177. SAGE Publication, USA (2009)
26. Linacre, J. M.: Understanding Rasch measurement: Estimation methods for Rasch measures. *Journal of Outcome Measurement* **3**(4), 382-405 (1999)
27. Linacre, J. M.: Data variance explained by Rasch measures. *Rasch Measurement Transactions* **20**(1), 1045-1054 (2006)
28. Lunz, M.E.: Comparison of item performance with large and small sample size. Measurement Research Associates: Test Insights, <https://www.rasch.org/mra/mra-03-10.htm>, last accessed 2023/10/10
29. Linacre, J. M.: Winsteps (Version 3.67.0) [Computer Software]. Winsteps.com, Chicago
30. Wright, B., Linacre, J. M.: A Rasch unidimensionality coefficient, <http://www.rasch.org/rmt/rmt83p.htm>, last accessed 2023/06/13
31. Fidalgo, A. M., Ferreres, D., Muniz, J.: Liberal and conservative differential item functioning using Mantel Haenszel and SIBTEST: Implications for Type I and Type II error. *Journal of Experimental Education* **73**(1), 23-29 (2004)

32. Zwick, R., Thayer, D. T., Lewis, C.: An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement* **36**(1), 1-28 (1999)
33. DeVellis, R.: *Scale development: Theory and applications*. Sage Publishing, USA (2003)
34. Mullis, I.V.S., Martin, M.O., Foy, P., Arora, A.: *TIMSS 2011 international results in mathematics*. TIMSS & PIRLS International Study Center, Boston College, Chestnut Hill, MA (2012)
35. Rajendran, N. S.: *Teaching and acquiring higher-order thinking skills theory & practice*. PenerbitUniversiti Pendidikan Sultan Idris, Malaysia (2008)
36. Cambridge University Press (2017). Questions: Wh-questions, <http://dictionary.cambridge.org/grammar/british-grammar/questions-and-negative-sentences/questions-wh-question>, last accessed 2023/09/09

## Appendix

### Rash Fit Statistics

**Table 7.** INPUT: 119 persons 24 items MEASURED: 118 persons 24 items 2 CATS

SUMMARY OF 107 MEASURED (NON-EXTREME) persons								
	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	16.6	24.0	1.16	.55	.99	.0	1.04	.1
S.D.	4.1	.0	1.09	.11	.20	.9	.65	.9
MAX.	23.0	24.0	3.68	1.05	1.60	2.8	5.08	3.6
MIN.	5.0	24.0	-1.66	.46	.63	-2.3	.31	-1.8
REAL RMSE	.58	ADJ.SD	.92	SEPARATION	1.59	person RELIABILITY	.72	
MODEL RMSE	.56	ADJ.SD	.94	SEPARATION	1.68	person RELIABILITY	.74	
S.E. OF person MEAN = .11								
MAXIMUM EXTREME SCORE:			11 persons					
LACKING RESPONSES:			1 persons					
SUMMARY OF 118 MEASURED (EXTREME AND NON-EXTREME) persons								
	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	17.2	24.0	1.51	.67				
S.D.	4.5	.0	1.51	.39				
MAX.	24.0	24.0	4.94	1.84				
MIN.	5.0	24.0	-1.66	.46				
REAL RMSE	.79	ADJ.SD	1.29	SEPARATION	1.64	person RELIABILITY	.73	
MODEL RMSE	.77	ADJ.SD	1.30	SEPARATION	1.68	person RELIABILITY	.74	
S.E. OF person MEAN = .14								
person RAW SCORE-TO-MEASURE CORRELATION = .94								
CRONBACH ALPHA (KR-20) person RAW SCORE RELIABILITY = .82								
SUMMARY OF 24 MEASURED (NON-EXTREME) items								

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT		OUTFIT	
					MNSQ	ZSTD	MNSQ	ZSTD
MEAN	84.8	118.0	.00	.27	.99	.0	1.04	.2
S.D.	18.5	.0	1.17	.08	.09	.7	.21	.9
MAX.	115.0	118.0	2.22	.60	1.19	1.1	1.44	2.1
MIN.	43.0	118.0	-2.95	.22	.84	-1.7	.69	-1.6
REAL RMSE	.28	ADJ.SD	1.14	SEPARATION	3.99	item	RELIABILITY	.94
MODEL RMSE	.28	ADJ.SD	1.14	SEPARATION	4.07	item	RELIABILITY	.94
S.E. OF item MEAN = .24								

UMEAN=.000 USCALE=1.000

item RAW SCORE-TO-MEASURE CORRELATION = -.97

2568 DATA POINTS. LOG-LIKELIHOOD CHI-SQUARE: 2368.21 with 2438 d.f. p=.8413

**DIF Output**

Name	CLASS	DIF		CLASS	DIF		DIFDIF	JOINT	Welch		MantelHanzl		item	
		MEASURE	S.E.		MEASURE	S.E.			CONTRAST	S.E.t	d.f.	Prob.		Prob.
C1	1	-2.95	.74	2	-2.98	1.02	.02	1.26	.02	98	.9858	.8575	+	1
	2	-2.98	1.02	1	-2.95	.74	-.02	1.26	-.02	98	.9858	.8575	-	1
C2	1	1.44	.30	2	1.12	.32	.32	.43	.74	103	.4609	.5189	.03	2
	2	1.12	.32	1	1.44	.30	-.32	.43	-.74	103	.4609	.5189	-.03	2
W3	1	-1.65	.46	2	-1.45	.54	-.20	.71	-.28	102	.7775	.6743	-.06	3
	2	-1.45	.54	1	-1.65	.46	.20	.71	.28	102	.7775	.6743	.06	3
W4	1	-1.45	.44	2	-2.98	1.02	1.53	1.11	1.37	84	.1734	.6582	+	4
	2	-2.98	1.02	1	-1.45	.44	-1.53	1.11	-1.37	84	.1734	.6582	-	4
C5	1	-.56	.35	2	-1.45	.54	.89	.65	1.38	96	.1708	.1334	-.59	5
	2	-1.45	.54	1	-.56	.35	-.89	.65	-1.38	96	.1708	.1334	.59	5
C6	1	.38	.30	2	.02	.36	.36	.47	.77	102	.4449	.5305	-.11	6
	2	.02	.36	1	.38	.30	-.36	.47	-.77	102	.4449	.5305	.11	6
C7	1	-1.10	.40	2	.02	.36	-1.12	.54	-2.09	104	.0386	.1704	.32	7
	2	.02	.36	1	-1.10	.40	1.12	.54	2.09	104	.0386	.1704	-.32	7
C8	1	.28	.31	2	.28	.34	.00	.46	.00	102	1.000	.9759	.14	8
	2	.28	.34	1	.28	.31	.00	.46	.00	102	1.000	.9759	-.14	8
C9	1	-.11	.32	2	-.57	.41	.46	.52	.89	100	.3782	.4914	.46	9
	2	-.57	.41	1	-.11	.32	-.46	.52	-.89	100	.3782	.4914	-.46	9
W10	1	-.44	.34	2	-.95	.46	.51	.57	.90	99	.3722	.3818	-1.02	10
	2	-.95	.46	1	-.44	.34	-.51	.57	-.90	99	.3722	.3818	1.02	10
W11	1	-.11	.32	2	.02	.36	-.13	.48	-.26	102	.7959	.8105	-.12	11

W11	2	.02	.36	1	-.11	.32	.13	.48	.26	102	.7959	.8105	.12	11
W11	1	-.11	.32	2	.14	.35	-.25	.48	-.53	103	.5984	.4753	.02	12
W12	2	.14	.35	1	-.11	.32	.25	.48	.53	103	.5984	.4753	-.02	12
W12	1	-1.45	.44	2	-.26	.38	-1.19	.58	-2.06	104	.0422	.0429	-.61	13
C13	2	-.26	.38	1	-1.45	.44	1.19	.58	2.06	104	.0422	.0429	.61	13
C13	1	-1.27	.41	2	-.75	.43	-.52	.60	-.86	103	.3901	.5975	.29	14
C14	2	-.75	.43	1	-1.27	.41	.52	.60	.86	103	.3901	.5975	-.29	14
C14	1	1.91	.31	2	1.61	.31	.29	.44	.66	104	.5092	.4494	.16	15
C15	2	1.61	.31	1	1.91	.31	-.29	.44	-.66	104	.5092	.4494	-.16	15
C15	1	.19	.31	2	.60	.33	-.41	.45	-.92	103	.3621	.7020	-.33	16
W16	2	.60	.33	1	.19	.31	.41	.45	.92	103	.3621	.7020	.33	16
W16	1	-.21	.33	2	-.41	.40	.19	.51	.38	101	.7055	.7985	-.67	17
W17	2	-.41	.40	1	-.21	.33	-.19	.51	-.38	101	.7055	.7985	.67	17
W17	1	-.11	.32	2	1.12	.32	-1.23	.45	-2.72	104	.0076	.0011	-1.10	18
W18	2	1.12	.32	1	-.11	.32	1.23	.45	2.72	104	.0076	.0011	1.10	18
W18	1	1.18	.30	2	1.02	.32	.16	.43	.36	103	.7187	.7424	.67	19
C19	2	1.02	.32	1	1.18	.30	-.16	.43	-.36	103	.7187	.7424	-.67	19
C19	1	.74	.30	2	-.12	.37	.85	.47	1.80	101	.0347	.0395	1.48	20
W20	2	-.12	.37	1	.74	.30	-.85	.47	-1.80	101	.0347	.0395	-1.48	20
W20	1	.74	.30	2	-.41	.40	1.15	.49	2.32	99	.0225	.0335	.04	21
C21	2	-.41	.40	1	.74	.30	-1.15	.49	-2.32	99	.0225	.0335	-.04	21
C21	1	1.01	.30	2	1.01	.32	.00	.43	.00	103	1.000	.9099	.08	22
C22	2	1.01	.32	1	1.01	.30	.00	.43	.00	103	1.000	.9099	-.08	22
C22	1	2.11	.32	2	2.34	.33	-.23	.46	-.50	103	.6174	.6751	-.24	23
C23	2	2.34	.33	1	2.11	.32	.23	.46	.50	103	.6174	.6751	.24	23
C23	1	1.26	.30	2	1.81	.32	-.55	.43	-1.26	103	.2099	.1205	-.99	24
W24	2	1.81	.32	1	1.26	.30	.55	.43	1.26	103	.2099	.1205	.99	24
W24														

-----  
 Size of Mantel-Haenszel slice: MHSlice = .010 logits

Size of Mantel-Haenszel slice: MHSlice = .010 logits

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

