INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 7, Issue 1, (2018) E-ISSN: 2226-6348

Detecting Item Bias in an Anatomy & Physiology Test for Nursing Students using Item Response Theory

Hishamuddin Ahmad, Siti Eshah Mokshein, Mohd Razimi Husin

Universiti Pendidikan Sultan Idris, Malaysia

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v7-i1/3904 DOI:10.6007/IJARPED/v7-i1/3904

Published Online: 22 February 2018

Abstract

This study aimed to assess whether or not there were bias items towards male or female examinees in the Anatomy & Physiology (A&P) test for the Diploma of Nursing in Ministry of Health (MOH), Malaysia. The study involved 971 examinees from the first semester cohort of the January 2013 session in which 867 examinees were females and 104 males. A differential item functioning (DIF) analysis was conducted with the help of *Xcalibre* software using *Mantel-Haenszel* coefficient (*M-H*) method. While 88.9% of the items did not indicate bias, three items were found to demonstrate bias, namely Item 17 (*M-H* = 0.28, $p < 0.05$), 18 (*M-H* = 0.51, $p < 0.05$), and 29 (*M-H* = 0.54, $p < 0.05$), from the topics of Cardiovascular System and Digestive System. All these three items favour female examinees where by female examinees tend to answer it correctly as compared to male students. These items need to be further revised, so that decisions can be made whether to improve or to remove them from the test.

Keywords: Item response theory, Differential item functioning, Assessment, Anatomy & Physiology

Introduction and Background

There are 16 colleges offering Diploma in Nursing programs in the Ministry of Health Malaysia (MOH). The colleges are located across Malaysia with 12 in the peninsular, two in Sabah and two in Sarawak in 2015. Students pursuing Diploma in Nursing at these colleges are the majority of female students. However, there are also minority students who follow the same study which are male students. Male students in the field of nursing are minority not only in Malaysia, but also at most regions of the world. The stigma arise is nursing study will always give an advantage to female students as compared to male students.

During the first semester, the subjects of Anatomy & Physiology (A&P) are among the subjects that need to be learned in addition to the other subjects. A&P subjects are subjects with the highest credit score and are essential as a basic knowledge in nursing fields regardless of male or female students.

In the assessment of A&P subject learning, multiple choice item items are among the methods currently being used. Multiple choice item have been fully accepted in most modern societies as the most objective method of decision making in schools, institutions of higher

learning, and industries. It is now applied not only in the field of education, but also includes the test of admission, recruitment, promotion, placement, evaluation, guidance and research. Because of the importance of multiple choice items in assessing student achievement in A&P subjects, specifically in the field of nursing, the items that are enacted should be fair to both groups of male and female students.

Fair items for subgroups of examinees who sit on a test are something that is rarely addressed or noted. In order to ensure an efficient measurement system, the fairness of the items for the examinee who sits on the test is among the issues to be considered. The aim of fairness is refer to unbiased items between two different groups, whether gender (male or female), race (Malay or non-Malay), religion (Muslim or non-Muslim) and so on. Originally known as a biased item (Lord, 1980), Hambleton, Swaminathan and Rogers (1991) stated that this phenomenon is related to biased elements against a group of examinees. Adedoyin (2010) has found that many researches in the field of educational measurement towards improving the test or examination fairness in various subgroups of the examinees have been carried out because of the test scores the examinee earned was very important to the provider of the examination in decision making. However, the presence of biased items is alarming as testing is usually used as a controller for educational opportunities. This means that for examinees who get a minimum achievement, their chances of continuing their studies will be obtained. But on the other hand for unsuccessful examinees to reach the minimum requirement, it may restrict the examinee's opportunity to continue his studies. This is a very important issue for test items to be fair to every examinee.

Fairness in item test is an ongoing assessment issues. Hambleton et al. (1991) found that issues related to the test and of course very important to the examinees is the fairness of test items. A test is considered as fair if the test gives all potential examinees the opportunity to demonstrate the skills and knowledge they have acquired in relation to the purpose of the test (Adedoyin, 2010). At the same time Hambleton et al. (1991) stated that an item is considered biased if examinees with the same abilities, but from separate groups, have different probabilities to get the correct answer.

Previously, many researches (Adedoyin, 2010; Abedlaziz, Ismail, & Hussin, 2011; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012; Sharp, Michonski, Steinberg, Fowler, Frueh, & Oldham, 2014) on biased items in gender have been conducted both internally and internationally. The studies have shown that there are biased items on a gendered group. However, there has been no research that investigates whether gender differences can contribute to the elemental weight of items in the A&P subject test in MOH. Therefore, it is supposed to be that test providers should conduct research on test items to demonstrate that they are fair and free from bias towards a group.

Biasness is the presence of several item features that result in different performance for individuals with the same abilities but different from the subgroups of the examinees. Biased items can also be defined as systematic invalidations or errors in how the test items measure one construct for a particular group member (Adedoyin, 2010).

Schmitt and Kuljanin (2008) stated that the measurement invariance refers to the consistency of a measurement of a group such as gender, ethnic groups, different groups of abilities and so on. They also found that, measurement equivalence is the basis of fairness in measuring by ensuring that every latent trait measured against an item or indicator is the same across each group studied.

However, with appropriate analysis, biased items that are present in the test of multiple choice items can be detected. What is needed to detect biased items is the analysis

of complex interaction patterns between subgroups and individual factors as well as item characteristics (McArthur, 1981).

Item Response Theory

IRT is related to the probability of answering an item correctly or reaching a specific response level modeled as an individual's ability function and item characteristics. IRT begins with the fact that individual responses to items or specific questions are determined by the mental nature of unobservable or latent examinees. IRT allows the latent properties measured on a scale of theta (ϑ) which has a zero center point in the range from negative infinity to positive infinity. However, the graphs of analysis results with software based on the IRT model, *Xcalibre* shows a range of ϑ scale from -4 to 4 (Guyer & Thompson, 2011). However in real practice, Hambleton et al. (1991) have suggested the range of examinees for a test is at the value of -3 to 3. For dichotomous items, there are three IRT mathematical equations known as the 1PL, 2PL, and 3PL models. The main difference between the models is the number of parameters used to describe the item. Although the 1PL model is easiest than the IRT 3PL, the selection of models depends on the mathematical modeling of the model.

The ultimate goal of the IRT application is to predict the probability of an examinee with a certain level of ability to respond correctly to an item with the parameters of difficulty, discrimination and guessing parameters. But, this study is will be focused in IRT applications that allow research on the fairness of items in two different groups.

Frequently, built tests contain unnoticeably biased items to a particular group that can raise issues to test fairness. Therefore, an analysis that has a feature of detecting unnoticeably biased item is needed. Hambleton et al. (1991) found that one of the indispensable features of the IRT based analysis is its ability to conduct bias-element investigation at the item level. Specifically, one of the privilege of IRT is its ability to detect biased items against two different groups (e.g. men vs. women). Van der Linden & Hambleton (2010) states that the IRT model is able to carry out more in-depth analysis of biased items by evaluating the difference between alternative alternatives for examinees from different groups. In this study, elements of biased items will be assessed by comparing the gender aspects of the examinees (male vs. female).

In the Classic Test Theory (CTT), the element of a biased item for a subgroup is tested with a significant mean difference based on the value of *p*. This *p* value however is only a single value that applicable to a test (consisting all items) as a whole, not to every single items. However, an IRT-based analysis in detecting individually biased items is known as differential item functioning (DIF). DIF is defined as, an item indicates DIF if examinees with similar abilities, but from different groups (e.g. men vs. women), do not have the same probabilities to answer something correctly (Hambleton, et al., 1991). Operationally, Hambleton et al. (1991) also define that an item indicates DIF if the item's response function or item characteristic curve (ICC) is not identical across different groups. Psychometric studies of DIF are generally concerned with the question of whether an item is fair to members of some focal group as opposed to members of a reference group. An item is considered unequal if the item is equally difficult for an examinee from a focal group and a reference group that has the equivalent competency in a test (De Boeck & Wilson, 2004). By all means, a good item should be unbiased when the assessment process is done (Azrilah, Mohd Saidfudin, & Azami, 2013).

Among the weaknesses of the CTT statistical test as compared to IRT is that it requires the assumption of normal data distribution that is usually difficult to obtain. In contrast to

CTT, an analysis with an IRT application does not require a normal distribution assumption for examinees' scores or parameter items (DeMars, 2010).

DIF is an approach that has been widely used to identify biased items (Ogbebor & Onuka, 2013; Sharp et al., 2014). Besides IRT, there are several other methods for analyzing DIF such as Logistic Regression method using *SPSS* software (Abedlaziz et al., 2011; Ogbebor & Onuka, 2013), Transformed Item Difficulty (Abedlaziz et al., 2011), and Rasch model with *Winsteps* (Rosseni, et al., 2012). Sharp et al. (2014) in their study, analyses DIF using another IRT based software which known as *IRTPRO*.

The important fact that, most techniques for DIF assessment have been developed in an educational environment where items are generally dichotomous (Abedlaziz et al., 2011). Moreover, Davidov (2008) argues that with the existence of IRT, simpler techniques such as DIF analysis can be used to assess the equivalence of items or measurements as compared to previous techniques. This study has applied IRT-based software known as *Xcalibre* in calibrating research dichotomous data.

Ahmadi and Thompson (2012) remind, it should be noted that fit issues on the IRT model will cause IRT to not apply for DIF investigation. In fact, the analysis using *Xcalibre* and *Iteman 4* software including others IRT-based DIF analysis cannot be applied when the administered test is found to be beyond the ability of the examinee, if the test is in the form of speeded or if the examinee is penalized for the wrong response (Ahmadi & Thompson, 2012).

DIF is said to occur when the performance of an item is different among the two groups of examinees who have been sitting for a same test (Guyer & Thompson, 2013). Furthermore, the DIF analysis is able to show potential items having a bias characteristic on one group versus the other group. According to Guyer and Thompson (2013), there are actually many ways to assess DIF. Among them is by comparing the ICC parameters for the corresponding groups in which DIF is considered to exist if the ICCs of the two groups differ (Lord, 1980). However, the DIF analysis in this study will be conducted using *Mantel-Haenszel* (*M-H*) statistics as suggested by Guyer and Thompson (2013). With *M-H* statistics, each group is separated into several levels of abilities, and thus the probability of providing the correct response compared to the reference group and the focal group at each level of abilities. This is a major advantage of using IRT analyses. But, using more advance analysis like this will need more complicated calculation.

According to Guyer and Thompson (2013), the *M-H* odds ratio of the group score *k* is defined as Equation 1;

$$\alpha_k = \frac{C_{R_k} I_{F_k}}{C_{F_k} I_{R_k}} \quad (1)$$

Where;

C and *I* respectively are notations for correct and incorrect responses,
R represents reference group,
F represents focal group.

The *M-H* DIF coefficient is the weighted average of the odds ratios group score and is defined by Guyer and Thompson (2013) as Equation 2;

$$\hat{\alpha} = \frac{\sum_k \left(\frac{C_{F_k} I_{R_k}}{N}\right) \propto_k}{\sum_k \left(\frac{C_{F_k} I_{R_k}}{N}\right)} \quad (2)$$

Where *N* is the number of examinees in the group score *k*.

Objective

The purpose of this study is to identify whether there are gender-biased selections of multiple choice item or multiple choice question (MCQ) in the A&P test for the January-June 2013 sessions.

Methodology

Sampling

The sample of this study consisted of the Nursing Diploma students in MOH colleges who attended the MCQ test item for A&P subject in January-June 2013 examination session. The number of students who had been in the test was 971 in which 867 are female and 104 are male. All students were selected as samples because the statistical analysis method with IRT application in this study did not require random sampling assumptions. This is because the value of item parameters with IRT is not considered to be dependent on the ability of candidates to respond to the item (Baker, 2001). Therefore, random sampling is not required to make generalization decisions (Abdu Bichi, Embong, Mamat, & Maiwada, 2015).

In this study, the 3PL model has been used to analyze responses from different subgroups (female and male) to the various subjects of A&P subjects. The sample distribution of this study according by gender and colleges is shown in **Table 1**.

Table 1
Sample Distribution By Gender And Colleges

No.	Colleges	Female	Male	Total
1.	Alor Setar	65	0	65
2.	Sungai Petani	63	0	63
3.	Pulau Pinang	53	0	53
4.	SAS, Ipoh	84	28	112
5.	Sungai Buloh	66	15	81
6.	Kuala Pilah	38	0	38
7.	Melaka	47	0	47
8.	Muar	0	0	0
9.	Johor Bahru	58	14	72
10.	Kuantan	0	0	0
11.	Kuala Terengganu	46	0	46
12.	Kubang Kerian	56	0	56
13.	Kota Kinabalu	103	25	128
14.	Sandakan	88	0	88
15.	Kuching	100	22	122
16.	Sibu	0	0	0
	Total	867	104	971

Instrument

The research instrument is a set of MCQ of A&P subjects that have been administered on 971 students composing the Nursing Diploma program at the MOH Training Institution. A&P MCQ item consists of 40 items of various options covering six domains; Body Integration, Musculoskeletal System, Cardiovascular System, Respiratory System, Digestive System, and Integument System. The achievement of examinees for each item is scored dichotomously (1 = correct, 0 = incorrect).

Model Assumption

Unidimensionality is the most important assumption for all IRT models because when the assumptions of unidimensionality are met, then another assumption of local independence is also obtained (Lord, 1980; Hambleton, Swaminathan, & Rogers, 1991). Awopeju and Afolabi (2016) also remind that, institutions and researchers that wish to use IRT in solving measurement problems should make efforts to conform to the assumptions before use especially property of unidimensionality. That means, item response theory analysis can only be performed only when the test scores are unidimensional.

In this study, after being tested, the data were found to meet the unidimensionality and local independence assumptions that were important in the analysis with the IRT model.

IRT Model Selection

According to Embretson and Reise (2000), with the IRT model application, the value of -2LL (-2 times loglikelihood) can be used to assess the fit of the comparable models. In this study, the -2LL parameter is used to test and compare the fit between the 1PL, 2PL, and 3PL models. Smaller -2LL parameter values indicate better fit to the data (de Ayala, 2009; Embretson & Reise, 2000; Guyer & Thompson, 2013).

From the results of the analysis as shown in **Table 2**, the value of -2LL parameter for 3PL model is the smallest as compared to 2PL and 1PL. Hence, the dichotomous data of the multiple choice items of this study are more suitable to be calibrated using the 3PL model. That means, compared to the 1PL and 2PL models, the 3PL model provides better fit over the data.

Table 2
IRT Model evaluation based on -2LL statistic

Model	1PL	2PL	3PL
-2LL statistic	42403	41951	41947

Data Analysis

In this study, DIF analysis was conducted with the help of *Xcalibre* software. By using *Xcallibre* software, *M-H* coefficients will be reported for each item as odds ratio in the DIF analysis. The *M-H* coefficient is a weighted average of the odds ratios for each theta level. According to Guyer and Thompson (2013), if the odds ratio is less than 1.0, then the item is more likely to be answered correctly by the majority group than the minority. On the other hand, if the odds ratio value is greater than 1.0, it indicates that the minority group has the advantage of answering something correctly compared to the majority group. Items with a value of $p < 0.05$ indicate that there is a significant DIF and needs to be revised to determine whether there is a real issue of bias (Guyer & Thompson, 2013). In addition to the *Xcalibre* software that is

applied based on the 3PL model in this study, others soffware are also able to analyze DIF items such as *BILOG-MG* that have been applied by Ibrahim and Mohamed Najib (2009) as well as *PARSCALE* which have been applied by Young, Morgan, Rybinski, Steinberg, and Wang (2013) in their study.

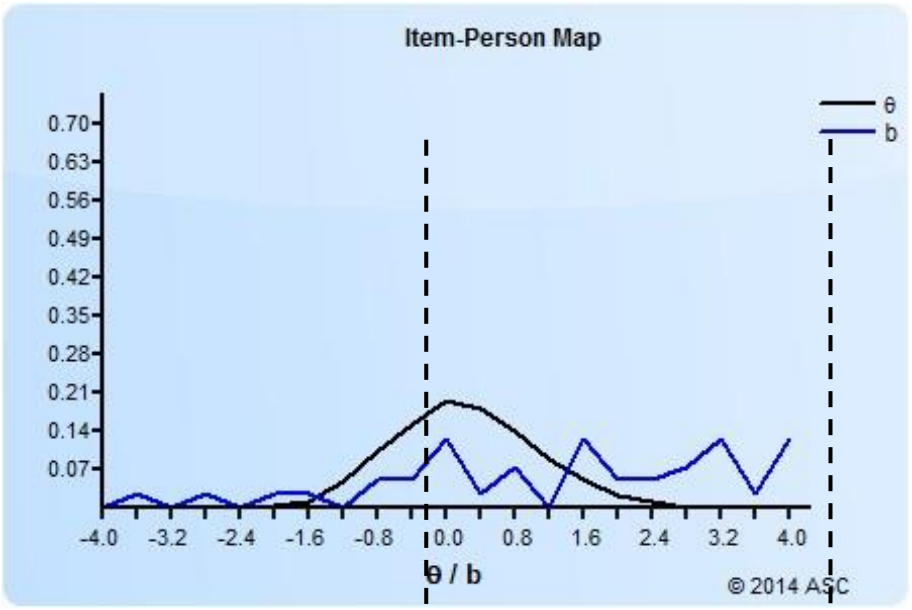
The DIF analysis attempts to show potential items that have a bias characteristic towards a group and DIF is said to exist when there is a difference in the performance of an item among the group of examinees who have been in the same test (Guyer & Thompson, 2013). This study has conducted a DIF analysis to detect whether there are biased items among prospective female and male students in the A&P test. The *Stats.csv* file of the analysis output have been reviewed from the calibration with the 3PL model.

Result

Among the examinees who attempted the A&P test, male examinees were a minority group as compared with female examinees who were the majority. As suggested by Schmitt and Kuljanin (2008), the proposed test items should be fair and impartial in any group.

However, calibration results on the January-June 2013 exam session data as illustrates in **Table 3** shows that none of the items gives advantage to male (minority) examinees. However, there were three items in the same test set, namely Item 17 (M-H = 0.28, $p < 0.05$), 18 (M-H = 0.51, $p < 0.05$) and 29 (M-H = 0.54, $p < 0.05$) that showing an bias evidence to female examinees.

Before the interpretation of the DIF analysis is made, items that have a fit issue on the model and are beyond the ability of the examinees should be noted as such items may cause IRT cannot be applied to investigate DIF (Ahmadi & Thompson, 2012). Therefore, item with fit issue and beyond the examinees' ability should be excluded from the DIF analysis. For this study, after excluding 13 items which were beyond the examinees' ability limit (including three items that misfit the model; Items 2, Item 32, and Item 38), there were 27 of 40 (67.5%) remaining items that found suitable for DIF investigations with the applications of IRT model. The calibration results (**Figure 1**) on the 27 items were obtained with three items in **Table 3** (Item 17, Item 18, and Item 29) show biased elements to female examinees. The DIF analysis results of Item 17, Item 18 and Item 29 are acceptable as they have no fit issues and those items are also within the limits of the examinees' ability. At the same time, the remaining 24 of 27 (88.9%) calibrated items for DIF are fair to both male and female students.



Outputs Range	Items		
	$b < \vartheta$	$-\vartheta \le b \le \vartheta$	$b > \vartheta$
$b \in (-3.887, 4.000)$ $\vartheta \in (-2.1839, 2.757)$	14, 25	1, 4, 6, 7, 8, 9, 10, 11, 12, 15, 16, 17, 18, 19, 20, 22, 23, 26, 28, 29, 30, 33, 34, 36, 37, 39, 40	2, 3, 5, 13, 21, 24, 27, 31, 32, 35, 38
	n = 2 (5%)	n = 27 (67.5%)	n = 11 (27.5%)

Figure 1 Item-Person Map vs. Items ID

Through items review, Item 17 is related to the artery name of ascending aorta that supplies blood to myocardium. Item 18 is about the name of the blood source structure pumped during ventricle systole, while Item 29 is related to the base layer name according to the sequence found in the gastrointestinal tract that starts from the inside out. That means, Item 1 and Item 18 are covered the subtopic of Cardiovascular System domain, while Item 29 is covered the subtopic Digestive System domain.

Table 3
DIF for A&P MCQ

Item ID	*M-H*	*M-H D*	*M-H SE*	*z-test*	*p*	*Bias Against*	*Theta 1 Odds-Ratio*	*Theta 2 Odds-Ratio*
1	1.00	-0.01	0.30	-0.01	0.99		0.82	1.23
4	1.29	-0.59	0.31	-0.81	0.42		1.25	1.33
6	0.73	0.75	0.30	1.07	0.29		0.92	0.55
7	1.24	-0.51	0.30	-0.72	0.47		1.50	1.03
8	0.98	0.05	0.34	0.06	0.95		1.29	0.53

Item ID	*M-H*	*M-H D*	*M-H SE*	*z-test*	*p*	*Bias Against*	*Theta 1 Odds-Ratio*	*Theta 2 Odds-Ratio*
9	0.80	0.53	0.32	0.70	0.48		0.71	0.96
10	1.05	-0.10	0.31	-0.14	0.89		0.72	1.52
11	0.67	0.94	0.34	1.17	0.24		0.71	0.60
12	1.28	-0.58	0.42	-0.59	0.56		1.13	2.04
15	1.58	-1.07	0.40	-1.14	0.26		1.32	2.28
16	1.01	-0.03	0.31	-0.04	0.97		0.71	1.40
17	0.28	2.98	0.34	3.71	0.00	Female	0.35	0.18
18	0.51	1.59	0.31	2.19	0.03	Female	0.60	0.39
19	1.00	0.01	0.30	0.01	0.99		1.18	0.86
20	1.46	-0.88	0.35	-1.07	0.29		1.57	1.39
22	1.07	-0.16	0.33	-0.20	0.84		0.90	1.42
23	1.27	-0.56	0.34	-0.71	0.48		0.74	1.98
26	1.29	-0.61	0.39	-0.66	0.51		1.71	0.25
28	1.12	-0.26	0.30	-0.36	0.72		1.12	1.11
29	0.54	1.43	0.29	2.12	0.03	Female	0.84	0.32
30	1.74	-1.30	0.35	-1.57	0.12		1.85	1.66
33	0.67	0.94	0.32	1.26	0.21		0.74	0.60
34	1.37	-0.74	0.34	-0.94	0.35		1.36	1.39
36	1.19	-0.42	0.32	-0.55	0.58		0.95	1.43
37	1.46	-0.89	0.31	-1.21	0.23		1.44	1.48
39	1.15	-0.32	0.35	-0.39	0.70		3.19	0.71
40	1.11	-0.25	0.31	-0.35	0.73		1.02	1.21

Female = 867; Male = 104

Discussion

The results of the study found that female's groups had the advantage of responding to Item 17, Item 18, and Item 29 correctly compared to the male examinees group. The advantages of female examinees may be due to the content of the items that require more reading besides memorizing the facts. This finding is consistent with the results of the study by Zalizan, Saemah, Roselan, and Jamil (2005) where they find that female students have an advantage in assignments that require memorization of facts. While the A&P subjects are well-known for topics that require a lot of memorization, it is undeniable that there are certain subtopics (e.g. Cardiovascular System and Digestive System) that involve complex fact-finding rather than other subtopics.

Since female examinees have shown that they have the advantage of Item 17 and Item 18 (subtopic of Cardiovascular System) as well as Item 29 (subtopic of Digestive System), therefore lecturers can use the advantages of female students in helping male students especially in Cardiovascular System and Digestive System topics. In this regard, the findings of this study not only tell about items with biased issues, but moreover, they can also inform educators about gender advantages over a subtopic so that mutual benefits and sustainability in learning can be obtained.

Conclusion

As a whole with applying DIF analysis, waiving items that have issues with examinees ability, misfit, and bias, there are 88.9% remaining items that does not indicate the problem as biased

item. This means that most of the items that have been enacted are fair to female and male examinees although male examinees are known as a minority group in the field of nursing studies. Based on a large number of items that does not show item bias, this study can generalize that the subject of A&P MCQ is ideal to be administered not only for female examinees, but it also suitable for male examinees.

However, reverting to the original purpose of an administered test when involving two groups (majority and minority), it must be fair. With the IRT model application, Ibrahim and Mohamed Najib (2009) recommend that items with bias elements be removed from the test set. Before being excluded, Guyer and Thompson (2013) stated that items showing significant DIF need to be revised to determine whether there is a true bias issue. When it is clear that there are items with biased issues against any of the subgroups studied, then it should be removed. These are also supported by Azrilah et al. (2013) where items need to be reviewed or considered for drops if there is a biased issue against a group or there is a group that is more successful in doing a task than the other group. In many cases, the biased item can be reviewed and improve.

Since this study has emphasized that item developers need to be aware of the possibility that there is a biased item in the test, item developers should ensure the minority group is any test treated fairly. Therefore, the main contribution of this study is to highlight that DIF analysis is necessary in analyzing dichotomous items particularly if involving two groups (majority and minority). The reality of bias item is a phenomenon that needs to be acknowledged and through the application of the IRT model shown in this study, it is clear that biased items can be easily identified.

References

Abdu Bichi, A., Embong, R., Mamat, M., & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: A review of empirical studies. *Australian Journal of Basic and Applied Sciences*, 549-556.

Adedoyin, O. O. (2010). Using IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in Mathematics. *Educational Research and Reviews, 5*(7), 385-399.

Ahmadi, A., & Thompson, N. A. (2012). Issues affecting item response theory fit in language assessment: A study of differential item functioning in the Iranian National University Entrance Exam. *Journal of Language Teaching and Research, 3*(3), 401-412.

Awopeju, O. A., & Afolabi, E. R. (2016). Comparative analysis of classical test theory and item response theory based item parameter estimates of senior school certificate mathematics examination. *European Scientific Journal, 12*(28), 263-284.

Azrilah, A., Mohd Saidfudin, M., & Azami, Z. (2013). *Asas Model Pengukuran Rasch: Pembentukan Skala & Struktur Pengukuran.* Bangi: Universiti Kebangsaan Malaysia.

Baker, F. B. (2001). *The Basic of Item Response Theory.* ERIC.

Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the European social survey. *Survey Research Methods, 2*(1), 33-46.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory.* New York: The Guilford Press.

De Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach.* USA: Springer.

DeMars, C. (2010). *Item Response Theory: Understanding Statistic Measurement.* New York: Oxford University Press, Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Guyer, R., & Thompson, N. (2011). *Item Response Theory Parameter Recovery Using Xcalibre™ 4.1.* Saint Paul, MN: Assessment Systems Corporation.

Guyer, R., & Thompson, N. A. (2013). *User's Manual for Xcalibre™ Item Response Theory Calibration Software, Version 4.2.* Woodbury MN: Assessment Systems Corporation.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* USA: Sage Publications, Inc.

Ibrahim, M., & Mohamed Najib, A. (2009). The analysis of Iran universities 2003-2004 entrance examination to detect biased items. *Jurnal Teknologi, 50 (E)*, 21-27.

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problem.* N.J., Erlbaum: Hillsdale.

McArthur, D. L. (1981). *Detection of Item Bias Using Analyses of Item Response Patterns.* Los Angeles: Graduate School of Education, University of California.

Moss-Racusin, C. A., Dovidio, J. F., Brescoll, V. L., Graham, M. J., & Handelsman, J. (2012). Science faculty subtle gender biases favor male students. *Proceedings of the National Academy of Sciences, 109*(41), 16474-16479.

Ogbebor, U., & Onuka, A. (2013). Differential item functioning method as an item bias indicator. *Educational Research, 4*(4), 367-373.

Rosseni, D., M Faisal, K., Helmi, N., Parilah, M., Aidah, A., Nur Ayu, J., & Verawati. (2012). Detecting gender biasness via gender differential item functioning analysis on integrated meaningful hybrid e-learning instrument. *WSEAS Transactions on Advances in Engineering Education, 9*(3), 63-71.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*, 210-222.

Sharp, C., Michonski, J., Steinberg, L., Fowler, J. C., Frueh, B. C., & Oldham, J. M. (2014). An investigation of differential item functioning across gender of BPD criteria. *Journal of Abnormal Psychology, 123*(1), 231–236.

van der Linden, W. J., & Hambleton, R. K. (2010). *Handbook of Modern Item Response Theory.* New York: Springer-Verlag New York Inc.

Young, J., Morgan, R., Rybinski, P., Steinberg, J., & Wang, Y. (2013). *Assessing the Test Information Function and Differential Item Functioning for the TOEFL Junior Standard Test.* Princeton, New Jersey: Educational Testing Service (ETS).

Zalizan, M. J., Saemah, R., Roselan, B., & Jamil, A. (2005). Prestasi Akademik Mengikut Gender. *Jurnal Pendidikan Malaysia, 30*, 93-111.