Jurnal Evaluasi dan Pembelajaran
Volume 5 Nomor 1, Tahun 2022
Available online at https://jepjurnal.stkipalitb.ac.id/index.php/hepi

Differential Item Functioning of Performance-Based Assessment in Mathematics for Senior High Schools

Abraham Gyamfi
Wesley College of Education, Kumasi, Ghana
E-mail: abrahamgyamfi84@gmail.com

Abstract

The purpose of the study was to validate performance-based assessment in mathematics (Quantitative Reasoning) for Senior High Schools for Differential Item Functioning (DIF). The study sought to found out if the five-items on a newly developed performance-based assessment in mathematics (quantitative reasoning items) has DIF. The study employed descriptive research design embedded in Graded Response Model (GRM). Stratified, census, simple random sampling and purposive samplings procedures were employed to select 750 SHS Three students in the Western Region from three categories of SHS. The Performance-based assessment test was used as the main data collection instruments. Data were analyzed with winGEN and independent t test using SPSS. It was also found that with the exception of the linear equation item that showed DIF for category A and C schools, there was no presence of DIF in the items for both gender and category of school. Based on the findings, it was recommended that performance-based assessment should be an integral part in the methods of assessment lessons and courses at both the colleges of education and universities and that major method of assessment strategy in teaching and learning of mathematics.

Keywords: *Differential item functioning, focal group, reference group, performance-based assessment, mathematics, quantitative reasoning*

INTRODUCTION

The ability to design a performance assessment in such a way that they can be presented on an individual student's level has been one of the challenges associated with performance assessments (Pegg, 2003). In today's classrooms, children exhibit wide range of abilities, and the teacher has the mandate of teaching these children. It is thinkable that a well-designed performance assessment could be used and still fail to provide relevant data if the assessment task is either too difficult or too easy for the student being assessed.

Classroom assessments in mathematics have faced a series of challenges to students' achievement in relations to PBA (Etsey & Abu, 2013). These challenges have been listed by Gao (2012) to include a focus on recall of isolated items of knowledge. To improve student achievement, mathematics, Gao (2012) suggested that assessment should be fused into planned instruction and relate to the students' real-world experiences.

Performance-based assessment as a contemporary form of assessment is perceived to address many of the challenges associated with the traditional assessment. The focus of performance-based assessment has to do with application of knowledge. According to Nitko (2004), PBA is a form of assessment that presents a hand on task which requires students to perform an activity that calls for application of the knowledge and skills from several learning. It allows students to show how well they have learnt. In its simplest term,

a PBA is an assessment which demands students' demonstration of the specific skills and competencies they have mastered by performing or producing something. Ainsworth and Viegut (2006) defined performance assessments as an "activity that requires students to construct a response, create a product, or perform a demonstration" (p.57). Performance assessment deals with the overall experience of a student in performing a learning target through the application of their knowledge and skills from several areas. Performance assessment also lends itself to multiple products to a task therefore resulting in multiple correct responses. The nature of performance-based assessment in mathematics for this study is what Shavelson, et al (2020) termed as Quantitative Reasoning (QR).

QR is seen as a "competence in interacting with myriad mathematical and statistical representations of the real world in contexts of daily life, work situations and the civic life" (Karaali, Hernandez & Taylor, 2016. P. 25). Shavelson (2008) posited that QR lean toward application with an emphasis on skills involved in dealing with messy, complex, real-world, everyday challenges with to a greater or lesser extent quantities and their various representations. Like all performance-based assessment, there is the application of knowledge and skills in real world situation. The difference therefore between PA and QR is that while PA could apply to any subject, QR is limited to mathematical and statistical concepts.

In validation of the traditional mathematics items in Ghanaian SHS, the psychometric properties of the computational items are not known. However, Brennan (2006) stated that the strength of an assessment procedure lies in its ability to meet acceptable psychometric values such as difficulty, discrimination, reliability and biases or differential item functioning (DIF). These psychometric values are better estimated using generalizability and item response theories.

Item response theory (IRT) on the other hand, allows the estimation of students' ability from any set of the items. Item response theory allows the difficulty and discrimination levels of each item on the test to be estimated. In the framework of IRT, item characteristics are independent of the sample and latent traits of the person are independent of the test on the account that the selected models perfectly fit the data. Therefore, scores that describe examinee performance are independent on test difficulty. The scores of the examinee may be lower on a difficult test and higher on easier tests, but the ability level of the examinee remains the same over any test at the time of testing or surveying (Le, 2013). The model of IRT allows the estimation of four item parameters (difficulty, discrimination, guessing and ceiling effect), DIF and the ability levels of the students on each item.

Item Response Theory (IRT) is also a statistical tool for estimating the difficulty and discrimination indices as well Differential Item Functioning (IRT). For multiple choice items, the classical approach, where the difficulty and discrimination indices are estimated with taking the ability levels of the examinees into consideration. The classical assumes that ability level is constant for all examinees. Again, the classical approach is not able to estimate the parameters for polytomous items or graded response items (Brennan, 2006). Le (2013) stated that some of the principal psychometrics of performance-based assessment could be estimated using a model, 1PLM, 2PLM, 3PLM or 4PLM of the IRT.

The model representing difficulty, discrimination, guessing and ceiling effect of each item to students' performance on the assessment for polytomous object (graded responses or multidimensional items) as in the case of performance-based assessment.

Using the IRT to estimate the item parameters of a performance-based assessment provide the best option to validate the newly developed performance-based assessment. This is to ensure that the newly developed performance-based assessment for Senior High Schools in the Western Region of Ghana could produce valid and reliable results. Performance-based assessment in mathematics have been proven by Arhin (2015); Brennan (2006) and Burkhardt and Swan (2008) to provide feedback to students that stimulates learning and also has positive effect on students learning. It also helps students to equip themselves for the WASSCE. In addition, students are able to apply knowledge in mathematics to real life situation. This calls for the development of performance-based assessment for Senior High Schools in Ghana.

Purpose of the Study

The general purpose of the study is to determine whether the newly developed performance-based assessment for Senior High Schools in the Western Region of Ghana exhibit significant differential item functioning base of gender and category of school.

Research Questions

The following research questions were formulated to guide the study:

1. Are there any DIF in the items on the performance-based assessment in respect to gender and category of school?

Differential Item Functioning (DIF)

Differential item functioning (DIF) refers to the extent to which an item functions differently for a group of the same ability level on a particular trait that is being measured (Linn, 2003). DIF is an "unforeseen difference among groups of examinees who are expected to be the same with respect to the trait being measured by the item and the entire test" (Dorans & Holland, 1993). "DIF is a statistical property, which states that matched-ability groups have differential probabilities of success on an item" (Annan-Brew, 2020, pg 57). The expectation of DIF is that "members of the two groups be compared on the important underlying ability before determining whether members of the two groups have different probability of responding correctly to the item" (Annan-Brew, 2020, pg 67).

Differential item functioning (DIF) happens when examinees belonging to different groups (gender or ethnicity or location) with the same underlying ability level have a different probability of responding correctly to an item or response to the item in a particular way (Bruckner, Forster, Zlatkin-Troitschanskaia, Happ, Walstad, Yamaoka & Asano, 2015). Group differences in item responses (or on latent variables) are not considered as DIF per se (for example when girls score higher than boys on mathematics items). However, DIF is present only when the girls and boys have the same ability level, yet the girls' scores are higher than the boys. DIF is present when examinees from different groups show different probabilities of responding correctly (or endorsing) the item after their underlying ability that the item is purported to measure is matched.

"DIF refers to differences in the functioning of items across groups, oftentimes demographic, which are matched on the latent trait or more generally the attribute being

measured by the items or test" (Cho & Cohen, 2010, pg. 5). When examining items for DIF, it is important to note that the groups must be matched to the same ability level on the measured attribute, otherwise, the result in the detection of DIF may be misleading (Camilli, 2006).

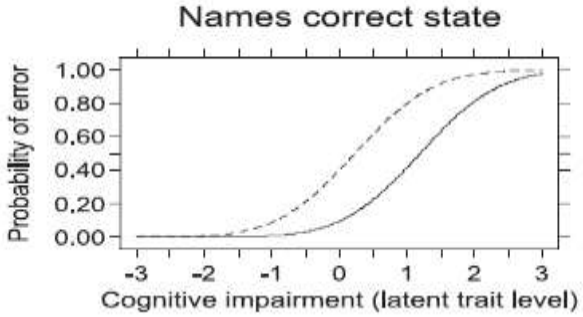
Types of DIF

According to Dorans and Holland (1993), there are two types of DIF: uniform and non-uniform.

Uniform

DIF occurs uniformly at all levels along the latent trait. Systematically, the item is more difficult for members of one group, even when examinees have the ability (theta). It is shown by a shift in b-parameter as shown in Figure 5.

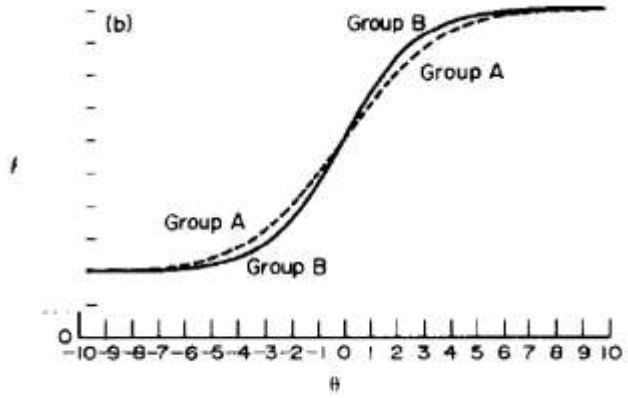
Figure 1 –*Uniform DIF*



Non-Uniform

DIF does not occur equally at all points on the latent trait (gender differences in response) may only be evident at high or low levels of the construct. The change in item difficulty does not follow a regular pattern across the ability spectrum. The increase/decrease in P for low-ability examinees is counterbalanced by the opposite for high-ability examinees. It is shown by a shift in a (and possibly b) as shown in Figure 6.

Figure 2-*Non-uniform DIF*



DIF Detection Methods

There are two main approaches to detect whether an item has DIF or not; Non-parametric and Parametric methods.

Non-parametric methods

These methods are particularly useful when the sample sizes are small for the groups of interest (Camilli, 2006). Most of them use contingency tables.

1. Mantel-Haenszel procedures
2. Chi-square methods
3. Proportion difference measure (Standardization)
4. Simultaneous Bias Test (SIBTEST)

Parametric methods

1. Logistic regression
2. Likelihood ratio test
3. Item response theory methods
4. Log linear models

Mantel-Haenszel

According to Clauser and Mazor, (1998), the Mantel-Haenszel (MH) test is the widely used method for detecting uniform DIF. The method conditions on raw score and statistical test of contingency tables. It has been proved to be effective with reliable statistical test and effect size. A popular software (SAS, SPSS) for the analysis is also available. However, "it does not test for non-uniform DIF, unlike LogR", before analysis could be carried matching score need to be put into bins (levels). "This is somewhat arbitrary and may affect the statistical decision regarding DIF".

Logistics regression

Logistic regression is "based on statistical modeling of the probability of responding correctly to an item by group membership (reference group and focal group) and a criterion or conditioning variable". The method conditions on raw score and models group-response relationship Clauser & Mazor, 1998). The logistics regression has "multivariate matching criteria (multidimensional matching), it can test for both uniform and non-uniform DIF and significance test (like IRT likelihood ratio) as well as effect size (T_2)". There are popular software (SAS, SPSS) available for use. The limitations (relative to IRT) are that it uses observed score metric and may need to purify matching score for scales with few items.

IRT DIF

These approaches compare item parameters, or ICCs, and condition on ability. IRT DIF analyses incorporate data from both groups to estimate cparameters, fix the c-parameters, and estimate a and b parameters in each group from different calibrations. Clauser and Mazor (1998) stated that if a multivariate DIF test is performed, two fully different calibrations may be performed (albeit this is not usually the case).

The ICCs in the two groups must be compared on the same scale before they may be compared (equated). Estimates of item parameters from the Focal group (minority) are often placed on the scale of the Reference group (majority). The reference group is the group that is thought to have a benefit, while the focal group is the group that is thought to be harmed by the program. The point is that if different ICCs occur after being matched

on ability, the item displays DIF. That is, if the group's chances of success are not the same. Even though one group is smarter than the other, there is no DIF if matched ability examinees have the same chance of responding correctly to the item.

IRT DIF Detection

There are two main methods for detecting DIF; compare Item Parameter Estimates and area methods.

1. ***comparing Item Parameter Estimates***: Using the comparison of item parameter method, a multivariate test to estimate b, a, and, c and a t-tests perform on b-values to find out there is a significant difference in the probability of success between the two groups.
2. ***Area Methods:*** here the Total Area, squared Differences and the weighted Areas and Differences are used in the comparison. The area methods "compare an ICC from one group against an ICC from the other and look at how much area is between the two", hence more common.

Importance of DIF analysis

Clauser and Mazor (1998) listed the following as the importance to DIF analysis in test development and validation:

1. Important first step in the evaluation of test bias
2. For construct validity items of a scale ideally should have little or no DIF
3. Items should function in the same way across subgroups of respondents who have the same underlying ability (or level on the latent trait)
4. Presence of DIF may compromise comparison across subgroups – give misleading results
5. Confound interpretation of observed variables

Graded Response Model (GRM)

The Graded Response Model (GRM) by Samejima (as cited in Park, 2012) belongs to the cumulative approach where all categories of scores are used to quantify the probability of success or failure (de Ayala, 2009). The GRM estimates probabilities based on the specification of 2PL. Separate b_i parameters are estimated for each step of the item. However, it uses one a_i parameter for all steps for each item. The GRM indicates m-1 "boundary" response functions which are an indication of the cumulative probability for a response category greater than the option of interest. It is represented by the equation:

$$Pij(\theta) = \frac{\exp\left[ai(\theta - bij)\right]}{1+\exp\left[ai(\theta - bij)\right]}$$

The reason for using GRM, or any model based on ordered response categories, with testlet-based scores (group of items based on the same or similar content developed as a unit with predetermined procedures that the examinee may follow) is that, theoretically, testlet-based scores can have an ordered quality if scores "correspond to the degree of completeness of the examinee's reasoning process within a testlet" (de Ayala, 2009, p. 58). Table 1 shows the summary of the IRT models.

METHODOLOGY

Research Design

A descriptive research design embedded in the Graded Response Model (GRM) of IRT was used for the study. The GRM -a two parameter model for polytomous items was used to describe the location (difficulty level) and slope (discrimination) of the PBA items developed by the researcher.

Population

The population for the study were all public SHS three students in in the western region of Ghana. There are 7498 SHS 3 from 35 SHSs in the region as at 2019. The accessible population comprise SHS 3 students and mathematics teachers selected from 30 SHS selected for the study. The accessible population also included the mathematics examiners in the region, WEAC mathematics zonal leader and assessment experts.

Sample and Sampling Procedures

A multistage sampling procedure was used for the selection of respondents for the study. The study made use of stratified, simple random and census techniques. In the first stage, a stratified sampling technique (Neuman, 2003) was used to select 15 SHSs. The Ghana Education Service's category of school was used as the strata. That is five schools from categories A, B and C were selected. In the next phase, a simple random sampling technique was used to select two SHS 3 classes from each school selected. The number of SHS 3 classes in the selected schools ranges from 7-19. **Each individual in the population of interest had an equal likelihood of selection.** Each unit in the population was identified, and each unit had an equal chance of being in the sample. Selection of one unit did not affect the chances of any other unit (Adjei & Tagoe, 2009; Cohen, Manion & Morrison, 2000). By census, all students in each class were selected for the study. In all, 750 SHS Three students in the western region was used for the validation phase of the instrument development.

Data Collection Instrument

The instruments for the data collection of the study were the performance-based items in mathematics which belongs to the quantitative reasoning items. The instrument was quantitative reasoning items developed by the researcher. The PBA comprised five mathematics computation items presented in real-life scenarios. Each item was designed to assess the proficiency of senior high school (SHS) students in these math domains: transformation, descriptive statistics, mensuration, geometric construction, and linear equations. Item 3, the mensuration item, is shown

> *Mr Mensah decided to put up a two-bedroom flat. The house has two bed rooms, living hall, dining hall, kitchen, two washrooms with toilet and a porch. The dimension of the bedrooms and kitchen are between 12-15ft, dining hall is 10-12ft, living hall is 25 -30ft, washroom with toilet is 5-7ft and the porch 7-12 ft. Mr. Mensah wants to tile the floor of all the rooms. Two sets of tiles are available, one measures 50×50mm and the other, 40 × 40mm. There are 7 pieces in the box of the 50×50mm and 15 pieces in the box of the 40 × 40mm.*

Choose an appropriate dimension of each room within the dimensions given, find how many boxes of each size will be needed to finish all tiling (explain your answer in either mathematical or everyday English).

Data Collection Procedures

The selected class of students sat for the performance-based test. This test was supervised under external examination conditions. The scripts were scored by three WAEC examiners, one for each category of school after a coordination on the scoring rubrics.

Data Processing and Analysis Procedure

Data on research question were analysed using the Graded Response Model (GRM) of the item response theory which estimates the difficulty and discrimination indices. Analysis was done using the winGen software. Using the comparison of item parameter method, a multivariate test to estimate b, and a t-tests perform on b-values to find out there is a significant difference in the probability of success between the two groups.

RESULTS AND DISCUSSION

The research question sought to find if the bias exist in the PBA administered to the students. Gender and category of school were used as the basis for the comparison. That is to find out if male and female at the ability level have different probability of responding above a specific threshold on the same item or a student in the Category C schools and a student at the Category B schools at the same ability level have different probability of responding above a specific threshold on the same item. The IRT method of DIF dictation was used. The winGen software of the IRT was used to generate the b values then independent t test was used to test if there is significant difference in the ability needed by each group to respond above a threshold with 50% probability. For gender, Male was treated as the reference group with female as the focal group. For location, Category a schools was the reference group with the Category B schools and Category C schools as the focal groups. Table 1 shows the comparison between and female.

Table 1- *DIF for Male and Female*

	Item 1 *Transformation*		*Item 2* *Descriptive statistics*		*Item 3* *Mensuration*		*Item 4* *Geometric construction*		*Item 5* *Linear equation*	
	M	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
1	-2.318	-2.949	-.977	-2.712	-2.315	-2.451	-2.226	-2.067	-2.845	-2.336
2	-1.071	.306	-.363	-2.641	-2.133	-1.680	-2.161	-2.004	-2.212	-1.142
3	.133	.515	-.246	-1.101	-1.984	-1.491	-2.007	-1.975	.062	-.720
4	.461	.674	-.013	-.591	-.239	-1.267	-1.719	-1.635	.400	.147
5	.494	1.495	.703	-.362	.010	-1.140	-1.047	-1.338	.961	.159
6	.561	1.619	2.148	-.220	.332	-.911	-.883	1.018	1.566	.217
7	2.121	1.700	2.287	.156	.623	-.354	-.041	1.034	1.996	.479
8	2.293	2.398	2.382	.512	.809	2.551	1.161	1.363	2.223	1.443
9	2.524	2.966	2.853	2.688	.870	2.687	2.750	1.798	2.764	2.230
Sig	t = -0.502, df = 16, p = 0.622		t = 1.987, df = 16, p =0.065		t = 0.004, df = 16, p = 0.997		t = -0.331, df = 16, p = 0.745		t = 0.623, df = 16, p = 0.542	

Table 2 shows the discrepancies (DIF) for male and female ability needed to respond to the items with 50% probability. The table show that there is no DIF in the items for male and female. There was no significant difference in the abilities needed by each group to respond to the items with 50% probability. All sig values were greater than 0.05. Item 1 (t_{16} = -0.502, p = 0.622), Item 2 (t_{16} = 1.987, p =0.065), Item 3 (t_{16} = 0.004, p = 0.997), Item 4 (t_{16} = -0.331, p = 0.745) and Item 5 (t_{16} = 0.623, p = 0.542). This means that female and male have equal chance of performing equally on all the items. The DIF for category of school is presented 2.

Table 2- *DIF for Category of School*

b	*Item 1* *Transformation*			*Item 2* *Descriptive statistics*			*Item 3* *Mensuration*			*Item 4* *Geometric construction*			*Item 5* *Linear equation*		
	A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
1	-2.793	-2.909	-.672	-2.545	-2.031	-.622	-2.915	-2.946	-.926	.207	-2.049	-.810	-2.708	-2.894	-.999
2	-2.689	-1.603	-.319	-2.091	-1.590	-.445	-2.372	-2.344	-.740	.758	-1.846	-.728	-2.080	-2.859	-.822
3	-2.315	-1.302	-.054	-2.087	-1.352	-.204	-1.633	-.718	-.674	.921	-1.788	-.725	-1.381	-1.807	-.756
4	-2.188	-.948	.275	-.983	-.371	.155	.145	-.089	-.361	1.315	.206	-.537	-1.212	-1.334	-.533
5	-.949	-.639	.362	-.325	.911	.163	.301	.139	-.361	1.404	.213	-.158	-.108	-.909	-.417
6	-.890	.330	.521	.028	1.353	.177	.826	.659	-.284	2.202	.254	-.123	.939	-.771	-.192
7	-.377	1.372	.725	.421	1.596	.292	1.830	.844	-.117	2.334	2.472	.656	1.146	-.669	-.080
8	-.307	2.480	.749	2.370	2.498	.864	1.926	1.967	.211	2.395	2.589	.805	1.203	.070	.165
9	2.929	2.805	.841	2.614	2.922	.917	2.232	2.665	.922	2.728	2.908	.968	2.736	1.192	.239
A/B	t_{16} = -1.160, p = 0.263			t_{16} = -0.830, p = 0.419			t_{16} = 0.020, p = 0.984			t_{16} = 1.745, p = 0.100			t_{16} = 1.273, p = 0.221		
A/C	t_{16} = -2.152, p = 0.051			t_{16} = -0.667, p = 0.515			t_{16} = 0.444, p = 0.663			t_{16} = 4.438, p = 0.000			t_{16} = 0.347, p = 0.733		
B/C	t_{16} = -0.472, p = 0.643			t_{16} = 0.461, p = 0.651			t_{16} = 0.436, p = 0.668			t_{16} = 0.573, p = 0.574			t_{16} = -1.584, p = 0.133		

Table 2 shows the discrepancies (DIF) for Category A schools, Category B schools and Category B schools at ability level of 1. The table shows that for $DIF_{catA/catB,}$ there was no significant difference in the abilities needed by each group to respond to the items with 50% probability. All sig values were greater than 0.05. Item 1 (t_{16} = -1.160, p = 0.263), Item 2 (t_{16} = -0.830, p = 0.419), Item 3 (t_{16} = 0.020, p = 0.984), Item 4 (t_{16} = 1.745, p = 0.100) and Item 5 (t_{16} = 1.273, p = 0.221). This means that students in the Category A schools and their colleagues in the Category B schools equal chance of performing equally on all the items.

The table shows that for $DIF_{catA/catC,}$ there was no significant difference in the abilities needed by each group to respond to the items with 50% probability except for Item 4. All sig values were greater than 0.05 except for Item 4. Item 1 (t_{16} = -2.152, p = 0.051), Item 2 (t_{16} = -0.667, p = 0.515), Item 3 (t_{16} = 0.444, p = 0.663), Item 4 t_{16} = 4.438, p = 0.000) and Item 5 (t_{16} = 0.347, p = 0.733). This means that students in the Category A schools and students in the Category C schools have equal chance of performing equally on all the items except for Item 4.

Table again shows that for $DIF_{catB/catC,}$ there was no significant difference in the abilities needed by each group to respond to the items with 50% probability. All sig values were greater than 0.05. Item 1 (t_{16} = -0.472, p = 0.643), Item 2 (t_{16} = 0.461, p = 0.651), Item 3 (t_{16} = 0.436, p = 0.668), Item 4 (t_{16} = 0.573, p = 0.574) and Item 5 (t_{16} = -1.584, p = 0.133). This means that students in the Category B schools and their counterparts in the Category C schools have equal chance of performing equally on all the items.

Discussion

There was no significant difference in the abilities needed by both male and female group to respond to the items with 50% probability. That is for $DIF_{catA/catB,}$ there was no significant difference in the abilities needed by each group to respond to the items with 50% probability. This means that students in the Category A and their colleagues in the Category B schools have equal chance of performing equally on all the items. For $DIF_{catA/catC,}$ there was no significant difference in the abilities needed by each group to respond to the items with 50% probability except for Item 4. This means that students in the Category A and students in the Category C schools have equal chance of performing equally on all the items except for Item 4. For $DIF_{catB/catC,}$ there was no significant difference in the abilities needed by each group to respond to the items with 50% probability. This means that students in the Category B and their counterparts in the Category C schools have equal chance of performing equally on all the items.

While this study found no DIF in almost all the items, Ani (2014) revealed that items functions differential in Economics among male and female students. The difference might lie in the use of multiple choice items in Ani (2014) and GR of this study. it can be said the this developed PBA might would be acceptable than the multiple choice of Ani (2014) as this study found no DIF among male and female.

Royal and Gonzalez (2016) found a psychometrically-sound instrument capable of producing valid and reproducible measures and that no DIF was found in the instrument. Royal and Gonzalez (2016) further found that items with no DIF are capable of producing valid and reproducible measures. It can therefore be concluded that this study which found no DIF in the items due to gender and category of school

has the potential of producing valid and reliable results as supported by Zubairi and Kassim (2006) items no biases produce item quality and test reliability.

Conclusion and recommendation

The study has reveals that the traditional items in mathematics for Senior High Schools could be modified a bit to make it a performance-based assessment, where students would be required to apply knowledge and skills acquired in mathematics to real life situation. It has also revealed that performance-based of this nature could be used in the Senior High Schools to have educational and catalytic effect require. This assessment is also feasible for use in the senior high schools in the Western Region of Ghana. This study would make a significant contribution to knowledge in the area of performance-based assessment for Senior High Schools in the Western Region of Ghana. The study would provide a guide on how to validate polytomous items using IRT. There is no known validation of polytomous items at the SHS in Ghana. Perhaps, assessors in Ghana such as WAEC and teachers do not know the procedures for validation of the polytomus items except the dichotomous items.

Based on the findings of the study, it is recommended that performance-based assessment should be an integral part in the methods of assessment lessons and course at both the colleges of education and universities where teachers are trained by the curriculum developers in mathematics education. This would help provide the knowledge and skills on PBA needed to have an effective and efficient assessment in mathematics.

References

Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice, 4*(22), 87 – 96

Adjei, E., & Tagoe, M. (2009). *Research methods in information studies*. Accra: IAE (UG)

Ainsworth, L., & Viegut, D. (2006). *Common formative assessments, how to connect standards-based instruction and assessment*. Thousand Oaks, CA, Corwin

Ani, E. N. (2014). *Application of item response theory in the development and validation of multiple-choice test in economics*. University of Nigeria, Nsukka: masters' Thesis.

Annan-Brew, R (2020). *Differential item functioning of West African senior secondary certificate examination in core subjects in southern Ghana*. UCC, Ghana: PhD thesis

Arhin, A. K. (2015). The effect of performance assessment-driven instruction on the attitude and achievement of senior high school students in mathematics in Cape Coast Metropolis , Ghana. *Journal of Education and Practice*, *6*(2), 112–114.

Baker, F. B. (2001). *The basis of item response theory*. USA: ERIC Clearinghouse on Assessment and Evaluation.

Bichi, A. A., Hafiz, H. & Bello, S. A. (2016). Evaluation of Northwest University, Kano Post-UTME Test Items Using Item Response Theory. *International Journal of Evaluation and Research in Education (IJERE), 5*(4), 261~270

Bichi, A. A; Embong, R; Mamat, M. & Maiwada, D. A. (2015). Comparison of classical test theory and item response theory: A review of empirical studies. *Austrian Journal of Basic & Applied Science 9(7): 549-556.*

Brennan, R. L. (Ed). (2006). *Educational measurement (4th ed)*. USA: American Council on Education, Praeger Series on Education.

Bruckner, S., Forster, M., Zlatkin-Troitschanskaia, O., Happ, R., Walstad, W. B., Yamaoka, M., & Asano, T. (2015). Gender Effects in Assessment of Economic Knowledge and Understanding: Differences Among Undergraduate Business and Economics Students in Germany, Japan, and the United States. *Peabody Journal of Education, 90* (4): 503-18. doi: http://dx.doi.org/10.1080/0161956X.

Burkhardt, H., & Swan, M. (2008). Designing assessment of performance in mathematics. *Educational Measurement: Issues and Practice, 9*(4), 1–24.

Carvalho, L. F., Primi, R. & Baptista, M. N. (2015). IRT application to verify psychometric properties of the Beck Depression Inventory (BDI) University Psychological. *Bogotá, Colombia, 14* (1), 91-102

Chalmers, P. R. (2012). Mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of statistical software, 48*(6), 231-246

Cohen, L., Manion, L. & Marrison, K. (2000). *Research methods in education*. London: Routedge Falmer.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: The Guilford Press.

Downing, S. M. (2003). Item response theory: applications of modern test theory in medical education. *Medical Education; 37*, 739–745

Egberink, I. J. L; Meijer, R. & Veldkamp, B. P. (2010) Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality* 44(2):232-244. DOI:10.1016/j.jrp.2010.01.007

Ellis, D. P (2011). Item-analysis methods and their implications for the ILTA guidelines for practice: A comparison of the effects of classical test theory and item response theory models on the outcome of a high-stakes entrance. instruction. *Topics in Language Disorders, 1*, 71-88.

Etsey, Y. K. A. & Abu, A. (2013). Colleges of Education tutors' capacity in classroom assessment in northern Ghana. *Journal of Educational Assessment in Africa, 8* , 101-109.

Gao, M. (2012) Classroom Assessments in Mathematics: High School Students' Perceptions *International Journal of Business and Social Science,.3*(2), 63-74

Karaali, G. E., Hernandez, H. V., & Taylor, J. A. (2016). What's in a Name? A Critical Review of Definitions of Quantitative Literacy, Numeracy and Quantitative Reasoning. *Numeracy 9* (1), 14-26. doi: http://dx.doi.org/10.5038/1936-4660.9.1.2.

Le, D. (2013). *Applying item response theory modeling in educational research*. Iowa State University: Masters' Dissertation.

Min, S. & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*(4) 453–477

Neuman, W. L. (2003). *Social research methods: Qualitative and quantitative approaches*. Boston: Allyn and Bacon.

Nitko, A. J. (2004). *Educational Tests and Measurements* (3rded.). USA: Prentice-Hall, Inc

Park, J. (2012). *Developing and Validating an Instrument to Measure College Students' Inferential Reasoning in Statistics: An Argument-Based Approach to Validation*. University of Minnesota: PhD dissertation

Pegg J., (2003). Assessment in mathematics: A developmental approach. In J. Royer (Ed.), *Mathematical Cognition* (pp. 227–259). Greenwich, CT: Information Age Publishing.

Royal, K. D & Gonzalez, L. M. (2016). An evaluation of the psychometric properties of an advising survey for medical and professional program students. *Journal of Educational and Developmental Psychology, 6*(1), 195 – 203

Shavelson, R. J. (2008). Reflections on Quantitative Reasoning: An Assessment Perspective. In *Calculation vs. Context: Quantitative Literacy and Its Implications for Teacher Education*, edited by Bernard L. Madison and Lynn Arthur Steen, 27-44. Washington: Mathematical Association of America.

Zanon, C., Hutz, C. S., Yoo, H. & Hambleton, R. K. (2016). *An application of item response theory to psychological test development*. Psicologia: Reflexão e Crítica, 29 (18), 345-367

Zubairi, A. M. and Kassim,N. L. A. (2006).Classical and Rasch analysis of dichotomously scored reading comprehension test items. *Malaysian Journal of ELT Research*, 2, pp.1-20.