JRIIE Journal of Research Innovation and Implications in Education
Center for Research Implications & Practice

Website:www.jriiejournal.com ISSN 2520-7504 (Online) Vol.9, Iss.3, 2025 (pp. 333 - 342)

Gender Differential Item Functioning on Entrance Examination: A Case of Malamulo College of Health Sciences

Victor Peter Nkungula
Basic Sciences Department, Malawi Adventist University,
Malamulo College of Health Sciences, Makwasa, Malawi
Email: nkungulav@mchs.adventist.org

***Abstract**: Differential Item Functioning (DIF) analysis is a key element in evaluating fairness and validity of tests. Gender is cited as a source of construct irrelevant variance. It plays a significant role in determining the volume of students who pass the tertiary level entrance examinations. By so doing, it causes bias and consequently challenges the inference made based on the instrument used for selection of the prospective students. This study aimed at investigating presence of DIF in terms of gender (an issue to do with examination validation) on an entrance examination for Malamulo College of Health Sciences, as triggered by observed low scores of female examinees. The participants (n = 615) were sampled randomly from examinees who were to sit for the 2017-2018 Malawi School Certificate of Education examinations. X-Calibre 4.2 software was used to produce item parameters (a, b and c) which were used in Raju formula. Gender DIF was detected in 77% of the items, biased towards male examinees. On the basis of the findings, it is concluded that the entrance examinations' test scores were not free of construct-irrelevant variance and the overall fairness of the test was likely compromised at it showed that it was heavily biased towards male examinees. Analysis of DIF items inform use and interpretation of test. This is dependent on item writing, review and the continuous analysis of the items to improve the validity of the instrument(s). By so doing, the examinees are protected and the information that is needed from the examinations is attained.*

Key words: Gender, Item, Bias, Impact, Differential Item Functioning, Validity.

How to cite this work (APA):

Nkungula, V. P. (2025). Gender differential item functioning on entrance examination: A case of Malamulo College of Health Sciences. *Journal of Research Innovation and Implications in Education, 9*(3), 333 – 342. **https://doi.org/10.59765/jriie.9.3.32**.

1. Introduction

Educational measurement and assessment aims at making correct and appropriate decisions for individuals. In most cases, tests (examinations) are used to guide performance-related decisions to be taken. Tests should be practical, valid and reliable, which may be affected by errors and bias. Item bias appears when individuals having the same level of skill have different possibilities of answering an item due to their certain characteristics (Zumbo, 1999). Detection of bias can be done by finding test items where one group performs much better than the other group: such items function differentially for the two groups, and this is known as Differential Item Functioning (DIF).

Any DIF study has two groups: the focal group, mostly the potentially disadvantaged group and the reference group, the group which is considered to be potentially advantaged by the test (McNamara & Roever, 2006; Angoff, 1993).

Whenever an item is flagged as displaying DIF, the source of DIF should be investigated to see if it is biased or not. Any item flagged as showing DIF is biased if, and only if, the source of variance is irrelevant to the construct being measured by the test (Messick, 1989, 1994), for the sake of making fair decisions.

Group differences on item performance that represent a difference in the construct measured are traditionally referred to as impact, representing a construct-relevant difference (Camilli & Shepard, 1994). Avoiding the confounding of "impact" and DIF has been and is still a permanent concern in item bias research. DIF analysis seeks to flag items for potential bias by identifying items on which differential group performance, beyond that expected by true group differences, is observed. Distinguishing DIF from impact and determining whether DIF items are measuring the intended construct or not, are fundamental validity issues in the pursuit of fairness in testing (Gomez-Benito, et al, 2017).

Item bias due to administration of examinations that are not validated may be a stumbling block to the endeavor of giving equal educational opportunities to males and females (SDG-goal 4 and MGDS III-key priority area 6.3). Poorly constructed or unvalidated tests can inaccurately reflect student knowledge. Validity improves the trustworthiness of the results obtained from examinations. Examination bias is a reality (McCollough, 2011), as such, no stone should be left unturned as it affects the performance of candidates.

This study therefore makes an investigation as to whether the gender of the examinee contributes to their performance, by exploring the possible biased items through DIF approach. Furthermore, the study is an addition to the limited local literature about gender DIF. This gives room for assessing the validity (and consequently, improving the reliability) of the testing instrument and it is a wakeup call to examiners to have in mind other forms of bias that might interfere with the quality of examinations beyond those that are used for selection purposes, for instance, continuous assessment tasks and end of semester examinations.

1.1Research Questions

1. To what extent do items in the entrance examinations function differentially by gender?
2. Is there any difference in the number of items that may function differentially by gender in favor of males and those in favor of females?

2. Literature Review

Differential Item Functioning (DIF) analysis is a key element in evaluating the fairness and validity of tests. The *Standards for Educational and Psychological Testing* (2014) (here after *Standards*) present DIF as an indicator of the extent to which different groups of test takers (in this case, males and females) who are at the same ability level have different frequencies of correct responses or, in some cases, different rates of choosing various item option. Gender is frequently cited as a source of construct irrelevant variance (Alavi & Bordbar, 2018), Bordbar & Alavi, 2020), thereby playing a significant role in determining the volume of examinees who pass examinations. By so doing, it causes bias and consequently challenges the intention or inference made based on the instrument.

2.1 DIF Investigation as an Exercise for Validation

DIF analyses may serve as validity evidence in the validation process of a test. The *Standards* refer to validity as the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests (*Standards*, 2014). The validation process is considered as an empirical evaluation of the meaning and consequences of measurement. This shows that validation combines scientific inquiry as well as rational arguments to justify (or nullify) score interpretation and use (Messick, 1995).

DIF analyses are closely related to construct-irrelevant variance which indicates that the test measures too many variables, many of which are irrelevant to the interpreted construct, which can make a test to be too easy or too difficult for examinees. Holland and Wainer (1993) advocate that this is a major source of bias when interpreting test scores. DIF analysis in this case provides evidence of possible invalid performance differences to alert test developers of potential bias in favour of a particular group on a test.

2.2 The ICC and Area Methods for DIF Detection

Educational measurement concerns underlying (latent) variables of interest and involves determining how much of such a latent trait a person possesses (Hambleton, 1994; Hambleton & Slater, 1997). A correct response depends on both the characteristics of the item and the person's ability. The probability of a correct response is expressed as a mathematical function of examinee ability and item characteristics – also known as the item characteristic curve (ICC).

The ICC graphically represents the regression of the item score on examinees' ability, which is known as the item response function. This function is plotted with the ability level of examinees along the x-axis (θ), against the probability of answering an item correctly on the y-axis (P(θ)). The graph takes the typical S-shaped form of the ICC (see Figure 1). Each item will have its own ICC.

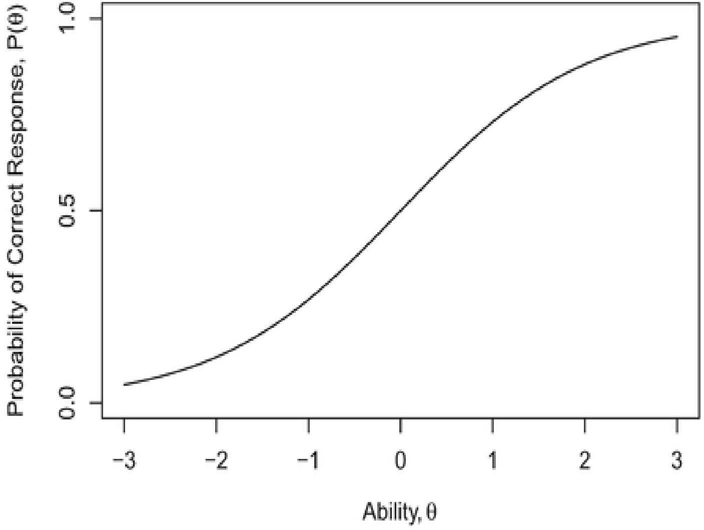

Figure 1: An example of an ICC (Adopted from De Beer, 2004, pp. 53)

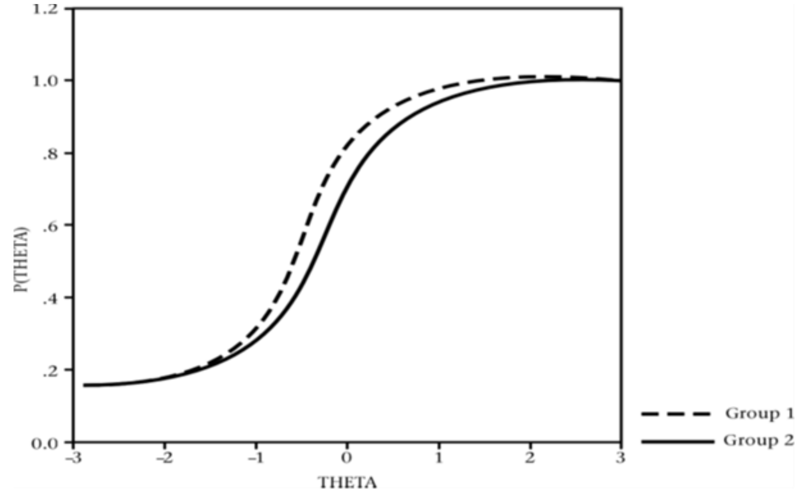

Figure 2: ICCs showing a question with DIF (Adopted from De Beer, 2004, pp. 55)

Area methods for DIF detection focus on comparing the area between ICCs, estimated for reference and focal groups after establishing a common or base metric (Raju, 1990). An item does not exhibit DIF when the reference group and the focal group have identical ICCs placed on the same scale. However, when the ICCs of two groups are different, the gap (area) between the two ICCs indicates an item with DIF (see Figure 2). The magnitude of the area can thus be determined through the Raju formula.

2.3 Malamulo College of Health Sciences Examination

Malamulo College of Health Sciences (MCHS), a constituent College of Malawi Adventist University, produces graduates who work in various health facilities as Medical Assistants, Clinical Officers, Medical Laboratory Technicians, Nurses or Public Health professionals. Every year, advertisements are made for people with at least Malawi Schools Certificate Education (MSCE) certificate who make applications to join the programme of their choice. Due to active enrollment of students who have to undergo a selection process guided among others by a tool for selection, usually an examination, MCHS gives a good platform for the study of item bias. The validity of the examination has to be sure as this is a high-stakes examination which decides the admission of successful students into tertiary level training. MCHS therefore has to strive in its endeavors to improve for the better, through continued analysis and validation of its tools used for selection of prospective students.

3. Methodology

3.1 Research Approach

The study followed a quantitative research tradition, because it gave room for the collected data to be subjected to statistical analysis as required by the research questions. For alternative or comparative analysis replication would be possible, and generalization and inference could be made from the analyzed results of the sample to the general population.

3.2. Design and Methodology

Non-experimental descriptive design was utilized, in an attempt to control threats to internal validity. Questions whether or not group differences exist were responded to. The conclusions that have been drawn are primarily descriptive in nature.

3.3 Participants

615 participants (330 females and 285 males) were randomly sampled from a population of 965 (confidence interval 2.38, confidence level 95%) to which the instrument was administered. Their role was to respond to test items which were given to them, that collectively made an examination.

3.4 Instrumentation and Administration

The main instrument was the 2017-2018 entrance examination paper for Malamulo College of Health Sciences. It contained a total of 60 multiple choice questions. The instrument was administered to the participants like any other closed book examination. Each student worked independently to find the solutions to the given questions.

3.5 Data Handling and Analysis

Answer sheets were marked by hand. Microsoft Excel 2013 application software was used to produce a dichotomous score sheet for all participants and to capture their actual responses. These were used on the SPSS version 20 application software to produce input and control files to be used on X-Calibre 4.2 application software and the computation of chi-square test (Fishers' exact test) to determine differences of DIF flagged items. Furthermore, the X-Calibre 4.2 was used to determine item parameters (a, b and c parameters) for DIF calculation.

3.6 Determination of DIF Items

The Xcalibre 4.2 output file(s), containing the item discrimination (a), difficulty (b) and pseudo guessing (c) parameters were used to detect DIF. The "a" and "b" parameters were substituted into the Raju formula, for calculation of DIF. The Raju formula is presented as follows:

$$\text{Area} = \left|2\frac{a2-a1}{Da1a2}ln\left[1+e^{Da1a2\frac{b2-b1}{a2-a1}}\right]-(b2-b1)\right|$$

Where: a1: discrimination parameter for males (reference group)

a2: discrimination parameter for females (focal group)

b1: difficulty parameter for males (reference group)

b2: difficulty parameter for females (focal group)

D = 1.7 (constant: scaling factor)

An item is said to possess DIF when the area index is greater than a critical value of 0.22, while an item does not possess DIF when the area index is zero or close to zero. Ling and Lau (2004).

The "c" parameter was used as a guide to detect possible DIF, having in mind that a group with significant "c" values (> 0.5) might be advantaged over the other, and that might be a source of bias.

3.7Determination of Group Differences in the Number of Flagged DIF Items

Differences between DIF flagged items were determined using a Chi-square test (Fisher's exact test) to determine if there was any difference (statistical significance) in the number of items functioning differentially by gender:

H_0: There is no significant difference in the number of items functioning differentially by gender in favour of males and those in favour of females.
H_1: There is significant difference in the number of items functioning differentially by gender in favour of males and those in favour of females

3.8 Ethical Considerations

Consent was obtained from all participants, the respondents to the examination. Their role, research methodology and possible outcomes were made clear. They did not give any form of identification as they responded to the questions with anonymity. They were only requested to indicate their sex, as this was crucial to enable the identification of male and female respondents. The scripts obtained were analyzed with confidentiality. There was no risk of harm to the participants and no other coercive or deceptive practices were used. The participants were at liberty to withdraw from the research at any point. They were not required to respond to all items given in the examination. Furthermore, authorization was obtained from MCHS for the use of the entrance examination paper.

4. Results and Discussion

4.1 Gender DIF Items in the Entrance Examinations

Research question one (1) focused on the extent to which items in the entrance examinations function differentially by gender. Table 1 shows the area indices for the examination items 2017-18 academic year entrance examinations for MCHS. It displays the items that exhibit DIF and the group it favoured. A total of 44 items were subjected to the DIF analysis. Sixteen (16) items were dropped (from a total of 60) as they had various forms of flags. These items were therefore deemed not fit for analysis.

The results in the table reveal that 34 items representing 77.3% functioned differently by gender with area indices greater than the critical value of 0.22 and 10 items representing 22.7% did not function differentially with area indices less than 0.22.

Of the 34 items that functioned differentially by gender, 28 items representing 82.4% were in favour of male examinees while 6 items representing 17.6% were in favour of the female examinees. In other words, 6 items (17.6%) functioned against the male examinees and 28 items (82.4%) functioned against the female examinees. This shows that male examinees were at an advantage of scoring better than their female counterparts, as they could ably respond to 86.4% of the items in the examination instrument (including the 10 non-DIF items).

Table 1: Summary of Area Indices of 2017-18 academic year entrance examinations for MCHS

SEQUENCE	ITEM	a1	a2	b1	b2	AREA INDEX	DECISION	FAVOURED GROUP
1	1	0.4693	0.3888	1.5516	1.7265	0.525	DIF	Male
2	3	0.4264	0.5012	0.6105	0.6532	0.204	NO DIF	
3	4	0.3793	0.4388	-1.2429	-1.5716	1.105	DIF	Female
4	5	0.4634	0.6772	-0.7139	0.2317	0.774	DIF	Male
5	6	0.4942	0.7255	0.3645	1.0708	0.728	DIF	Male
6	7	0.5388	0.8745	-2.7545	-1.5324	0.630	DIF	Male
7	9	0.7396	0.9518	-0.9985	-0.5798	0.297	DIF	Male
8	10	0.9698	1.0787	2.3105	2.6661	0.013	NO DIF	
9	12	0.8584	1.0117	2.5316	2.9686	0.103	NO DIF	
10	13	0.5463	1.035	-0.5354	0.4975	0.875	DIF	Male
11	14	0.5265	0.8559	0.8663	0.9108	0.175	NO DIF	
12	15	0.907	0.915	3.0257	3.0324	0.010	NO DIF	
13	16	0.5229	0.4529	1.1216	1.3417	0.675	DIF	Male
14	17	0.8872	0.7744	0.2583	1.2497	1.630	DIF	Male
15	18	0.605	0.7587	1.0489	1.2165	0.312	DIF	Male
16	19	0.6056	0.7484	1.5161	1.7066	0.312	DIF	Male
17	20	0.7182	0.7118	-2.6843	-1.706	1.065	DIF	Male
18	22	0.674	0.5767	-0.9753	-1.1226	0.256	DIF	Female
19	23	0.6469	0.9807	0.3518	0.475	0.332	DIF	Male
20	27	0.4156	0.3582	-1.0235	-1.2902	0.615	DIF	Female
21	28	0.6142	0.6393	1.1659	1.682	0.237	DIF	Male
22	30	0.7054	0.9249	2.9312	2.7029	0.532	DIF	Female
23	31	0.8109	0.6446	0.7835	1.1334	0.880	DIF	Male
24	32	0.581	0.5429	0.9199	0.8661	0.121	NO DIF	
25	33	0.7634	0.9318	-0.3771	0.3525	0.034	NO DIF	
26	35	0.8851	0.8246	0.8481	2.1845	1.758	DIF	Male
27	38	0.9813	0.9587	1.8948	2.9285	1.187	DIF	Male
28	40	0.8065	0.808	2.9144	3.2965	0.364	DIF	Male
29	41	0.7528	1.0315	1.2258	1.7148	0.366	DIF	Male
30	42	0.5909	0.4817	0.9829	0.9987	0.105	NO DIF	
31	43	0.7854	0.6759	2.0543	2.4212	0.848	DIF	Male
32	44	0.8307	0.6781	1.0239	1.3514	0.813	DIF	Male
33	47	0.7528	0.8135	0.9209	1.4269	0.132	NO DIF	
34	48	0.5133	0.5482	1.2514	1.1614	0.258	DIF	Female
35	49	0.8316	0.9712	2.5046	3.0642	0.014	NO DIF	
36	51	0.8807	0.8127	-0.8405	0.0493	1.311	DIF	Male
37	52	0.7125	0.6302	0.2489	0.5139	0.648	DIF	Male
38	53	0.5001	0.3578	-0.5298	0.1624	2.303	DIF	Male
39	54	0.6135	0.5857	-0.2423	0.2242	0.776	DIF	Male
40	55	0.6979	0.704	2.3048	2.125	0.241	DIF	Female
41	56	0.6401	0.4652	-0.0307	0.2191	0.467	DIF	Male
42	57	0.423	0.4322	0.6164	1.5738	0.664	DIF	Male
43	58	0.5827	0.7773	1.6926	2.0217	0.448	DIF	Male
44	60	0.5623	0.4939	0.3801	0.838	1.117	DIF	Male

4.2 Difference in Number of Flagged Items between Genders

Research question two (2) focused on determining if there was any significant difference in the number of items functioning differentially by gender in favor of males and those in favor of females to test the null hypothesis and alternative hypotheses as presented in the methodology.

Analysis of summary statistics for all calibrated items show that the items did not give room for students to guess the correct response. The calculated pseudo guessing parameter (c) values were 0.354 and 0.314 for males and females respectively. These were below the threshold value of 0.500. As such, any other possible bias was not related to guessing of the correct response.

Table 2 shows a chi-square (Fishers exact test statistic) result where 82.4% (28/34) of items favoring male examinees is compared to 17.6% (6/34) for those favoring females (p = 0.00), at alpha level of 0.05. As such, there was significant difference in the number of items functioning differentially by gender in favour of males and those in favour of females. Male examinees were at an advantage of performing better that the female examinees.

Table 2: Chi-square Test Summary of Differential Item Functioning in Favour of Male and Female Students

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	34.000[a]	1	.000		
Continuity Correction[b]	27.467	1	.000		
Likelihood Ratio	31.688	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	33.000	1	.000		
N of Valid Cases	34				

a. 3 cells (75.0%) have expected count less than 5. The minimum expected count is 1.06.

b. Computed only for a 2x2 table

The distribution of the DIF items according to subjects of study and gender is presented in Figure 3. It should be noted that the equality in the number of DIF flagged items for Physical Science and Biology and for English Language and Mathematics was a mere coincidence.

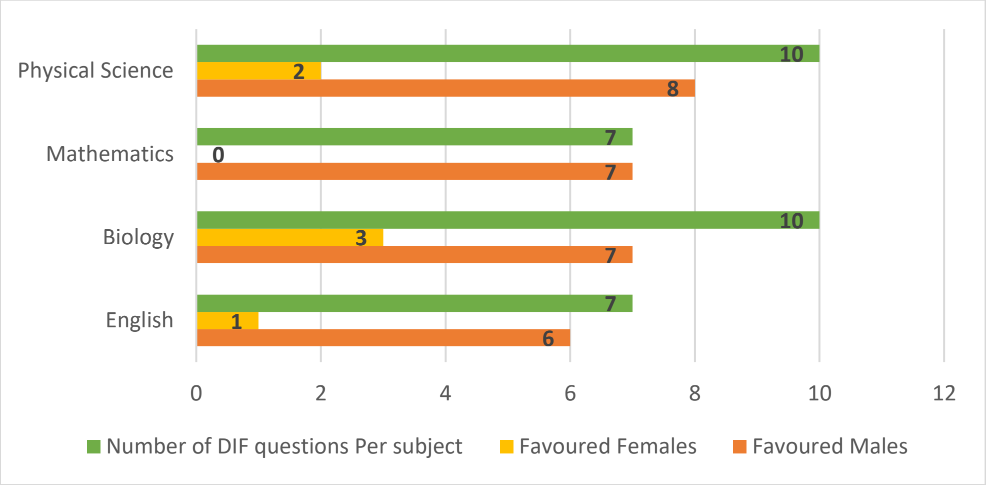

Figure 3: Distribution of the DIF items by subject of study and sex

4.3 Relationship of Findings to Prior Research

The results that were obtained by Kalaycioglu & Berberoglu (2011) from a DIF analysis of an entrance examination containing science and mathematics items are similar to what this study has obtained, where most items favoured male examinees. Science items (Biology and Chemistry) that favoured females were those that basically required the examinees to recall information, whilst those that required more elaborative skills (like Physics items) favoured male examinees. Similar results were also obtained by studies that were conducted by Mendes-Barnett & Ercikan (2006) and Zenisky et al., (2004).

In terms of the number of items that were flagged as having DIF, there is a contradictory result between this research and that which was conducted by Ahmadi & Bazvand (2016), who found equal numbers of DIF items for the focal and reference groups, leading to a concept of DIF cancellation. However, Pae & Park (2006) and Zumbo (2003) observe that DIF cancellation is a complicated issue that depends on the number of items indicating DIF as well as the magnitude of DIF.

5. Conclusion and Recommendations

5.1 Conclusion

This study analyzed the entrance examination for the year 2017-2018 for Malamulo College of Health Sciences for gender DIF. Most items analyzed suggested that male examinees were favoured for and female examinees were favoured against. Male examinees might have an advantage of performing better than their female counterparts. The validity of the instrument was likely compromised, thereby affecting female examinees negatively. In as much dimensions as are considered when selecting examinees who are considered to have passed an entrance examination, the tool for testing (the instrument, in this case, the examination) plays a greater role, as such, its contribution to the final decision made cannot be underestimated.

The entrance examination is a high-stake examination. Due to bias, some candidates (especially females) might have not been selected for training (were disadvantaged), which is against the principle of fairness in examinations. Results from this study will help to guide other researchers in identifying items that can disadvantage examinees. These results have implications for item writing, item review, and the continuing analysis of the items as an evidence of improving validity of the instruments as a whole.

5.2 Recommendations

Literature proposes different methods for detecting items with possible DIF. Interestingly, the procedures give results which may be contradictory or not (Ercikan, et al, 2004; Kim & Cohen, 1995). DIF detection is therefore dependent on the procedure used. It is therefore recommended to use complementary procedures for more reliable findings and consequently, interpretations. Luckly enough, no matter which method is taken into consideration, comparison of the probabilities of correct responses of the students at the same ability level but belonging to different groups is the major approach in detecting items as DIF.

Researchers have questioned and challenged the parameter of gender as a fixed, binary variable, (e.g. Sunderland, 2000). It is claimed that rather than being a fixed, biological variable, gender is predominantly a socially constructed variable within specific cultural and situational contexts (Davis & Skilton-Sylvester, 2004). In other words, different results are likely to be obtained if such factors are considered when making the DIF analysis. Breland & Lee (2007) propose studies that classify the examinees into several major cultural, national, and or educational subgroups and conduct a separate gender DIF study within each of these subgroups.

Furthermore, future DIF analyses should consider using 4 Parameter Logistic Model (4PLM). Currently, the fourth parameter (the slip parameter) is rarely used, because the interpretation of the higher asymptote (on ICC) regarding slip, carelessness, disinterest or boredom is questionable. Further research may be employed to clear this questionable mist. Nevertheless, the 4PLM gives room for explaining why respondents with high ability may miss the correct endorsement to an easy item.

References

Ahmadi, A & Bazvand, A. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research*. Retrieved from www.urmia.ac.ir/ijltr

Alavi, S., & Bordbar, S. (2018). Differential Item Functioning Analysis of High-Stakes Test in Terms of Gender: A Rasch Model Approach. *MOJES: Malaysian Online Journal of Educational Sciences, 5(1),* 10-24. Retrieved from https://mojes.um.edu.my/article/view/12631

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

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning.* Hillsdale, NJ: Lawrence Elrbaum Associates.

Bordbar, S. & Alavi, S. M. (2020). Investigating gender-biased items in a high-stakes language proficiency test: Using the Rasch model measurement, *Applied Linguistics Research Journal, 4(5)*: 1–21.

Breland, H., & Lee, Y.W. (2007). Investigating uniform and non-uniform gender DIF in computer-based ESL writing assessment. *Applied Measurement in Education*, *20*, 377- 403.

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. United States: SAGE Publications.

Davis, K. A., & Skilton-Sylvester, E. (2004). Looking back, taking stock, moving forward: Investigating gender in TESOL. *TESOL Quarterly, 38(3),* 381– 404.

De Beer, M. (2004). Use of Differential Item Functioning (DIF) Analysis for Bias Analysis in Test Construction. *SA Journal of Industrial Psychology*, *30 (4),* 52-58

Ercikan, K., Gierl, M. J., Mc Creith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*, 301-321.

Gomez-Benito, J., Sireci, S., Padilla, J., Hildago, D & Benitex, I. (2017). Differential Item Functioning: Beyond Validity Evidence Based on Internal Structure. *Psicotherma (30):1*, 104-109

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment,10 (3),* 229-244.

Hambleton, R.K. & Slater, S.C. (1997). Item response theory models and testing practices: Current international status and future directions. *European Journal of Psychological Assessment,13 (1),* 21-28.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum

Kalaycioglu, D.B. & Berberoglu, G. (2011). Differential Item Functioning Analysis of the Science and Mathematics Items in the University Entrance Examinations in Turkey. *Journal of Psychoeducational Assessment, 29(5), 467– 478.*

Kim, S., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detecting of differential item functioning. *Applied Measurement in Education, 8*, 291-312.

Ling, S. E and Lau, S. H. (2004). *Detecting differential item functioning (DIF) in standardized multiple-choice test: An application of item response theory (IRT) using three parameter logistic model.* Journal of Applied *Psychology, 94 (7)*, 452-459.

Malawi Government (2017, August). *Building a Productive, Competent and Resilient Nation: The Malawi Growth and Development Strategy III* (Report). Lilongwe: Ministry of Economic Planning & Development.

McCullough, L. (2011). Women's Leadership in Science, Technology, Engineering and Mathematics: Barriers to Participation. *Forum on public policy online, 2011(2).* Retrieved from https://files.eric.ed.gov/fulltext/EJ944199.pdf

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics Assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19(1)*, 289-304.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23(2).*

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' performance as science inquiry into score meaning. *American Psychologist, 50(9)*, 741-749.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement, Edition 3.* New York: American Council on Education & Macmillan.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned area between two item response functions. *Applied Psychological Measurement, 14(1),*197-207

Sunderland, J. (2000). Issues of language and gender in second and foreign language education. *Language Teaching, 33*, 203–223.

UN General Assembly (2015, October). *Transforming our world: The 2030 Agenda for Sustainable Development*. Retrieved from http://www.refworld.org/docid/57b6e3e44.html

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2004). DIF detection and interpretation in large-scale science assessments: Informing item writing practices. *Educational Assessment, 9(1)*, 61-78.

Zumbo, B. D. (2003). Does item-level DIF manifest itself in scale-level analyses? Implications for translating language tests. *Language Testing, 20(2),* 136-147.

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