Gender-Related Differential Item Functioning of 2015 Wassce Core Mathematics Results in Southern Ghana Using Logistic Regression Procedure

Ruth Annan-Brew, PhD
Andrews Cobbinah, PhD
Department of Education and Psychology,
University of Cape Coast, Cape Coast, Ghana

Doi:10.19044/esj.2020.v16n16p188 URL:http://dx.doi.org/10.19044/esj.2020.v16n16p188

Abstract

Differential item functioning (DIF) occurs when individuals of the same ability level from separate groups have a different probability of answering an item correctly. This study was conducted in two parts: in the first part a real 2015 West African Senior Secondary Certificate Examination (WASSCE) core mathematics test data were analyzed for uniform and non-uniform DIF using binary logistic regression (LR) procedure and in the second part, content analysis of items identified as DIF were classified under the levels of the cognitive domain by experts. Three research questions were formulated for the study. A sample of 4,285 male and 3,712 female candidates were selected from a population of 15,258 candidates who sat for the examination in 2015 from 20 selected schools in Southern Ghana. The instrument for the study was the 50 multiple-choice core mathematics items. The findings showed that there was 43 significant gender differential item functioning items of which 9 were uniform and 34 non-uniform. Also, the content analysis revealed that items that favoured males were mainly number and numeration, algebraic processes, probability and statistics and mensuration whiles plane geometry and coordinate geometry revealed DIF in favour of females. It was concluded that test items used were not free from gender DIF. It was recommended that DIF studies should be conducted by test developers in order to be review or exclude DIF items to enhance fairness in assessment.

Keywords: Differential item functioning, Southern Ghana, Logistic regression, Content analysis, West African Senior Secondary Certificate Examination

Introduction

Reliability and validity are two characteristics that all measurement instruments must have, including educational and psychological tests. The American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council on Measurement in Education, NCME (1999) define "validity as the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). Thus, any test parameter that is different between two or more subpopulation groups, like item difficulty or test length, maybe a sign of a threat to test validity because the test results would need different interpretations for each group. In this context, differential item functioning (DIF) becomes an important validity and bias issue of test analysis.

Notwithstanding, the consistency of test results can be affected by the test takers' demographic characteristics. Every test taker belongs to a subgroup. Test Takers' answer to items tends to be influenced by their membership.

Standardized tests and measurements are used mainly to discriminate between ability levels of examinees. As a part of the determination o0f validity for these tests, differential item analysis is employed to evaluate the degree to which measurements discriminate true abilities among examinees in an impartial manner. Psychometricians and test developers are to use differential item functioning (DIF) analysis to determine if there is a possible bias in a given test or examination. DIF involves a two-step process: The first step is the comparison of two groups' outcome on an item and determining the presence of DIF and the second step includes a decision of whether there is a large enough difference between the groups to eliminate or change the item of interest.

Differential item functioning (DIF) is "an indicator of bias observed when test takers from different groups have different probability or likelihood of responding correctly to an item, after controlling for ability" (Awuor, 2008). DIF occurs when individuals from different subgroups have unequal expected item scores or matching on the primary trait, attribute, or ability the test is intended to measure (Kilmen, 2016). For example, the situation of child development, females tend to develop fine motor skills at an earlier age than males do. Males, on the other hand, tend to outperform females when using gross motor skills. In that regard, items or activities that require the use of gross motor skills would display differential item functioning for males and females and that constitutes an example of the actual difference. We might also consider the issue of crying. In some cultures, crying is considered an acceptable way of showing pains for males and in others, it is not. Therefore, any item related to crying could be understood differently in one subculture than in another. Items that refer to crying would likely demonstrate differential item functioning when responded by individuals of various subcultures.

Differential item functioning (DIF) procedures are currently the dominant psychometric methods for addressing fairness in standardized achievement, aptitude, certification, and licensure testing (Clauser & Mazor, 1998; Millsap & Everson, 1993). These procedures reflect, in large part, a response to the legal and ethical need to ensure that comparable examinees are treated equally. Generally, examinees are split into two groups namely the reference and focal groups. The reference group consists of majority or advantaged group members and the focal group consists of minority or disadvantaged group members. DIF analysis, then, involves matching members of the reference and focal groups on a measure of ability and implementing statistical procedures to identify group differences on test items. These group differences may take two forms. Most DIF procedures are designed to identify uniform (unidirectional) DIF, which occurs when an item favours one group over another throughout the ability continuum. Occasionally, DIF procedures may identify non-uniform (crossing)DIF, which occurs when there is an Ability × Group Membership interaction, but generally DIF procedures are not designed to do so. Swaminathan and Rogers (1990) applied the logistic regression (LR) procedure to DIF detection. This was a response, in part, to the belief that the identification of both uniform and non-uniform DIF was important. The strengths of this procedure are well documented. It is a flexible model-based approach designed specifically to detect uniform and non-uniform DIF with the capability to accommodate continuous and multiple ability estimates. Furthermore, simulation studies have demonstrated comparable power in the detection of uniform and superior power in the detection of non-uniform DIF compared to the Mantel–Haenszel (MH) and Simultaneous Item Bias Test (SIB) procedures (Li & Stout, 1996; Jodoin & Gierl, 2001; Narayanon & Swaminathan, 1996).

Ong, Williams and Lamprianou (2015) explored crossing differential item functioning (DIF) in a test drawn from a national examination of mathematics for 11-year-old pupils in England. An empirical dataset was analyzed to explore DIF by gender in a mathematics assessment. A two-step process involving the logistic regression (LR) procedure for detecting uniform and nonuniform DIF was applied to identify crossing DIF. The results showed 36 uniform and 19 nonuniform statistically significant gender DIF items. Out of the 19 nonuniform DIF items, 10 items were crossing DIF. Their study was consistent with Abedalaziz (2010) who also used LR method to identify DIF on the mathematical ability scale for 30 items. Eighteen items or 60% of the items revealed DIF of which 10 were uniform DIF and 8 non-uniform DIF.

Research has repeatedly reported gender differences in mathematics performance on several standardized mathematics tests such as the SAT-M

(Scholastic Assessment Test-Mathematics) (Willingham & Cole, 1997; Song, Cheng & Klinger, 2015). The test scores on these standardized tests have been regarded as an important measure of abilities to do mathematics problems (Halpern, 2000; Stumpf & Stanley, 1998). However, results from these studies are not consistent: Halpern (2000) found that males generally outperformed females on mathematical tasks while (Voyer, Voyer, & Bryden, 1995) found that differences exist based on gender depending on the types of mathematical tasks. Hyde, Fennema and Lamon (1990) suggested that there was a very small or null gender difference in mathematical ability on these tests. Caplan and Caplan (2005) even argued that the link between gender and mathematical ability was very weak. Battista (1990) conducted a study of 145 high school geometry students from middle-class communities. This research examined the role that spatial visualization and verbal-logical thinking played in gender differences in geometric problem-solving in high school. The findings suggested that males and females differed in the level of discrepancy between spatial and verbal abilities. Gallagher, De lisi, Holst, McGillicuddy-De Lisi, Morely, and Cahalan (2000) suggested that males tend to be more flexible than females in applying solution strategies. This study, therefore, sorts to provide an opportunity to examine issues in gender-related differential item functioning of 2015 WASSCE core mathematics in specific and for the logistic regression procedure in the detection of DIF.

Research Questions

The following research questions guided the study:

1. What is the nature of those items identified as exhibiting uniform and non-uniform DIF?
2. How do gender differences link to content areas within mathematics?
3. What is the nature of the cognitive ability level of those items identified as showing DIF?

Method

This study aims to determine if the items in 2015 WASSCE core mathematics exhibited item bias about the variation of gender. Since the research was intended for determining an existing situation, it employed the descriptive design. The research population was the examinees who schooled in the southern part of the country and took the 2015 WASSCE exam. Secondary data from WASSCE was used for the analysis. A simple random technique was used to select examinees from southern Ghana. The sample consisted of 4,285 males and 3,712 females.

The data used in this study were obtained from WAEC. In determining the differential item functioning, Mantel Haenszel (MH) and logistic

regression (LR) methods, were considered. These two techniques are based on classical test theory. According to (Clauser & Mazor, 1998; Monahan et al., 2007; Swanson, Clauser, Case, Nungester, & Featherman, 2002; Hidalgo & Lopez-Pina, 2004), the LR is the most preferred method for determination of DIF because it has easy application and statistical interpretation, gives effective results for small groups and is most efficient to determine uniform and non-uniform DIF. Therefore, the LR was preferred at determination phase of DIF in this study. The STATA programme was used for the LR analysis.
LR uses the examinee as the unit of analysis and has the following form: Let L_1, L_2, and L_3 be the log-likelihood values associated with the following models, respectively,

$$\text{logit}\ \{\Pr(y)\} = \tau_0 + \tau_1 t + \tau_2 g + \tau_3 (t \times g) \quad (1)$$

$$\text{logit}\ \{\Pr(y)\} = \tau_0 + \tau_1 t + \tau_2 g \quad (2)$$

$$\text{logit}\ \{\Pr(y)\} = \tau_0 + \tau_1 t \quad (3)$$

where y is a vector of responses for a given item; t is the latent trait, most commonly represented by the observed total score; and g is a dichotomous variable representing the focal group.

In the present study, likelihood-ratio tests are used to compare the nested models. The test for non-uniform DIF compares models that is, equations 1 and equation 2 which is given by $LR_1 = -2\ (L_1 - L_2)$. The logistic regression, LR_1 is distributed as χ^2_1. If the null hypothesis of no nonuniform DIF is rejected, we do not proceed to the test for uniform DIF. The test for uniform DIF compares equations 2 and 3 and is given by $LR_2 = -2\ (L_2 - L_3)$. The logistic regression, LR_2 is distributed as χ^2_1. The item reveals DIF in favour of males when the significant odd ratio is greater than one, whereas the item reveals DIF in favour of females when the significant odd ratio is less than one ($\alpha = 0.05$).

Also, the effect size that was used in this study was based on Dorans (2004) classification {i.e., -2.35ln (odds ratio)}. In this classification, three main categories are used. Category “A” depicts items with negligible or nonsignificant DIF which is defined by LR-D-DIF and not significantly different from zero or an absolute value less than 1.0. Category “B” depicts items with slight to the moderate magnitude of statistically significant DIF which is also defined by LR-D-DIF significantly different from zero and an absolute value of at least 1.0 and either less than 1.5 or not significantly greater than 1.0. Category “C” depicts items with moderate to large magnitude of statistically significant DIF and defined by the absolute value of LR-D-DIF of at least 1.5 and significantly greater than 1.0.

Results and Discussion

Nature of items identified as exhibiting uniform and non-uniform DIF

Table 1 shows the summary results of the LR method to identify DIF on the 2015 WASSCE core mathematics exams for each of 50 items. Forty-three (43) items or 86% of the items revealed DIF. Out of this forty-three (43) items, nine (9) items revealed statistically significant uniform DIF, whereas the thirty-four (34) items revealed statistically significant non-uniform DIF. This finding is in support of Ong, Williams and Lampriaou (2015) who found out of 60 items which showed DIF, that 36 were uniform while 19 showed non-uniform. Again, this current finding confirms that of Abedalaziz (2010) who had 10 uniform and 8 non-uniform items out of 30 items which showed DIF. This current finding perhaps is so probably because of the settings and content areas studied.

Gender differences and it link to content areas within mathematics

Again, from table 1, seven items (5, 12, 14, 17, 22, 25 and 39) out of 50 were DIF free, thus these items did not function differently among examinees being it male or female. The items 1, 2, 3, 4, 7, 8, 9, 15, 16, 19, 20, 23, 27, 29, 31, 35, 37, 38, 42, 44, 45, 46, 49 and 50 were in favour of males, whereas the items 6, 10, 11, 13, 18, 21, 24, 26, 28, 30, 32, 33, 34, 36, 40, 41, 43, 46, 47 and 48 were in favour of females.

Table 1: Summary Results of the LR analysis

	Non-uniform		Uniform		(a)		
Item	χ^2	Prob.	χ^2	Prob.	Odds	-2.35ln(a)	ETS
					Ratio		
1	24.86	0.0000			1.9567	-1.5775	C
2	7034	0.0067			6.9943	-4.5710	C
3			19.26	0.0000	1.2867	-0.5924	A
4	580.59	0.0000			1.8626	-1.4616	B
5			0.56	0.4537			
6	87.65	0.0000			0.0671	6.3487	C
7			50.57	0.0000	1.7375	-1.2983	B
8	49.81	0.0000			1.9306	-1.5459	C
9	138.47	0.0000			8.4073	-5.0034	C
10	157.34	0.0000			0.2798	2.9931	C
11	241.94	0.0000			0.5250	1.5142	C
12			0.41	0.5209			
13	11.31	0.0008			0.6952	0.8544	A
14			0.92	0.3384			
15	4.03	0.0446			1.2867	-0.5924	A
16	47.52	0.0000			1.3396	-0.6871	A
17			0.27	0.6063			
18	29.76	0.0000			0.6698	0.9418	A
19			107.27	0.0000	2.9166	-2.5155	C
20	96.38	0.0000			2.0035	-1.6330	C
21	265.95	0.0000			0.5315	1.4853	B

22			2.25	0.1339			
23	93.85	0.0000			6.4727	-4.3888	C
24	71.89	0.0000			0.1085	5.2194	C
25			1.42	0.2326			
26	231.55	0.0000			0.5307	1.4889	B
27			144.02	0.0000	3.0366	-2.6102	C
28	97.22	0.0000			0.2391	3.3626	C
29	64.36	0.0000			1.8867	-1.4918	B
30	136.56	0.0000			0.1441	4.5525	C
31	160.64	0.0000			4.5225	-3.5463	C
32			129.06	0.0000	0.3179	2.6931	C
33	77.56	0.0000			0.0849	5.7958	C
34	3.98	0.0459			0.3992	2.1580	C
35	146.88	0.0000			3.0405	-2.6133	C
36			316.26	0.0000	0.1747	4.1000	C
37			62.58	0.0000	3.4821	-2.9319	C
38	15.32	0.0001			1.9451		C
						-1.5635	
39			3.24	.0720			
40	120.34	0.0000			0.4963	1.6464	C
41	8.09	0.0045			0.3184	2.6895	C
42	10.84	0.0010			6.6763	-4.4616	C
43			170.15	0.0000	0.2669	3.1041	C
44	184.81	0.0000			2.0276	-1.6611	C
45	158.32	0.0000			2.885	-2.4899	C
46			49.15	0.0000	0.2456	3.2995	C
47	452.29	0.0000			0.4626	1.8116	C
48	9.76	0.0018			0.3283	2.6175	C
49	71.18	0.0000			2.1060	-1.7503	C
50	491.98	0.0000			2.0475	-1.6841	C

The nine items that revealed statistically significant uniform DIF had 5 in favour of males and 4 in favour of females, whereas the thirty-four items that revealed statistically significant non-uniform DIF had 18 in favour of males and 16 in favour of females. With effect size measure based on Dorans (2004) classification system, there were 1 negligible, 1 moderate and 7 large uniform whiles nonuniform DIF items identified were measured as 4 negligible, 4 moderate and 26 large items.

According to the results of the LR analysis that was conducted to see if the item function of the 2015 WASSCE core mathematics exams changed regarding the gender difference, it was found that 1 item at the category “A”, 1 item at the category “B” and 7 items at the category “C” with a total of 9 items out of the 50 questions included uniform DIF. It is seen that 3 out of the 7 items containing uniform DIF at the category “C” worked in favour of males. Also, for the 33 non-uniform DIF items identified, there were 4 items at category ‘A’, 4 at category ‘B’ and 26 at category ‘C’ out of which 14 favours male examinees. In general, using the effect size of B and C, it is seen that 20 out of 50 items indicated DIF in favour of males whiles 18 out of 50 items exhibited DIF in favour of females.

Looking at the first step the content analysis which was to describe the content and skills of the items that favoured males or females. It can be stated that the items working in favour of males were about numbers and numeration, algebraic processes, mensuration and statistics and probability whereas the items functioning in favour of females were plane geometry and coordinate geometry of straight lines. These results are consistent with those found in earlier gender studies of multiple-choice tests where items that measure reasoning and problem solving generally favoured males.

The findings are inconsistent with the findings obtained from the study by Demirtasli (2015). where male students are more successful in items about nature in the field of mathematics and the questions in the field of geometry and included DIF in favour of male students shows parallelism with the findings obtained from the research of Abedalaziz (2010) on the investigation of the differential item functioning according to variation of the gender of the items in mathematics tests. This finding contradicts the finding obtained from that previous study of Ding, Song and Richardson (2007) who emphasized that male students were more successful at primary education level, whereas female students were more successful at secondary education or university level, especially in problem-solving and application. It can be also seen that the 23rd question, which measured trigonometry in the same test worked in favour of males.

Nature of the cognitive ability level of those items identified as showing DIF

The third step of content analysis was to examine the items that were not flagged for DIF to determine whether flagged items represented differences in the cognitive ability of examinees to use mathematical processes. Even though, studies of differential item functioning have been done in terms of gender and also the nature of items identified as exhibiting uniform and non-uniform DIF. The nature of the cognitive ability level of those items identified as showing DIF has not yet been reported in the literature. However, it was found in this study that items that worked in favour of males were at the knowledge and comprehensive levels whereas females outperformed males in items that functioned at the application and analysis levels of the cognitive ability.

Conclusion

This study provides evidence that there are gender differences in performance on test items in core mathematics that vary according to content even when content is closely tied to the curriculum. Furthermore, assuming that males performed better on algebraic processes, mensuration and numeration system is an indication of reliance on algorithmic learning. Females on the other hand might profit even more than males from an instructional strategy that relies less on teaching algorithms and more on teaching problem solving and effective means of approaching non-routine problems. The study also indicated that items that worked in favour of males were at the knowledge and comprehensive levels whereas females outperformed males in items that functioned at the application and analysis levels of the cognitive ability.

The presence of differential item functioning is a serious threat which affects the validity of test items or test scores which must have kept some candidates at a disadvantaged position. Most candidates who aspired to study science-oriented courses at the University or any tertiary institutions have been denied admission or must have found themselves into programmes they never applied for.

Recommendations

It is recommended that all examination bodies, test experts in WAEC and people charged with the responsibility of developing, validating and administering of test need to carry out differential item functioning analysis for all items before administering the test. During teaching, illustrations should be drawn from the learners' environment owing to the diversified background of learners while students should ensure that they make adequate preparation for their examinations. Finally, teachers should ensure adequate coverage of their curriculum and boys and girls should be given the same opportunity and treatment as well as same challenges in the mathematics class.

References :

1. Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: an IRT-based montecarlo study with SIBTEST and Mantel-Haenszel procedures* (Doctoral dissertation, Virginia Tech).
2. Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education, 14*(4), 329-349.
3. American Educational Research Association (AERA), American Psychological Association (APA) & the National Council on Measurement in Education (NCME) (1999), Standards for educational and psychological testing, AERA Publications.

4. Abedalaziz, N. (2010). A Gender-related differential item functioning of mathematics test items. *The International Journal of Educational and Psychological Assessment, 5,* 101-116.
5. Battista, M. T. (1990). Spatial visualization and gender differences in high school geometry. *Journal for Research in Mathematics Education, 21(1),* 47-60.
6. Caplan, J. B., & Caplan, P. J. (2005). The perseverative search for sex differences in mathematics abilities. In A. M. Gallagher, & J. C. Kaufman (Eds.), Gender differences in mathematics: An integrative psychological approach. Cambridge: Cambridge University Press.
7. Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and practice, 17*(1), 31-44.
8. Ding. C. S., Song, K., & Richardson, L. (2007). Do mathematical gender differences continue? A longitudinal study of gender difference and excellence in mathematics performance in the U.S. Educational Studies 40 (3), 279-295.
9. Dorans, N. J. (2004). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement, 41*(1), 43–68.
10. Demirtasli, Ç. (2015). A Study on Detecting of Differential Item Functioning of PISA 2006 Science Literacy Items in Turkish and American Samples. *Eurasian Journal of Educational Research*, *58*, 41-60.
11. Gallagher, A. M., De lisi, R., Holst, P. C., McGillicuddy-De Lisi, A. V., Morely, M., & Cahalan, C. (2000). Gender differences in advanced problem-solving. *Journal of Experimental Child Psychology, 75,* 165-190.
12. Halpern, D. F. (2000). Sex differences in cognitive abilities (3rd Ed.).
13. Mahwah, N. J.: Lawrence Erlbaum Associates.
14. Hidalgo, M. D., & Lopez-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between Logistic Regression and Mantel-Haenszel Procedures. *Educational and Psychological Measurement, 64*(6), 903-915.
15. Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107(2), 139-155.
16. Kilmen, S. (2016). Effect of DIF Magnitudes, Focal Group Sample Size, and DIF Ratio on the Performance of SIBTEST. *International Journal of Social Sciences and Education*, *6*(1), 91-98.
17. Li, H. H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, (4), 647-677.

18. Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement*, *17*(4), 297-334.
19. Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. *Journal of Educational and Behavioral Statistics, 32*(1), 92–109
20. Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274.
21. Ong, Y. M., Williams, J., & Lamprianou, I. (2015). Exploring crossing differential item functioning by gender in mathematics assessment. *International Journal of Testing*, *15*(4), 337-355.
22. Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test. *Language Testing and Assessment, 4*(1), 97–124.
23. Stumpf, H., & Stanley, J. C. (1998). Standardized tests: Still gender-biased?
24. Current Directions in Psychological Science, 7, 335-344.
25. Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.
26. Swanson, D. B., Clauser, B. E., Case, S. M., Nungster, R. J., & Featherman,
27. C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics, 27*(1), 53–75.
28. Voyer, D., Voyer, S., & Bryden, M. P. (1995). The magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. Psychological Bulletin, 117, 250-270.
29. Willingham, W. W., & Cole, N. S. (1997). Gender and fair assessment. Lawrence Erlbaum Associates, Publishers.