Analysis of Differential Item Functioning (DIF) for Grade 10 mathematics test at secondary schools in Port Moresby, Papua New Guinea (PNG)

Jerome Oko

Abstract

This study examines gender related differential item functioning for Grade 10 students' mathematics test in secondary schools in Port Moresby, PNG. The sample of the study consisted of 355 Grade 10 students and the study employed a quantitative research method. To achieve the purpose of the study, the researcher used the students' responses for a 40 item mathematics test. The final 36 test items were used in the study after removing the other four items through validation due to violation of Rasch model rules. The results revealed that items 27, 29 and 31 were consistently found to have DIF between female and male students, both with item threshold and group plot approach. Items 27 and 29 were biased in favour of males, and item 31 is biased in favour of females. This study concluded that there was a significant difference between the performances of male and female students in the Grade 10 test.

Key words: Differential Item Functioning, test fairness, biased items, mathematics test, Rasch model, Item threshold approach, and Item fit approach

Introduction

Differential item functioning (DIF) analysis was conducted to confirm the fairness and equity of the test items and survey questionnaires due to the different gender groups surveyed (male and female). DIF analysis was utilised because it allows the researcher to determine if individual items or groups of items function differently for specified groups by looking at both the item level difference and the group difference on the items (Abou El-Komboz et al., 2014; Bond & Fox, 2015; Wu et al., 2016). For example, the Grade 10 respondents with the same ability level (both males and females) may have a different probability of a correct response due to items having different difficulties. This provides evidence of different probabilities or likelihoods of success on an item when matched on the ability or interest of different groups (Bansilal, 2015; Bond & Fox, 2015; Hagquist & Andrich, 2015). Therefore, "in order to preserve the unidimensionality trait of the construct under measurement, an important aspect of Rasch analysis was the examination of the presence of DIF in the various items" (Bansilal, 2015, p. 6). This procedure allows the researcher to determine if certain items were problematic when item bias is suspected, or if there is a response difference to certain types of items for gender subgroups. In this study, DIF analysis was employed to investigate whether individual items have similar psychometric properties among males and females who are from the same population (Abou El-Komboz et al., 2014; Guilera, Gómez-Benito, & Hidalgo, 2010; Hagquist & Andrich, 2015). The rationale for this is to confirm the item bias if different test-takers with equal ability who are from different subgroups do not have the same chance of success on an item (Brodersen et al., 2007).

DIF analysis investigates the quality of a measurement instrument (Bond & Fox, 2015; Bond et al., 2007; Wu et al., 2016). The principle underlying the detection of DIF is to investigate whether there is invariance in the person and item plots across different test situations (Bond et al., 2007; Brodersen et al., 2007; Wu et al., 2016). The procedure for detecting DIF items is necessary in the process of examining the fairness of a test and identifying the problematic items for revision or elimination before the administration of the test (Abou El-Komboz et al., 2014; Hagquist & Andrich, 2015; Wu et al., 2016). If item difficulties of a

group are plotted against another group, the slope that describes the invariance of the item difficulties should be equal to one (Abou El-Komboz et al., 2014; Bond & Fox, 2015; Wu et al., 2016).

Literature review

In the psychometric measurement and evaluation, the detection of biased items needs to be done in the interests of the equity and fairness of a test or an instrument for the disadvantaged subgroups such as female students or ethnic minority groups (Abedalaziz, 2011; Adams, 1992). Items that are biased can be rejected while constructing a test or selecting items for a specific test or during the process of evaluating the predictive validity (Adams, 1992). One of the major problems in the development of a test is to detect biased items (Boone & Scantlebury, 2005). Biased items are defined as items which differently function among the sub-group of examinees population (Boone & Scantlebury, 2005). There are two types of bias: external bias and internal bias (Adams, 1992). External bias, which is called test bias, is related to the fairness of the whole test during the procedure of test or item selection (Adams, 1992). Such bias can exist when test scores are correlated with other independent variables in the test (Osterlind, 1983, cited in Adams, 1992). It is regarded with the predictive and construct validity (Adams, 1992). The detection of internal bias is for examining the psychometric properties of items and that of a whole test (Adams, 1992). The internal bias is termed as item bias (Adams, 1992). The main reason of detecting item bias is to investigate whether or not individual items have similar psychometric properties among different sub-groups which are from the same population (Adams, 1992). Bond and Fox (2007) state that item will be biased if different test takers who are from different subgroups and with the equal ability do not have the same chance of success on an item (cited in Adams, 1992). DIF is the way to investigate the quality of a measurement instrument (Bond & Fox, 2007).

Three methods of identifying DIF in the Rasch model were carried out in many studies: *the item threshold approach*, *item fit approach*, and *group plot methods* to detect measurement invariance. These methods are now discussed below.

Item threshold approach

The item threshold approach is the first approach used to identify DIF in this study, and a common procedure used for evaluating DIF within the context of the Rasch model (Abd-El-Fattah, Al-Sinani, El Shourbagi, & Fakhroo, 2014; Hungi, 2005; Wu et al., 2016). This approach focuses on the difference between the threshold values (difficulty levels) of item in sub-groups. If the difference in the item threshold values is noticeably large, this implies that the item is particularly difficult for members of one of the groups being compared (Abd-El-Fattah et al., 2014; Hagquist & Andrich, 2015). This is not because of their different levels of the underlying latent trait, but due to other factors probably related to being members of that group (Abd-El-Fattah et al., 2014; Le, 2006; Strobl, Kopf, & Zeileis, 2015). With the item threshold approach, an item found to be more difficult for a group than the other items in a test is considered biased against that group (Abd-El-Fattah et al., 2014; Strobl et al., 2015). Those items with the largest differences in scale value are the suspect items.

A biased item can be detected by the difference of the threshold values (difficulty levels) of the items in the two groups. This is because the difficulty of an item (*d* parameter) is estimated separately for each group (Hungi, 2005; Meade & Fetzer, 2009; Scheuneman & Bleistein, 1999). Scheuneman and Bleistein (1999, p. 231) highlight that the difference in item difficulty between groups can be calculated with t statistics with the given formula:

$$t_i = \frac{(d_{1\_}d_2)}{\sqrt{SE_{i1}^2 + SE_{i2}^2}}$$

where: SE represents the standard error of d.

d_1 item difficulty value for group 1

d_2 item difficulty value for group 2

This formula can predict whether DIF is present when the t-value is large (Scheuneman & Bleistein, 1999). For example, in Table 5.6 p.135 in Chapter 5 all t-values or the standardised difference in item threshold are calculated following this calculation for item 01:

$$t_i = \frac{(-0.107 - 0.107)}{\sqrt{0.12^2 + 0.12^2}} = \frac{-0.214}{\sqrt{0.0288}} = \frac{-0.214}{0.169} = -1.27$$

Further, if there is a noticeable difference in the t-value, that particular item is recognized as being more difficult for a certain group than another group (Hungi, 2005; Scheuneman & Bleistein, 1999). As mentioned earlier, this is not due to the groups' different performance levels, but is a result of other factors related to the different features of the group members (Andrich & Hagquist, 2015; Hungi, 2005; Le, 2006). Both the unexpectedly difficult items and unexpectedly easy items for a specific group can be identified with this process. Bond and Fox (2015) point out that, in a high-stakes test, if item threshold difference is greater than 0.5 logits, this can be used as a criterion for detecting the DIF. This view is supported by Hungi (2005), who stresses that the absolute value of the difference of ± 0.5 logits indicates the difficulty difference of two sub-groups of one-year of school learning. The formula shown below is utilised.

$$-0.05 < \mathrm{d}_1 - \mathrm{d}_2 < 0.50$$

where: d_1= the item's difficulty value in group 1, and

d_2 = the item's difficulty value in group 2

Hungi (2005, p.146) states that the specified range value between -2.00 to +2.00 is the acceptable standardized difference of item difficulty: "Items whose differences in standardised item threshold between any of the groups fall outside a predetermined range do not confirm to the model and can be identified biased items". The formula below illustrates this scenario:

$$-2 < \mathrm{std}(\mathrm{d}_1\_\mathrm{d}_2) < 2.00$$

where: std = standardized difference of item difficulty

d_1= the item's difficulty value in group 1, and

d_2 = the item's difficulty value in group 2

The standardised difference in item threshold is important to test the premise that the difference in difficulty between males and females is statistically significant. This process is carried out through checking a set of criteria of a range of values of the standardised difference in item threshold among genders, as specified by Adams and Khoo (1993; cited in Hungi, 2005).

Furthermore, an item can be flagged as a DIF item when the absolute difference of item threshold for an item is greater than 0.25 logits (Abd-El-Fattah et al., 2014; Le, 2006). This absolute difference value is equal to approximately half of the school year to learn a distinct content area. However, when the difference in the item threshold value of an item between male and female respondents is outside the ±0.25 logit range, such a difference may cause significant concern related to item DIF (Abd-El-Fattah et al., 2014; Le, 2006).

Item fit approach

The item fit approach is the second approach in the Rasch model used to detect DIF. This approach investigates whether the items have equal discrimination power, allowing the infit mean square value for all items within the acceptable range (Hungi, 2005; Scheuneman & Bleistein, 1999). However, if the INFIT MNSQ values of the items are outside the range, the items can be assumed to be biased. This demonstrates that the items cannot equally discriminate in all different sub-groups (Bond et al., 2007; Hungi, 2005). In other words, non-biased items would fit the model in each group (Scheuneman & Bleistein, 1999). The acceptable range of item fit is between 0.77 to 1.30 for test items, and 0.60 to1.40 for Likert scales (Bond & Fox, 2015; Hagquist & Andrich, 2015; Hungi, 2005). This INFIT MNSQ range is useful to identify whether all items are satisfactorily fitting to the models when comparing subgroups (Hungi, 2005).

Group plot methods

The group plot method is the final approach for Rasch model, and it compares the area of the item characteristic curves (ICC) of difference in sub-groups. According to Alagumalai et al. (2005) and Wu et al. (2016), for the ICC, the gradient/slope of the ICC is positive when the probability (p) is 0.5. This is because p=0.5 describes the latent trait for item location for the specific group. The comparison of the area of the ICC of difference in two groups is equivalent to the difference between the item difficulties (Andrich & Hagquist, 2015; Hagquist & Andrich, 2015; Wu et al., 2016). This is because item difficulty is only one parameter of Rasch model. It can be concluded that the results of the group plot method will be similar to those of the item threshold approach discussed earlier. Results of the group plot method use graphs to demonstrate the difference between the two groups.

Methods

This section of the paper discusses the methods and procedures used to collect and analyse the mathematics test.

Development of Grade 10 test items

Mathematics test items were developed in reference to the Grade 10 mathematics curricula in PNG. The units/topics in the syllabus were utilised by formulating a table of specification for this study according to the distribution of the domains (knowledge, comprehension and higher order) stipulated in PNG Grade 10 table of specification. Grades 10 mathematics items were drawn from past examination papers of the PNG curriculum.

The Grades 10 test items were adopted and developed from PNG's past standardised examination papers. Adoption and development of the items required certain steps to ensure that the participants responded to the items with clarity within the given time frame. The researcher provided a draft of the test to Associate Professor Nicholas Buchdahl from School of Mathematical Sciences (the University of Adelaide) for a content check and suggestions. After adjustments, the test items were again given as a trial to the same Grade 10 students. After finalising these procedures, the study was ready to be conducted.

The Mathematics Examination consisted of 40 items for Grade 10 students. The Grade 10 mathematics students attempted the same items, because some topics that they study are similar. This was also convenient for the researcher and the schools in terms of scheduling the research. The examination items were developed from a Table of Specification (TOS) that highlighted the topics that were studied at Grade 10 level.

Participants

Prior to data collection, it was important for this study to clarify the population to be analysed. The population of the study was defined as students formally enrolled in Grades 10 secondary schools in Port Moresby. This definition was possible because mathematics is a core compulsory subject studied by students of the specified cohort. After that, the researcher collected data in Port Moresby, based on Roiser and Roos' (cited in Keeves, 1992) stratified random sampling technique. The primary data collected was from 354 Grade 10 students from the different secondary schools in Port Moresby. The genders in Grades 10 were in proportion to males and females in Port Moresby within the selected type of schools. The reason for selecting 16 schools in each region was based purely on the amount of research work that was scheduled, the availability of the schools and financial considerations. In order to carry out the data collection procedure, class lists were obtained from each of the schools to indicate the number of students selected. In each school, sampled data were collected from an intact classroom to randomly select the respondents (Roiser & Roo cited in Keeves, 1992). This was possible through the stratification process involved at another stage of selecting the sample population, though there were challenges faced by the researcher in accessing the schools.

Prior to the administration of this study, it was necessary to obtain ethical research approval from the University of Adelaide's Human Research and Ethics Committee (UAHREC) (Ethics Approval No H-2017-133). The committee's approval was granted on five conditions to which the researcher must conform: 1) every participant be provided with an information sheet about the study; 2) every participant must read, sign and return the consent form to the researcher to participate in the study; 3) consent from parents/guardians was to be obtained for participants below the age of 18 years; 4) the identity of every participant was to be kept confidential when conducting survey questionnaires, and 5) participation was voluntary and participants were free to discontinue at any time. Consideration of these conditions was important in carrying out the study. The conditions were made clear to participants through the participants' consent letter before administration of the survey and interview to clarify doubts pertaining to ethics requirements.

Differential item functioning (DIF) analysis was conducted for only the 36 items that fitted the Rasch model, in order to assess the fairness and equity of the test items in terms of the different gender groups (male and female). Where DIF is present in an item, this suggests that the item functions differently in different contexts; in other words, that two groups of people with the same ability have different probabilities of success in responding to an item (Hagquist & Andrich, 2015, 2017; Wu et al., 2016). For example, in this study, both male and female Grade 10 students with similar average mathematics abilities were given a test, in which an item was administered with the context of building a house. From this, it is observed that males in this group performed considerably better than females on this item, even though girls and boys performed similarly on other items. This is because the males were more familiar with the context of the question than the females, and so males found the item easier than females did. This example demonstrates an item that exhibits DIF for the two gender groups. This research therefore checked DIF both on the test level and the item level for clarity on the fairness and equality of the test items on gender.

Results and Discussion

Overall test level DIF for gender

This section analyses and compares the overall mathematics test performance for Grade 10 male and female students in Port Moresby. As shown in Table 1.1, the INFIT MNSQ values of the overall test for both males and females is within the acceptable value used in this study of 0.70 to 1.30. Additionally, Table 1.1 also shows the item difficulty parameter estimates for each of the 36 items; the negative sign for the 17 items on the parameter estimates indicates that they were easier for male students, while the 19 items parameter estimates with positive signs indicates that those items were more difficult for females (Wu & Adams,

2007; Wu et al., 2016). Generally, it can be observed that there was not much difference in difficulty levels between male and female respondents.

Table 1.1 Test level gender difference statistics

Item No	Estimate	Error	INFIT MNSQ	Item No	Estimate	Error	INFIT MNSQ
Item 01	-1.15	0.12	1.0	Item 22	1.38	0.15	0.98
Item 02	0.72	0.13	0.94	Item 23	-0.63	0.11	0.99
Item 05	-0.49	0.11	1.03	Item 24	-0.21	0.11	0.93
Item 06	0.68	0.13	1.04	Item 25	-0.52	0.11	0.92
Item 07	-0.70	0.11	0.97	Item 26	1.10	0.14	1.01
Item 08	-0.83	0.11	1.01	Item 27	-0.16	0.11	1.05
Item 09	0.24	0.12	0.91	Item 28	-0.14	0.11	0.92
Item 10	0.31	0.12	1.04	Item 29	1.41	0.15	1.08
Item 11	0.74	0.13	1.02	Item 30	0.82	0.13	1.01
Item 12	-1.41	0.12	1.00	Item 31	-0.76	0.12	1.01
Item 13	-1.01	0.12	0.97	Item 32	-0.13	0.12	1.06
Item 14	-0.56	0.11	0.94	Item 33	0.44	0.12	0.92
Item 15	0.50	0.12	0.96	Item 34	-0.07	0.12	1.00
Item 16	0.00	0.12	1.05	Item 35	-0.24	0.12	0.98
Item 17	-1.08	0.12	1.00	Item 36	-0.19	0.12	1.02
Item 18	0.35	0.12	0.94	Item 37	0.34	0.13	1.04
Item 20	0.17	0.11	0.98	Item 39	0.35	0.12	1.03
Item 21	0.25	0.12	1.09	Item 40	0.48	0.13	1.12

Table 1.2 Test gender differences in ability estimates

	Sex	Estimate	Error	INFIT MNSQ	ZSTD (t)
1	Female	0.042	0.042	1.03	0.3
2	Male	-0.042*	0.042	1.07	0.7

Chi-square test of parameter equality = 0.99, df = 1

The results of the mean estimates of male and female respondents for the overall test are shown above in Table 1.2. This table shows estimates for gender differences in ability estimates at test level. From these results, it is apparent that the test is easier for male students than for female students, indicated by the parameter estimate of -0.042 for males. The actual parameter estimate for male students is one (0.042/0.042), equal to its standard error estimate, and so the difference between the male and female means is obviously insignificant. However, if the actual parameter estimate for male students is two or three times larger than its standard error estimate, then the difference between the male and female means is obviously significant (Wu & Adams, 2007; Wu et al., 2016). This difference is associated with the chi-square value of 0.99 on one degree of freedom as shown in Table 1.2. Therefore, it can be concluded that the male

students' mean performance is the same as the female students. As seen in Table 1.2, the difference between the actual parameter estimate of male and female students shows that males scored 0.084 logits lower than female students. Hence, a difference of 0.084 logits, which is much smaller than 0.5 logits, implies that the average performance levels of male and female students are not substantially different. The overall test INFIT MNSQ and t-values are within the range of 0.70 to 1.30 and -2 to 2, respectively. Specifically, females and males have INFIT MNSQ values of 1.03 and 1.07 with t-values of 0.3 and 0.7, respectively. Together, these statistical findings suggest that there is no DIF in the overall mathematics test for Grade 10 students. In the next discussion, individual items on the test are checked to ascertain whether there exist DIF items for male and female students at the item level.

Item level DIF for gender

Following analysis for DIF at the overall test level, individual items from the mathematics test were analysed for DIF between male and female participants. This step was carried out due to the insignificant difference found for gender in the overall test analysis. At the item level of analysis, on the other hand, it is expected that DIF would be detected on the items for gender. Item level analysis was conducted through three methods of identifying DIF in Rasch model: the item threshold approach, item fit approach, and group plot methods (Hungi, 2005; Scheuneman & Bleistein, 1999). These methods were used as they underpin the assumptions of item response theory (IRT), which makes it useful to investigate DIF. The estimated parameters of the item response function (the probability that persons with lower ability have less of a chance to give the correct answer, while persons with higher ability are very likely to answer correctly) are unchanged for different samples drawn from the same population (Scheuneman & Bleistein, 1999). "Therefore, if parameters are estimated separately for two groups, the resulting item response functions of an item which is functioning equivalently for those groups should be the same" (Scheuneman & Bleistein, 1999, p. 229). In other words, the probability of a correct response for respondents at a given ability level is the same for males and females, since a true ability scale is used rather than observed test scores. This, therefore, allowed the researcher to use the three methods to detect DIF in Rasch model discussed in the literature review section instead of other methods utilised by other researchers in classical test theory (CTT).

Item threshold approach

In Table 1.3, negative values of difference in item threshold and difference in standardised item threshold, signify that the item is relatively easier for female students than for male students, while positive values for males imply the opposite (Abd-El-Fattah et al., 2014; Le, 2006). These differences are apparent in the Grade 10 test (see Table 1.3) with 17 items having negative values of item threshold difference and standardised item threshold difference. This implies that these items apparently favors female students over male students. The other 19 items have a positive value of difference in item threshold and difference in standardised item threshold, indicating that they are relatively more difficult for male respondents than for the female respondents.

Furthermore, the difference of the item threshold for an item d_1-d_2 <-0.5 and d_1-d_2>0.5 logits (Abd-El-Fattah et al., 2014; Le, 2006) indicates an item with DIF. This difference value is approximately equal to an extra one year of school to learn a distinct content area. When the difference in the item threshold value of an item between male and female respondents is below or above the predefined range of d_1-d_2 <-0.5 and d_1-d_2>0.5 logit range, such a difference may cause a significant concern regarding the item DIF (Abd-El-Fattah et al., 2014; Le, 2006).

In this study, as shown in Table 1.3, items 27, 29 and 31 are not within the acceptable item threshold logit value range of d_1-d_2 <-0.5 and d_1-d_2>0.5, and 34 items are within the acceptable range of d_1-d_2 <-0.5 and d_1-d_2>0.5 logits (Hungi, 2005; Le, 2006). It is observed that items 09, 17, 32, 36 and 39 are more difficult for male respondents than for female respondents, with differences of item threshold values less than -0.25 logits (Hungi, 2005). According to Hungi (2005), this indicates that male respondents require more time to learn some Grade 10 mathematics content compared to female respondents. On the other hand, items 08 and 37 are significantly more difficult for female students than male students with the differences of item

threshold values are greater than 0.25 logits. In other words, female respondents need more time to learn this Grade 10 content compared to the male respondents (Hungi, 2005). Added to this, items 27, 29 and 31 posed a substantial amount of DIF and were deleted, with absolute difference values above and below + 0.5 logits to -0.5 logits threshold, as recommend by Hungi (2005). This would indicate that items 27 and 29 favored male respondents, and item 31 favored female respondents. Both genders require more time to learn the specific content area about these DIF items.

Table 1.3 DIF results for the 36 mathematics test items by item threshold

Item No	Female		Male			
	Estimate (d1)	Error (e1)	Estimate (d2)	Error (e2)	d1-d2	st (d1-d2)
Item 01	-0.11	0.12	0.107	0.12	-0.21	-1.26
Item 02	0.05	0.13	-0.05	0.12	0.09	0.51
Item 05	0.07	0.11	-0.07	0.11	0.14	0.89
Item 06	0.05	0.13	-0.05	0.12	0.10	0.55
Item 07	-0.08	0.11	0.08	0.11	-0.16	-1.02
Item 08	0.19	0.11	-0.19	0.11	0.39	2.00
Item 09	-0.13	0.12	0.13	0.11	-0.27	-1.62
Item 10	0.11	0.12	-0.11	0.11	0.22	1.32
Item 11	-0.05	0.13	0.06	0.12	-0.11	-0.66
Item 12	0.07	0.12	-0.07	0.12	0.15	0.78
Item 13	0.11	0.12	-0.11	0.11	0.23	1.39
Item 14	0.10	0.11	-0.10	0.11	0.202	1.28
Item 15	0.05	0.12	-0.03	0.12	0.09	0.54
Item 16	-0.11	0.12	0.11	0.11	-0.23	-1.39
Item 17	-0.20	0.12	0.20	0.11	-0.41	-2.00
Item 18	0.01	0.12	-0.01	0.11	0.03	0.16
Item 20	-0.09	0.11	0.09	0.11	-0.18	-1.12
Item 21	0.01	0.12	-0.02	0.11	0.03	0.19
Item 22	-0.04	0.15	0.04	0.14	-0.09	-0.41
Item 23	-0.09	0.11	0.09	0.11	-0.19	-1.19
Item 24	-0.09	0.11	0.09	0.11	-0.19	-1.20
Item 25	-0.08	0.11	0.08	0.11	-0.16	-1.05
Item 26	0.03	0.14	-0.03	0.13	0.06	0.31
Item 27	0.32	0.11	-0.32	0.11	0.64*	4.03
Item 28	0.02	0.11	-0.02	0.11	0.03	0.22
Item 29	0.29	0.15	-0.29	0.15	0.58*	2.68
Item 30	0.07	0.13	-0.07	0.13	0.14	0.76
Item 31	-0.24	0.12	0.25	0.11	-0.49*	-2.92
Item 32	-0.13	0.12	0.13	0.11	-0.27	-1.63
Item 33	-0.00	0.12	0.01	0.12	-0.00	-0.05
Item 34	0.05	0.12	-0.05	0.11	0.09	0.56

Item 35	0.07	0.12	-0.07	0.11	0.15	0.89
Item 36	-0.16	0.12	0.16	0.11	-0.33	-1.94
Item 37	0.16	0.13	-0.16	0.12	0.32	1.79
Item 39	-0.15	0.12	0.151	0.12	-0.30	-1.74
Item 40	-0.03	0.13	0.03	0.12	-0.07	-0.38

* = difference in item difficulty outside the range d_1-d_2 <-0.5 and d_1-d_2>0.5, Total (N=355)

Item fit approach

INFIT MNSQ values between the ranges of 0.70 to 1.30 are used to detect DIF in items for male and female respondents. The items with INFIT MNSQ values outside this acceptable range for males and females are assumed to be misfitting items, or items with DIF. However, it is evident from Table 1.4, that the final 36 items appearing in the Grade 10 mathematics test recorded INFIT MNSQ values that are within the predetermined range (0.70 to 1.30) for all respondents, both male and female. Therefore, based on INFIT MNSQ criteria, it is clear that gender DIF is not a significant problem in the 36 Grade 10 mathematics test items. That said, 16 items are easier for females and twenty items are more difficult for males; this, however, does not mean that there is DIF between male and female students. Instead, it indicates that Grade 10 male students in Port Moresby performed better compared to their female counterparts on the mathematics test.

Table 1.4 Test item fit statistics for gender (male and female)

	Female			Male		
Item No	Estimate (d1)	Error (e1)	INFIT MNSQ	Estimate (d2)	Error (e2)	INFIT MNSQ
Item 01	-0.11	0.12	0.99	0.11	0.12	1.01
Item 02	0.05	0.13	1.00	-0.05	0.13	0.88
Item 05	0.07	0.11	1.06	-0.07	0.11	1.01
Item 06	0.05	0.13	1.02	-0.05	0.13	1.06
Item 07	-0.08	0.11	0.96	0.08	0.11	0.98
Item 08	0.19	0.11	1.04	-0.19	0.11	0.98
Item 09	-0.13	0.12	0.88	0.13	0.12	0.95
Item 10	0.11	0.12	1.05	-0.11	0.12	1.02
Item 11	-0.06	0.13	1.08	0.06	0.13	0.96
Item 12	0.07	0.12	0.99	-0.07	0.12	1.01
Item 13	0.11	0.12	0.91	-0.11	0.12	1.03
Item 14	0.10	0.11	0.92	-0.10	0.11	0.96
Item 15	0.05	0.12	0.95	-0.05	0.12	0.97
Item 16	-0.12	0.12	1.13	0.12	0.12	0.98
Item 17	-0.20	0.12	0.98	0.2	0.12	1.02
Item 18	0.01	0.12	0.92	-0.01	0.12	0.96
Item 20	-0.09	0.11	1.04	0.09	0.11	0.93
Item 21	0.02	0.12	1.07	-0.02	0.12	1.10
Item 22	-0.04	0.15	0.95	0.04	0.15	1.01

Item 23	-0.1	0.11	1.04	0.1	0.11	0.95
Item 24	-0.1	0.11	0.92	0.1	0.11	0.95
Item 25	-0.08	0.11	0.94	0.08	0.11	0.91
Item 26	0.03	0.14	1.01	-0.03	0.14	1.01
Item 27	0.32	0.11	1.03	-0.32	0.11	1.08
Item 28	0.02	0.11	0.9	-0.02	0.11	0.94
Item 29	0.29	0.15	1.06	-0.29	0.15	1.09
Item 30	0.07	0.13	1.03	-0.07	0.13	0.99
Item 31	-0.25	0.12	0.99	0.25	0.12	1.02
Item 32	-0.14	0.12	1.06	0.14	0.12	1.06
Item 33	0.00	0.12	0.98	0.00	0.12	0.88
Item 34	0.05	0.12	1.01	-0.05	0.12	0.99
Item 35	0.07	0.12	1.01	-0.07	0.12	0.95
Item 36	-0.16	0.12	1.00	0.16	0.12	1.03
Item 37	0.16	0.13	1.11	-0.16	0.13	0.99
Item 39	-0.15	0.12	1.05	0.15	0.12	1.02

Group plot methods

The third approach used in Rasch model to identify DIF in this study is the group plot method. As with the above-discussed analysis, the group plot method is used to detect DIF due to gender. Figure 1.1 shows an example of a non-suspect item (item 06) while Figures 1.2, 1.3 and 1.4 show the item characteristics curve of items 27, 29 and 31, which are identified to be suspects of DIF in the previous section (item threshold approach). Figure 1.1 shows a comparison between male (light green curve) and female (dark blue curve) average scores on item 06 at each autonomy level, and indicates non-suspect DIF between female and male respondents on that item.

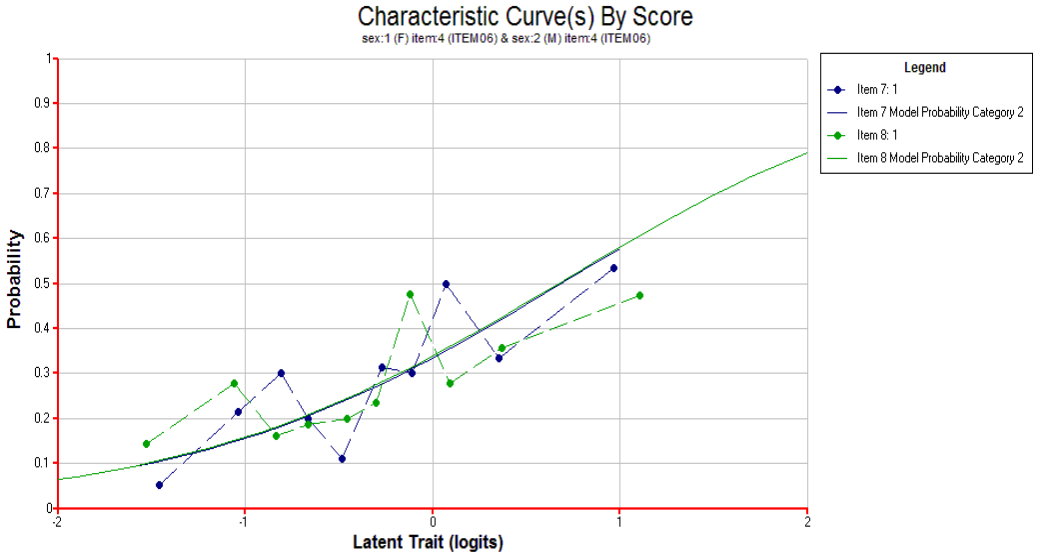

Figure 1.1 ICC for Item 06 showing non-bias between male and female respondents

It can be seen from Figure 1.2 (item 27) and Figure 1.3 (item 29) that the ICCs for males (light green curve) are clearly higher than those for females (dark blue curve), which means that the males stand a greater chance than females of getting these items correct at the same ability level. These two items are suspected of DIF, with absolute difference values above +0.5 logits. This indicates that these two items are more in favour of males than females. On the other hand, the ICC shown in Figure 1.4 (item 31) is mostly higher for females than for males. This means that the item is biased in favour of the female respondents.

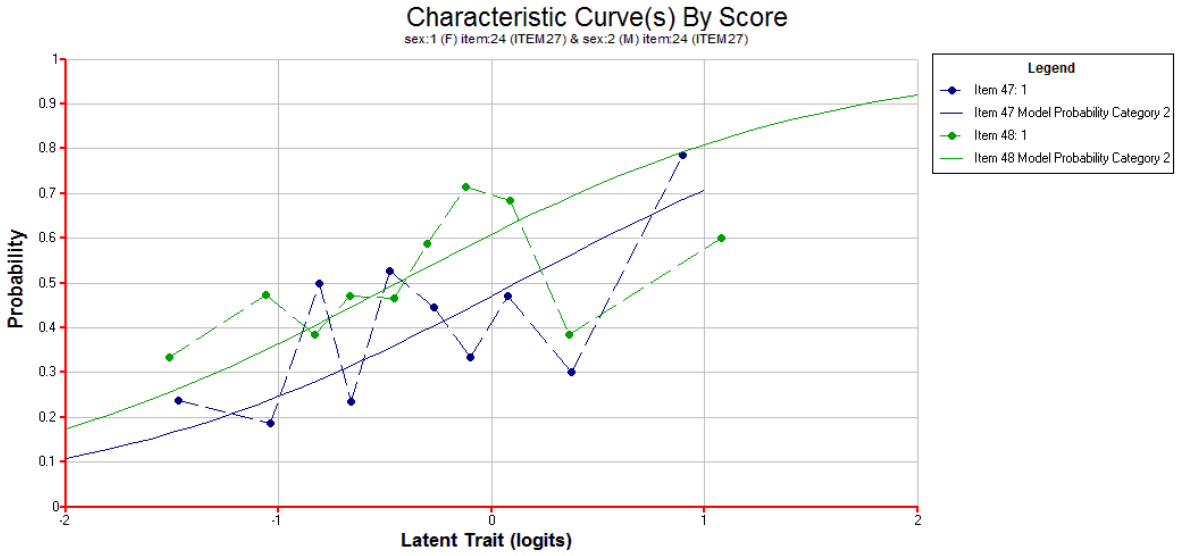

Figure 1.2 ICC for items 27 biased in favour of male respondents

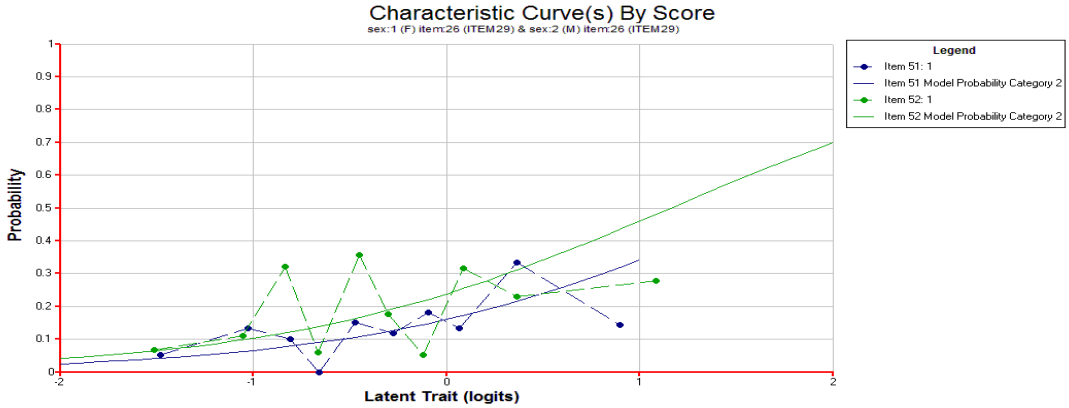

Figure 1.3 ICC for items 29 biased in favour of male respondents

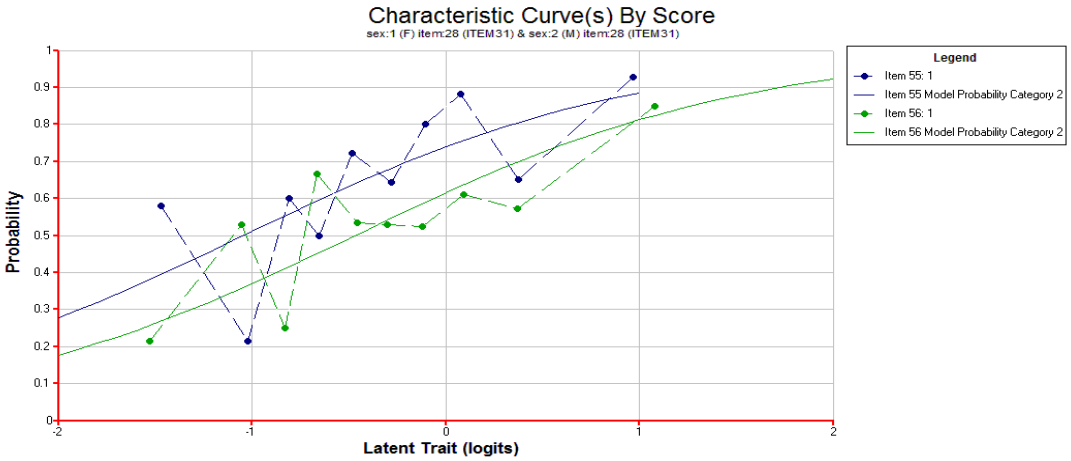

Figure 1.4 ICC for items 31 biased in favour of female respondents

The DIF in Figure 1.2 (item 27), Figure 1.3 (item 29) and Figure 1.4 (item 31) shows that, though respondents have the same location (as evidenced by the analysis of the items), they scored differently depending on their gender. More importantly, the aforementioned figures show that, given the same total score, there is a difference by gender. According to Andrich and Styles (2004), this total score causes the items to show an opposite effect to some degree; a few items show a small amount of DIF, and a few items show the opposite to what is shown in Figures 1.2, 1.3 and 1.4. Therefore, if these items are to be studied further in terms of their relationships with the content of teaching, further investigations along these lines would need to be carried out.

Overall DIF analysis provided information associated with equity and fairness of item functioning and respondent ability difference, thereby providing equity and fairness for disadvantaged groups. This analysis was carried out through three different methods (INFIT MNQS, item threshold and group plot approach) that identified DIF items with gender. These methods were utilised because they are not sample dependent.

The item fit approach examined the INFIT mean square statistics of different sub-groups within the acceptable range of 0.70 to 1.3. This approach could not identify any item as DIF for two significant reasons; (1) the INFIT mean square statistics of the 36 items for the Grade 10 sample already existed within the acceptable range; and (2) the INFIT mean square statistics are for examining all items' fit within the Rasch theoretical curve. The item threshold approach and group plot approach provide similar results. Between these two, it is observed that the item threshold approach was the best method to provide the details of item thresholds and ability for male and female cohorts. The group plot approach presented graphs for item function through the range of ability for males and females.

The empirical findings obtained through the different methods, the quality of item in regard to fairness, and the ability gap between males and females, are all associated with DIF. The item threshold approach provided more specific findings; namely, that there were some items that favoured males and some that favoured females. Items 27, 29 and 31 were consistently found to have DIF between female and male students, both with item threshold and group plot approach. Items 27 and 29 were biased in favour of males, and item 31 is biased in favour of females. Therefore, this result indicates that there was a significant difference between the performances of male and female students in the Grade 10 test.

Though the overall test statistics indicate no significant difference in DIF for gender, however, the individual item analysis indicated that items 27, 29 and 31 have DIF within the Rasch model approaches. These three items were removed from the test for three reasons: a) the three items were not within the set statistical criteria for DIF, b) with a sample size of 355, the effect of DIF on the item results would be

visible; and c), there were 36 items that were tested for DIF, and the results of these could have been affected considerably by the three items with DIF. This is because the number of items was small and the DIF influence in favour and against males and females was not evenly distributed, with items 27 and 29 favouring males and item 31 favouring females (Wu & Adams, 2007; Wu et al., 2016).

The three items with DIF detected featured content related to trigonometric functions under the topics Pythagoras theorem and trigonometric application, respectively. The observed gender differences on these topics may have been influenced by culture rather than content competence, because males are more dominant in the application aspect of mathematics in PNG, compared to females (Leder et al., 1996; Sukthankar, 1995). These three items (27, 29 and 31) show that there was difference between the male and female students in understanding and solving trigonometric functions.

Conclusion

Analysis of the DIF in relation to gender was necessary, to examine the fairness and equity of the test in identifying biased items. The analysis was carried out through three different methods (INFIT MNQS, item threshold and group plot approach) that identified DIF items with gender. These methods were utilised because they are not sample dependent. The investigation of DIF showed a significant difference on the item level analysis and detected three items (27, 29 and 31) with DIF; two items (27 and 29) biased towards males, and one (31) item towards females. The Rasch model's item threshold and group plot approaches were used to identify the DIF items. This result indicates that there was a significant difference between the performances of male and female students in the Grade 10 test. The overall item analysis was important for test and examination question improvement for students.

References

Abd-El-Fattah, S. M., Al-Sinani, Y., El Shourbagi, S., & Fakhroo, H. A. (2014). Using Rasch Analysis to examine the Dimensionality Structure and Differential Item Functioning of the Arabic version of the perceived physical ability scale for children. *Australian Journal of Educational & Developmental Psychology, 14*, 29-44. https//www.newcastle.edu.au/journal/ajedp/

Abedalaziz, N. (2011). Detecting DIF using Item Characteristic Curve Approaches. *The International Journal of Educational and Psychological Assessment.* 8(2) 1-15.

Abou El-Komboz, B., Zeileis, A., & Strobl, C. (2014). Detecting Differential Item and Step Functioning with Rating Scale and Partial Credit Trees. *Department of Statistics*: *Technical Reports*, No.152, Department of Statistics University of Munich. doi: 10.5282/ubm/epub.17984

Adams, R. J. (1992). Item bias. Keeves, J. P. (Ed.) (1992). *The IEA technical handbook*. The Hague, The Netherlands: The International Association for the Evaluation of Educational Achievement.

Alagumalai, S., & Curtis, D. D. (2005). Classical test theory. In S. Alagumalai, D. D. Curtis, & N. Hungi (Eds.), Applied Rasch measurement: A book of exemplars (139-157). Springer.

Andrich, D., & Styles, I. (2004). Final Report on the Psychometric Analysis of the Early Development Instrument (EDI) using the Rasch model: A technical paper commissioned for the development of the Australian Early Development Instrument (AEDI). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.495.5875&rep=rep1&type=pdf

Bansilal, S. (2015). A Rasch analysis of a Grade 12 test written by mathematics teachers. *South African Journal of Science, 111*(5/6), 68-69. doi:10.17159/sajs.2015/20140098

Boone, W. J., & Scantlebury, K. (2005). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. doi:10.1002/sce.20106

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. (2nd ed.). Taylor & Francis Group.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*: Routledge.

Brodersen, J., Meads, D., Kreiner, S., Thorsen, H., Doward, L., & McKenna, S. (2007). Methodological aspects of Differential Item Functioning in the Rasch model. *Journal of Medical Economics, 10*(3), 309-324. doi: 10.3111/13696990701557048.

Guilera, G., Gómez-Benito, J., & Hidalgo, M. D. (2010). Citation analysis in research on differential item functioning. *Quality & Quantity, 44*(6), 1249-1255. doi:10.1007/s11135-009-9274-3

Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF–a study based on simulated polytomous data. *Psychological Test and Assessment Modeling, 57*(3), 342-376.

Hagquist, C., & Andrich, D. (2017). Recent advances in analysis of Differential Item Functioning in health research using the Rasch model. *Health and quality of life outcomes, 15*(181),1-8. doi: 10.1186/s12955-017-0755-0

Le, L. (2006). Analysis of Differential Item Functioning. Paper presented at the annual meeting of American Educational Research Association, San Francisco CA. https://www.acer.org/files/analysis

Leder, G. C., Forgasz, H. J., & Solar, C. (1996). Research and intervention programs in mathematics education: A gendered issue. *International handbook of mathematics education* (945-985). https://doi.org/10.1007/978-94-009-1465-0_26

Scheuneman, J., & Bleistein, C. (1999). *Item bias*. *Advances in measurement in educational research and assessment,* Pergamon.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting Differential Item Functioning in the Rasch model. *Psychometrika, 80*(2), 289-316.

Sukthankar, N. (1995). Gender and mathematics education in Papua New Guinea. *Equity in mathematics education: Influences of feminism and culture, 51*(2)135-140.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*: Educational Measurement Solutions Melbourne. http:www.edmeasuement.com.au

Wu, M., Tam, H. P., & Jen, T.-H. (2016). *Educational Measurement for applied researchers: Theory into practice*. Springer. doi:10.1007/978-981-10-3302-5

About the author

Jerome Oko, PhD, is currently the Campus Administrator for Divine Word University (DWU) at Port Moresby Campus in Papua New Guinea. He obtained his Bachelors in Education-Technical degree from Don Bosco Technological Institute in Papua New Guinea. He received his Master and Doctoral degrees in Mathematics Education from the University of Adelaide, South Australia. His Research areas include Mathematics and Science Education. His research covers measurement and evaluation in Education. He mainly employs in his data analysis newer psychometric techniques such as Rasch model and multi-level modelling.

---oo0oo---