Research Journal of Educational Sciences ISSN 2321-0508
Vol. **5(1),** 1-10, January **(2017)** *Res. J. Educational Sci.*

Dimensionality analysis of students' performance in 2013 BGCSE agricultural examination: Implications for differential item functioning

Moyo S.E.[1*] and Nenty H.J.[2]
[1]Public Health Department, Kanye Seventh Day Adventist College of Nursing, Kanye, Botswana
[2]Educational Foundations Department, University of Botswana, Gaborone, Botswana
sellmoyo@gmail.com

Available online at: www.isca.in, www.isca.me
Received 25th May 2016, revised 1st December 2016, accepted 14th December 2016

Abstract

Assessing desirable changes in learners' behaviour are applied only if the measures in use provide valid outcome data for different subgroups. The deterioration of student performance in the Botswana General Certificate of Education (BGCSE) examination results is a disturbing trend that bothers parents, teacher, policy makers and government. This problem prompts this study on dimensionality analysis of students' performance in 2013 BGCSE Agriculture Examination, implication for differential item functioning, to determine its fairness to all learners. The population for the study was all the 12784 students' responses who sat for the 2013 BGCSE agriculture examination. Differential Item Functioning (DIF) analysis for each item was done according to gender and location type using logits test for t-test significance (p < .05). The findings from this study on gender based DIF indicated that twenty nine (29) out of the 40 items were DIF, seventeen (17) items favoured boys whereas twelve (12) favoured females. With location based, the DIF findings indicated that eighteen (18) were DIF, ten (10) favoured rural and eight (8) favoured urban students. In conclusion, the results of this study, as it explored the national assessment tool, showed that 2013 BGCSE Agriculture Examination was not fair to all students. It was, therefore recommended that test developers and examination bodies should consider improving the quality of their test items by conducting IRT psychometric analysis for validation of DIF purpose among others.

Keywords: Equity in education, Botswana General Certificate of Education examination, Agriculture, Differential item functioning, Item bias.

Introduction

Equity and excellence of assessing desirable changes in the learners' behaviour are applied only if the measures used provide valid outcome data for different subgroups. The desirable change in learner's behaviour is defined through quantity (amount) and quality (desirability) of the newly acquired behaviour by bringing in the essence of assessment, measurement and evaluation[1]. The use of examination is widely accepted as a psychological measurement instrument which determines the extent to which an individual has acquired the intended desirable behaviour.

In every society testing through educational examination serves as the vehicle through which human cognitive behaviours are displayed or exhibited and documented. The realization of testing for desired behaviour yields valid results and is interpretable, if the examination is designed and used to measure one and only underlying behaviour. This is especially true, when the development of assessment and analysis is conducted within an item response theory framework. This demand on assessment instrument still remains a core challenge to examinations in African education. In Botswana, since the localisation of examinations, the secondary education system has failed to realize its main academic objective for every student. Oftentimes, local newspapers report unpleasant interpretations of students' achievement in public examinations[2]. Thus, year after year, students' performances in the national examination are declining. Some students may be more susceptible to fail due to unfair assessment procedures and interpretations of the results from national examination. Once again an interesting phenomenon which occurred over a period of time was that the Botswana General Certificate of Secondary Education (BGCSE) replaced Cambridge Overseas School Certificate (COSC). COSC was an international examination administered by University of Cambridge Local Examination Syndicate (UCLES).

This change was necessary since COSC seemed to be irrelevant as it was not designed for local issues and cultural set-up in mind[3]. On the other hand, BGCSE was to be a quality assurance measure since it would examine the local syllabi which were based on the philosophy of Botswana's education system[3].

Nevertheless, one may be tempted to speculate that the alarming decline in public examination results over the last couples of years started immediately after the shift of examinations from international to local setting. It is therefore not by coincidence for one to associate the deterioration of student performance to the standard of education and particularly the quality of the test items used in the national examination. This was attested also by [4]who revealed that Botswana Examination Council (BEC) only review structured items and ignoring other components. It is appears to be obvious that a number of candidates who are strong in other components are disadvantaged.

The quality of test items in any public examinations is always examined through item analysis of examinees' responses[5]. As such apsychological instrument cannot be assumed to provide accurate information without proper psychometric evidence to support claims of what the instrument purports to measure[6]. The conversion of tests scores to equal interval measures is particularly important. The reason is that many education reforms efforts focus on monitoring the performance of underachieving and underrepresented students[7,8].

If the examination data do not convert to equal interval measures, then results of such analysis may provide incorrect/or incomplete information on student's performance. BEC does not seem to subscribe to the modern way of analysing student results, instead it seems to opt for analysis of raw scores. It is upon Botswana measurement specialists to pursue and to embrace change to seek a transformed solution for the declining of results using a proper way of constructing and interpreting the public examination, which will enhance the quality of education.

Theoretical foundation: Item response theory (IRT): Successful application of fundamental measurement in educational research and much current of academia speak, writing, advice and practice has come to rescue all society who are stagnant in educational transformation. Through the knowledge of concerned theory and practice of fundamental measurement which exist, has of course transformed other nationals in the developed states like the West and Asia to better society in education.

Item Response Theory (IRT) is among the major progressive models in educational measurement. IRT is probabilistic model for expressing the association between an individual's response to an item and the underlying latent variable being measured by the item. Such an item (task, question, statement) may elicit the exhibition of appropriate cognitive, effective or psychomotor trait or attribute[9]. The latent variable, expressed as theta (θ), is a continuous unidimensional construct that explains the covariance among item responses [9]. Individual persons at higher level of θ have a higher probability of responding correctly an item and otherwise vice versa.

IRT model uses item responses to obtain scaled estimates of θ, as well as to calibrate items and examine their properties[10]. Each item is characterized by one or more model parameters. The three parameters associated with the item are the item discrimination parameter (a); the difficulty parameter (b); and the guessing parameter (c)[9].

Most IRT models like the one, two, and three parameter logistics assume that the normal ogive or logistic function describes the relationship accurately and fits the data. The logistic function is similar to the normal ogive function and is mathematically similar to use and, as a result is predominantly used in research.

The item characteristics curve (ICC) can be viewed as the regression of item score on the underlying variable θ[11]. ICC shows that examinees with different amount of the latent trait have different probabilities of getting the item correct. The probability of an examinee scoring the item correctly depends on the person's parameter and the item parameters[12]. Usually, an ICC has one, two or three parameters that are called parameters. IRT make some confining operational demands to ensure that scores that emanate from testing can be validly interpreted as representing the value of the trait or behaviour under measurement possessed by the testees. Only on such scores are valid policy decision based.

Determination of DIF with IRT: IRT has brought about significant changes in psychometric theory and test development. In its most basic form, it postulates that a single ability underlies examinee performance on a test and that the probability of a correct response on an item is a monotonically increasing via the curve[13]. IRT offers a powerful method of investigating item bias- which also referred to as differential item functioning (DIF). There are many factors that constitute sources of invalidity in a test impinge on the unidimensionality of a test. Such factors can be psychological or physical which include, the nature of the subject matter content, nature of human cognitive level, language skills and demands mention just a few[14,15].

Some subject contents are complex and need a lot of competence, skill and expertise to carry out a good analysis necessary to break them into bits and pieces of tasks based on which items could be developed. For example, the boundaries between the different subjects are too blurred to sustain the development of item that measure purely one specific subject matter. Hence an item developed to measure, for example, history also taps into English language and this is likely to introduce a second dimension in a history test. Similarly, the boundaries between the different levels of Bloom taxonomy are also too blurred to sustain the development of item that measure purely one specific level of the taxonomy.

The level of language with which items are expressed sometimes constitute a dimension in addition to that of the content. The language with which test items are expressed if not basic enough for the maturity level of the testees may constitute a dimension different from the content knowledge and cognitive behaviour under measurement. When a test item provokes the ability under measurement possessed by the testees, it also provokes some related psychological traits[14,15]. Significant differential item functioning as a result of item bias constitutes a dimension different from what the item was developed to measure and hence violates the unidimensionality assumption of IRT in test items[9].

Among them is the use of ICCs for DIF detection concerns the comparison of differences in the ICCs for different subgroups. With ICC only two groups can be compared at a time, but a particular sample can be divided into various subgroups for such comparisons. In fact the area between the equated ICCs is an indication of the degree of bias present in a considered test item [16]. The use of ICC is predominately used by measurement specialists in their attempt to detect DIF in test items. Conversely, there is other modified method of testing DIF, which is through group difference for ability expressed over standard deviation to get t-values. Thus, it is calculated by using the mean difference of the logits (b-value) of IRT. The t-test is attained through comparison of the variation of the ability in the data- expressed as the standard deviation of the difference between means of logits. The level of significance is set at (p <.05) to test significance difference when the t value is greater than 1.98 of the critical value[17]. The t-values are also used to detect DIF and hence were useful in assessing the dimensionality of agriculture examinations.

Statement of the problem and purpose of the study: Exiting evidence shows that BGCSE results for students in public secondary schools are not as good as they used be. In the past the results at times are deteriorated yearly across all schools and across levels. The published preliminary analysis of Botswana's national results from year 2000 to 2012 showed a decline in the A, B and C of BGCSE 74.7% in 2011 compared to 77.3 % in 2000 including agriculture [18]. Decline of high grades appears to be indicative of fall in standard of education. It has been revealed that Botswana Examination Council (BEC) only review structured items and ignoring other components[4]. This means that multiple choice items are among components which are not reviewed, despite the equal weighting the components contribute to the BGCSE. For instance, in agriculture, the weighting of the papers are; Component 1 (40%) and Component 2 (40%)[19].

The failure to review other set of item is a major gap which disadvantages a number of candidates who are stronger in those un-reviewed components. Accuracy and fairness in testing and quality of examination are to some extent compromised. Hence decisions based on such test results are prone to be invalid.In an attempt to contribute a solution to this problem, the current study purposeful to examine the dimensionality of agriculture examination as a means of generating information with which contribution could be made to the improvement of test development. The specific questions of the study are:

Research questions: i. To what extent do gender influence DIF among students responses in the 2013 BGCSE agriculture multiple choice items? ii. To what extent does location influence DIF among students' responses in the 2013 BGCSE agriculture multiple choice items?

Literature review: Gender and differential item functioning (DIF): On a study carried with the aim to detect differential item functioning (DIF) items across gender groups; analysed item content for the possible sources of DIF; and eventually investigated the effect of DIF items on the criterion-related validity of the test scores in the quantitative section of the university entrance examination (UEE) in Turkey[20]. The study evaluated DIF on items which came from subject matter related factors, cognitive skills measured, and item format characteristics. It seemed that higher order cognitive skills and figural or graphical representations used in item content were the two sources of DIF for favouring male students, whereas routine algorithmic calculations could produce DIF against males. Among the factors considered, cognitive skills assessed by items seem the most effective factor in producing gender DIF. However, DIF items did not create a threat to the criterion-related validity of the quantitative section of the UEE.

Notably, consideration of the DIF items and gender differences revealed gender differences in item selection on any measure that is used for a similar purpose should be considered[20]. This study shed light on DIF but did not spell out the DIF analysis and sample sizes used.

A similar study was done (1) to identify gender DIF in a large-scale science assessment, and (2) to look for trends in the DIF and non-DIF items due to content, cognitive demands, item type, item text, and visual-spatial/reference factors. To facilitate the analyses, DIF study was conducted at three grade levels, and for two randomly-equivalent forms of the science assessment at each grade level (administered in different years). A variant of the standardization procedure was applied to very large sets of data (six sets of data, each involving 60,000 students), and has the advantages of being easy to understand and to explain to practitioners[21]. Several findings that would be useful to pass on to test development committees emerged from the study. For example, when there is DIF in science items, multiple-choice items tend to favour males and open-response items tend to favour females. Compiling DIF information across multiple grades and years increases the likelihood that important trends in the data will be identified and item writing practices are informed by more than anecdotal reports about DIF[21].

On another study a comparative method analysis for DIF, used the chemistry test data of junior secondary school students in Philippines to demonstrate the difference between CTT and IRT. Cronbach's alpha and Rasch model were used to for analyse the data for the study through IRT and CTT respectively. It was found among others that IRT estimates of item difficulty do not change across samples as compared with CTT which was inconsistent [22]. The study also found that the difficulty indices were more stable across forms of test in IRT than CTT approach. This study shed some light on Rasch model[22].

In a study in Botswana, used IRT approach to detect gender biased items in public examinations. The author randomly selected a sample of 4000 students response to Mathematics Paper 1 of the Botswana Junior Certificate Examination were selected from the 36, 000 students who sat for the examination. Out of 36,000 students set for examination, 2000 were males and 2000 were females. The examination paper consisted of 38 items. To detect gender bias items, test generated the item characteristics curves (ICC for the male/female). The study compared the ICC curves for the male and female groups, and found that, out of 16 test items that fitted the 3PL item response theory (IRT) statistical analysis, 5 items were gender biased. The research concluded that through the application of IRT methodology, it was clear that the biased item were detected, hence called for further need to detect gender bias test items from other subjects in any public examinations, through the use of item response approach (ICC curves). The sample used in this study was relatively large; to detect differential item functioning using IRT model and it will be useful to the current study[23].

Differential item functioning analysis by location: DIF study was carried out to investigate items that were bias in relation to school type (private and public schools), school location (urban and rural schools) using National Examinations Council (NECO) Biology questions for 2012 examination. The research design employed in this study was a comparative research type of design. The sample comprised candidates in Taraba State, Nigeria. Four hundred and forty seven (447) candidates were used and the NECO Biology test contains 60 items which was administered to the students[24].

They used logistic regression to analysis the data. The research findings showed that out of sixty items in test, 10 items were biased in relation to school type and 8 items in relation to school location. The implication of these findings is that NECO Biology examinations questions have incidences of differential item functioning (DIF). From the result of the findings, it was then recommended that test experts and developers should explore the use of DIF approach to detect biased items[24]. Though this study used a comparatively small sample considering the massive numbers of examinees in nationwide examinations, it provides fundamental guidance to carry out DIF analysis.

The three-parameter logistic model study was done to identify DIF items in Mathematics Paper 1 of 'SijilPelajaran Malaysia (SPM)' Trail Examination for Sri Aman/Betong Division for the year 2003 across urban and rural students. The study flagged only Item 15 as DIF across urban and rural students [25]. The positive area for Item 15 indicated that the item was in favour of the urban group. The difference between the signed and unsigned area for the item shows that it was a non-consistent DIF and was not in favour of urban group over the entire ability range[25]. This study shed some light that location has an impact on item functioning, though the sample size of study was not indicated.

Similarly a quantitative study to identify location biased items with respect to rural and urban schools in the 2010 Botswana Junior Certificate Examination Mathematics paper 1 using IRT Item Response Characteristics Curves was also done [26]. The study further identified rural/ urban location biased items with respect to gender of students. The 2010 Botswana Junior Certificate Examination Mathematics examination Paper 1 consisted of forty (40) multiple choice test items. The sample for this study comprised of 4000 students randomly selected from a population of 36940 who sat for 2010 Botswana Junior Certificate Examination Mathematics Paper 1.

The sample of students randomly selected consisted of 2000 male students, of which 1000 were from rural schools and 1000 were from urban schools. The remaining 2000 students were females, 1000 from urban schools and 1000 from rural schools.3PL (Multilog software) Item Response Theory (IRT) statistical analysis was used to generate the Item Characteristics Curves (ICCs) for the corresponding groups rural/urban, rural / urban with respect to gender. The ICCs for the corresponding groups were compared to identify rural/urban location biased items. The findings of the study revealed that from the 24 items that fitted the IRT (3PLM) model, six (6) items were rural /urban location biased items.

The study further found out that three (3) items were rural /urban location biased with respect to males and six (6) items were rural /urban location biased with respect to females. They reached a conclusion those students who attended schools in urban area outperformed students who attended schools in rural areas. It is recommended that test developers in Africa should always endeavour to create bias free items for testing and examination purposes and the connotations reflected in test or examination items should be relevant to the life experiences of examinees responding to the items[26].The study shed some light that location has significant impact on DIF, hence it is of useful foundation to the current study.

Methodology

In an attempt to reach valid findings, this study employed exploratory design assesses the underlying dimensionality property of multiples choice items for 2013 BGCSE Agricultural Examination. This design is much appropriate to detect items which were DIF across examinees, because it has the strength to diagnose the validity and underlying dimension of 2013 BGCSE Agriculture Examination. This assumed to provide the researcher's direction on ways to improve and to monitor the instrument from one test administration to the next[27]. Consequently results from this study would provide insight of the dimensionality of Botswana examination (in this case 2013 BGCSE Agriculture multiple choice item).

This study target 12784 Form 5 candidates responses to items in Paper 1of BGCSE Agriculture Examination administered to the 32 public senior secondary students both government and government aided schools in Botswana. The multiple choice component (Paper 1) of the examination carries the same 40 percent weight as for Paper 2 (constructed response items) contribution to the whole BGCSE Agriculture Examination.

In this study, every student's responses to multiple choice items in BGCSE Agriculture were given equal chance to be selectedand this enhanced the external validity of the study. In effect, students' academic records in agriculture examinations for 2013 were available. The researchers retrieved the entire student's responses to every item for 2013 agriculture multiple choice examination. The subjects consisted of boys and girls to enable assessment of gender-based performance of each item and the dimension of the examination test performance. Student performance represents students' responses to items in which a score was awarded one to correct item and zero to the incorrect respond.

Location was also another independent variable with a distinguishing parameter in which student responses to an item could have influenced the examination dimension. The students' responses to the 40 multiple choice items on the basis of the three locations were as follow; 3945 for rural, 5722 for peri-urban and 3945 for urban. The wide difference observed between the students who attended schools at peri-urban and other two locations, implicated to drop the peri-urban in the location analysis. That is only students' responses from the two extreme locations were used because they were relatively equal groups for analysis purpose and it was also a means to maximize the variance of the variables of the research hypothesis. This means that the current study only used rural and urban locations in order to maximize the variance in the analysis of detecting DIF among the items in the examination.

Permission from BEC was requested to retrieve students' academic records on agriculture examination for 2013 final year. The scores for BGCSE Agriculture are assumed to be valid, on the basis that BEC has intensive panel-base who deals with content analysis and face validation for every subject. It also assumed that the instrument was reliable in which the examination scores for students were attained.

BILOG-MG V3.0 software for IRT model was used to obtain 1-parameter logistic and thereafter logistics converted t-values which were tested for significance for gender and location to establish differential item functioning of items in 2013 Agriculture Examination. This was to justify whether items were fair or unfair to all students.

Results and discussion

In this study the researcher used a total of 12734 students' responses in the 40 multiple items for agriculture examination. The study consisted of 5995 (47%) of the total respondents were boys and 6739 (53 %) were girls. The result showed many females than males students sat for 2013 BGSCE agriculture examination. The students' responses were also classified into school location where they attended namely, urban 3067 (24.1%) peri-urban 5722 (44.9%) and rural 3945 (31%) of the total number of students' responses. The total scores of students responses were approximately normally distributed with the skewness of .326 (SE = .022) and a kurtosis of -.169 (SE = .043).

Q1. To what extent do gender influence differential item functioning in the 2013 BGCSE agriculture multiple choice items?: The question dealt with testing whether gender did or not significantly influence item functioning in the 2013 BGCSE Agriculture multiple choice items on the basis of the estimate of item parameters to generate logits mean difference. The logits of the male and female were used to estimate the item parameter on the 1PL to generate the logits mean difference (t-values) (Table-1).

Table-1 showed the DIF statistics in logits mean difference (t-values) for girls and boys on each of the 40 items in 2013 BGCSE Agriculture Examination. The t-test comparing of logit for girls and boys flagged 29 items with significantly ($p <.05$) DIF and 11 non-DIF items. A sign on t-value of DIF reflected both direction and magnitude of DIF. It was obtained by attaching a positive sign to DIF in favour of females and a negative sign if the item revealed DIF in favour of the males, only when the t-value was greater than 1.98 critical values ($p < .05$). In this study, 12 out of 29 items favoured girls. These were Items3, 4, 11, 12, 14, 16, 21, 24, 29, 36, 31 and 38. While17 out of 29 items were in favour of boys, these were 5, 6, 7, 8, 9, 15, 18, 20, 22, 23, 25, 26, 30, 32, 33, 39 and 40 (Table-2). The significant difference found between the logits or b-values of boys and girls implied that they were other factors apart from ability under measurement influenced responses to the items in favour one of group over the other.

Q2.To what extent does location influence differential item functioning in the 2013 BGCSE Agriculture multiple choice items?: To answer the question, mean difference for logits (t-values) was alsoused for DIF analysis of student’s responses to agriculture items. The mean difference showed significance at alpha .05 (t-values > 1.98). Here, the analysis of the DIF for students’ responses to agriculture items showed that all but 18 of the 40 items were flagged the DIF on the basis of location (Table-3). The 18 DIF items were: Items 1,3,4,5, 6, 9,10, 11, 15, 16, 17, 25, 28, 33, 36, 37, 39 and 40.Ten out of 18 items favoured rural student's responses while the remaining 8 items were in favour of the urban student responses (see Table 4). This implied that location had influence on some items. Thus one group was favoured over the other. Like in this case of the study, more of DIF items favoured students' responses who attended schools in the rural area than those who attended in the urban area.

Table-1: Analysis of gender DIF in 2013 BGCSE agriculture multiple-choice test items using logits (b-values) for t-value

	Gender				DIF Index		
	Girls		Boys				t-value
Item #	Logit	SE	Logit	SE	Logit	SE	
	(b-value)		(b-value)				
1	0.745	0.044	0.766	0.046	-0.021	0.064	-0.328
2	0.160	0.043	0.235	0.047	-0.075	0.063	-1.190
3	-1.778	0.046	-2.828	0.059	1.05	0.075	14.000*
4	-3.132	0.057	-4.34	0.08	1.208	0.098	12.327*
5	-1.117	0.043	-0.983	0.046	-0.133	0.062	-2.145*
6	1.123	0.045	1.555	0.049	-0.432	0.067	-6.448*
7	0.006	0.042	0.271	0.045	-0.265	0.062	-4.274*
8	-1.208	0.042	-0.835	0.044	-0.374	0.061	-6.131*
9	1.631	0.045	2.201	0.051	-0.570	0.068	-8.383*
10	0.827	0.045	0.533	0.470	0.294	0.065	0.452
11	-0.660	0.041	-0.211	0.043	-0.448	0.059	7.593*
12	-1.139	0.043	-1.326	0.048	0.187	0.064	2.922*
13	-1.633	0.046	-1.520	0.049	-0.112	0.067	-1.750
14	2.136	0.051	1.826	0.051	0.310	0.072	4.306*
15	-0.331	0.042	-0.121	0.045	-0.210	0.061	-3.443*
16	-0.689	0.043	-0.841	0.046	0.152	0.063	2.413*
17	2.620	0.053	2.726	0.055	-0.106	0.077	-1.377
18	2.957	0.057	3.144	0.059	-0.187	0.082	-2.280*
19	-0.113	0.041	-0.429	0.044	0.317	0.600	0.528
20	-6.527	0.134	-5.776	0.117	-0.751	0.178	-4.219*
21	-1.588	0.046	-2.678	0.058	1.090	0.074	14.730*
22	0.125	0.042	0.346	0.045	-0.220	0.061	-3.607*
23	-1.084	0.043	-0.914	0.046	-0.170	0.063	-2.786*
24	-0.252	0.041	-0.506	0.044	0.254	0.060	4.233*
25	-3.577	0.062	-3.203	0.064	-0.374	0.089	-4.202*
26	-2.232	0.049	-1.825	0.050	-0.407	0.070	-5.771*
27	0.632	0.042	0.589	0.044	0.043	0.061	0.705
28	-0.124	0.043	-0.247	0.046	0.123	0.063	1.952
29	-1.351	0.044	-1.993	0.052	0.642	0.068	9.441*
30	1.059	0.044	1.357	0.047	-0.298	0.064	-4.656*
31	-1.151	0.045	-1.703	0.051	0.552	0.068	8.118*
32	0.172	0.040	0.606	0.042	-0.434	0.058	-7.483*
33	-2.676	0.052	-2.414	0.055	-0.262	0.076	-3.447*
34	-0.269	0.042	0.066	0.045	-0.336	0.062	-0.274
35	-0.480	0.042	-0.122	0.045	-0.359	0.062	-0.297
36	2.322	0.053	2.167	0.054	0.155	0.076	2.039*
37	-0.021	0.043	-0.004	0.046	-0.017	0.063	-0.270
38	-1.494	0.045	-2.314	0.054	0.820	0.070	11.714*
39	-3.016	0.056	-2.570	0.058	-0.446	0.081	-5.506*
40	-0.791	0.043	-0.602	0.047	-0.189	0.064	-2.953*

*The item selected with t-value greater than 1.98 is significant

Table-2: Number of items favoured gender -type

Item	Items favoured girls	Items favoured boys	Total
Biased Item	3, 4, 11, 12, 14, 16, 21, 24, 29, 36, 31, 38	5, 6, 7, 8, 9, 15, 18, 20, 22, 23, 25, 26, 30, 32, 33, 39, 40	
Total number of Items	12	17	29

Table-3: Analysis of location DIF in 2013 BGCSE agriculture multiple-choice test items using logits (b-values) for t-value

	Location				DIF Index		
	Rural		Urban				t-value
Item #	Logit	SE	Logit	SE	Logit	SE	
	(b-value)		(b-value)				
1	0.067	0.055	0.959	0.064	-0.288	0.085	-3.388*
2	0.118	0.055	0.108	0.063	0.011	0.083	0.133
3	-2.556	0.067	-1.851	0.069	-0.704	0.096	-7.333*
4	-3.689	0.084	-3.067	0.086	-0.622	0.120	-5.183*
5	-1.152	0.055	-0.867	0.062	-0.285	0.083	-3.434*
6	1.455	0.059	1.074	0.065	0.382	0.087	4.391*
7	-0.027	0.053	0.067	0.062	-0.094	0.087	-1.080
8	-1.081	0.054	-0.931	0.061	-0.150	0.081	-1.852
9	1.694	0.058	2.022	0.067	-0.328	0.088	-3.727*
10	0.468	0.055	1.061	0.066	-0.592	0.086	-6.884*
11	-0.462	0.051	-0.677	0.059	0.215	0.078	2.756*
12	-1.106	0.055	-1.190	0.063	0.085	0.084	1.012
13	-1.501	0.058	-1.519	0.067	0.019	0.089	0.213
14	1.631	0.060	2.001	0.072	-0.369	0.094	-3.926*
15	-0.056	0.053	-0.486	0.062	0.430	0.081	5.309*
16	-0.444	0.054	-1.005	0.065	0.561	0.084	6.679*
17	2.640	0.069	2.636	0.075	0.004	0.101	0.040
18	2.867	0.071	3.063	0.080	-0.197	0.107	-1.841
19	-0.261	0.052	-0.147	0.060	-0.114	0.079	-1.443
20	-6.155	0.164	-5.784	0.169	-0.370	0.236	-1.568
21	-1.880	0.060	-2.022	0.072	0.142	0.094	1.511
22	0.335	0.053	0.233	0.060	0.102	0.080	1.275
23	-0.883	0.054	-0.963	0.064	0.080	0.084	0.952
24	-0.222	0.052	-0.391	0.060	0.169	0.079	2.139*
25	-3.065	0.074	-3.289	0.089	0.225	0.116	1.940
26	-1.950	0.061	-1.867	0.069	-0.084	0.092	-0.913
27	0.615	0.053	0.866	0.060	-0.251	0.080	-3.138*
28	-0.201	0.054	-0.266	0.063	0.065	0.083	0.783
29	-1.391	0.057	-1.553	0.067	0.162	0.088	1.841
30	1.232	0.057	1.128	0.063	0.103	0.084	1.226
31	-1.278	0.050	-1.347	0.067	0.069	0.088	0.784
32	0.292	0.057	0.416	0.057	-0.123	0.076	-1.618
33	-2.262	0.064	-2.646	0.079	0.384	0.101	3.802*

	Location				DIF Index		
	Rural		Urban				t-value
Item #	Logit	SE	Logit	SE	Logit	SE	
	(b-value)		(b-value)				
34	-0.258	0.053	-0.113	0.061	-0.012	0.081	-0.148
35	-0.258	0.053	-0.352	0.062	0.095	0.082	1.159
36	2.386	0.068	2.116	0.074	0.270	0.100	2.700*
37	0.207	0.054	-0.248	0.063	0.454	0.083	5.470*
38	-1.722	0.059	-1.662	0.067	-0.060	0.089	-0.674
39	-2.535	0.068	-2.919	0.084	0.384	0.108	3.556*
40	-0.562	0.055	-0.797	0.064	0.235	0.085	2.765*

*The item selected with t-value greater than 1.98 is significant.

Table-4: Number of items favoured location -type

Item	Items favoured rural students	Items favoured urban student	Total
Biased Item	6, 11, 16, 17, 25, 33, 36, 37, 39, 40	1,3,4,5,9,10,15,28	
Total Number of Items	10	8	18

Differential item functioning by gender -type: The findings of this study showed that 29 items out of the total of 40items for the 2013 BGCSE Agriculture Examination items functioned significantly different between boys and girls. More of the items favoured boys than females. For instance, out of 29 items which were identified DIF, 17 items favoured boys over girls. Testing through public examination like BGCSE, has been fully accepted in most modern societies as the most objective method of decision making in schools, industries and government establishments. It is for this reason that it is assumed that test has to be fair to all groups who undertake the public examinations. It is revealed from findings that girls were disadvantaged in their attempt to answer the tasks placed before them because they were other demand beside knowledge of agriculture tied to the items. The findings were in line with those researchers who detected DIF on some items favoured males' students over females in some items[20],[21]. There were other factors attributed significant systematic variance which favoured boys over girls. The foregoing studies even though were from mathematics and science fields respectively, nevertheless they revealed similar findings to corroborate the current study in agriculture. Agriculture is also an applied science subjects and hence it appears that agriculture items were favouring boys than girls' students. This provoked a thinking to associate that science either favoured boys against girls or was the instrument error measurement.

However as observed, that test is bias if it contains language or content that is differentially familiar for different subgroups of the examinees; it contains sources of difficulty that are irrelevant or extraneous to the construct being tested; that a test is biased if it contains clues that would increase the performance of one group over another[15]. Thus boys might have had a sound command of language related to agriculture which enabled them to outperform their females' counterparts. Culturally, it appears boys to be better than girls in agriculture, not as subject of study in school, but as farming practices. This empowers boysat early age to interact and gain exposure through most of agricultural activities, so this could have influenced their performance in agricultural items. The unfairness of the items revealed in this study appeared to be dealing with incomplete use of standards when design instruments for national examination. This was confirmed by the researchers who observed that BEC only reviewed the structured questions for agriculture and ignored the multiple choice items[4].

The unreviewed items may have had disadvantaged other candidates who were strong in multiple choice items. The outcome of this study indicated that accuracy and fairness in testing to some extent was compromised. As stressed that the existing absence of the regenerative feedback through large scale assessment is a handicap to attaining and maintaining high standard in education[28].

Differential items functioning by location: The findings showed that agriculture items significantly functioned differently among the student responses from rural and urban area are in line with other researchers[24]. Their study revealed that out of sixty items in test, 10 items were biased in relation to school type and 8 items in relation to school location [24]. In this study out of the total of 18 items which were found to function significantly differently at alpha level .05, that 56% (10) of the items favoured rural student’s response. This implied that students from urban schools and students from rural schools with the same latent ability in agriculture responded in different ways to the 18 out of 40 items and such items were said to be biased. There are many factors that could have influenced examinees from different subgroups to respond differently to

45% of total 40 items for 2013 BGSCE Agriculture Examination by location influence which among other was exposure.

One attempted to speculate that assessment tool favoured much rural folks of students. Given the perception that the Botswana society tend to associate agriculture as rural life, this might have boosted the students from rural setting because that the main mode of exposure they might have gained at the farm and hence had influenced their performance. Contrary to the students who attended schools in the urban area, they underperformed in agriculture due to somelikely influence of the parental guidance who consider agricultural as low class and industrial subject, hence provoked negative attitude on students towards it. This was corroborated with the researcher who holds that in measurement, an item is biased if its construction, setting, language, idea or interest portrayed, picture/diagram used, relevance and illustration are giving an undue advantage or disadvantage to a particular group of testees over the other group[14]. One of these factors mentionedabove might have had an influence of students in rural and urban perform differently in some items despite their same ability level.

In the contrary, study revealed the existence of location bias in mathematics examination was shifted towards the students who attended schools rural areas[26]. Thus, students who attended schools in urban area outperformed students who attended schools in rural areas. Similarly the inverse of the current study was also attested by[25] in their study DIF items favoured urban group.

Implications for differential item functioning: The results obtained through testing and test scores have an important use for people in Botswana. It is through the test administered to people that test scores used for promotion, selection for various jobs, placed in various institutions, given awards, scholarship and appointment into various positions are obtained. Their use also applied to education sector, social-economic sector, and both political and non-political sectors. All these sectors make an informed decision based test scores. Botswana is a heterogeneous state with diverse geographical locations features. The test items which are administered to students at all levels either at schools or national examination must be fair to all. Otherwise, if the test items are biased like what is revealed in this study, then there is a major concern regarding the validity of scores to warrant a decision making process.

If decision to develop any new programs is made on performance on a biased test then such programs will also be biased. For instance, the study had revealed that they were gender based-DIF items which favoured boys than girls, while with the location based-DIF items more items favoured students attended in the rural than urban students, then a performance developed based on such scores will also tend to be biased accordingly. The stakeholders should be concerned with what factors attributed to the systematic extraneous variance of the items. As such one is tempted to speculate that DIF factors such as language, un-equal access to natural laboratories like farms or fields, practical exposure to agriculture activities to mention but a few. This implication is that if education intends to put in place a corrective measure regarding agriculture on the basis of the biased tests, obviously some learner will be are disadvantaged. It is silent warning, if the issue of test bias is not properly addressed as they avail in test analysis of our national examination then some of the vision 2016 pillars 'educated and informed nation', and 'a prosperous, productive and innovative nation' would remain an unattainable dream.

Conclusion

It is apparent that BGCSE results for students in public secondary schools are not valid as they should be. The gender-DIF items favoured mostly the male students. This implies that the measurement was ineffective generating valid scores because a particular group of testees were given an undue advantage or disadvantage over the other group. Location also detected DIF in which more items favoured the students who attended rural schools. Students who attended school in the rural areas outperformed those students who attended schools in the urban areas though both groups may have the same ability in agriculture. These students in the rural schools had an upper practical exposure to agriculture before starting examination.

The researchers recommend that BEC should put in place the in-service training through workshops, conference and other available mechanisms in order to update the test/examination developers on issues of assessment bias. These will courage the uptake and intensive application of DIF items analysis among teachers and test/examinations developer in Botswana.

References

1. Nenty H.J. (1985). Fundamental of measurement and evaluation in education. Unpublished monograph, University of Calabar, Nigeria.
2. Hunyepa J. (2014). Khama must be petitioned over problems in education. The Botswana Gazette, News, Botswana (pty) Ltd, Gaborone, Botswana.
3. Utlwang A. (2003). The localization of Cambridge School Examinations as a quality assurance measure. *AEAA Conference*, Cape Town, South Africa.
4. Thobega M. and Masole T.M. (2008). Predicting students’ performance on agricultural science examination from forecast grades. *US-China Education Review*, 5(10), 45-52.
5. Adedoyin O.O. and Mokobi T. (2013). Using IRT psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science*, 3(4), 992-1011.

6. Brian F.F., Daniel H.B. and William E.F. (2007). The psychometric properties of the agricultural hazardous occupation order certification training program on written examinations. *Journal of Agricultural Education*, 48(4), 11-19.

7. Lee J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher*, 31(1), 3-12.

8. R Paige, E Hickok (2004). No child left behind: A toolkit for teachers. U.S. Department of Education, Office of the Deputy Secretary, Washington DC: Author.

9. Nenty H.J.O. A. Afemikhe& J. C. Adewale (Eds.) (2004), .Issues ineducational measurement and evaluation inNigeria (in honour of Professor WoleFalayajo) (pp.371–383). Ibadan. Institute of Education, University of Ibadan, Nigeria, From CTT to IRT: An introduction to a desirable transition

10. Mellenbergh G.J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, 29, 223-236.

11. Lord F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

12. Pido S. (2012). Comparison of item analysis results obtained using item response theory and classical test theory approaches. *Journal of Educational Assessment in Africa,* 7, 192- 207.

13. Hambleton R.K., Swaminathan H. and Rogers H.J. (1991). Fundamentals of item response theory. Sage: Newbury Park

14. Nenty H.J. (2008). Constructing, administering, scoring and interpreting from educational Instruments. Educational Foundations, University of Botswana, Gaborone, Botswana.

15. Nenty H.J. (2010). Gender-bias and human resources development: Some measurement considerations. *Ilorin Journal of Education, 29, 13-26.*

16. Oche E.S. (2012). Issues in test item bias in public examinations in Nigeria: Implications for testing. *International Journal of Academic Research in Progressive Education and Development,* 1(1), 179 -187.

17. Nenty H.J. (2013). Fundamentals of quantitative research education. Book under preparation, University of Botswana, Gaborone.

18. Republic of Botswana. Ministry of Education (2013). Request for quotation for the evaluation of declining results in basic education sector (primary, junior and senior secondary) since 2007 to date- 2013. Gaborone, Botswana: Government printer.

19. Republic of Botswana (2002). Botswana examination council act. Gaborone, Botswana

20. Kalaycioglu D.B. and Berberoglu G. (2010). 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.

21. Robin F, Zenisky A.L and Hambleton (2003). DIF detection and interpretation in large scale science assessments: Informing item writing practices. University of Massachusetts, Amherst and Frederic Robin Educational Testing Service.

22. Magno C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment,* 1(1), 1-11.

23. Adedoyin O.O. (2010). Using IRT approach to detect gender biased items in public examinations: A case study from the Botswana junior certificate examination in mathematics. *Educational Research and Reviews,* 5(7), 385-399. http://www.academicjournals.org/ERR2.

24. Amuche C.I. and Fan A.F. (2014). An assessment of item bias using differential item functioning technique in Neco biology conducted examinations in Taraba State Nigeria. *American International Journal of Research in Humanities, Arts and Social Sciences,* 6(1), 95-100.

25. Eng L.S. and Hoe L.S. (2005). Detecting differential item functioning (DIF) in standardized multiple-choice test: An application of item response theory (IRT) Using three Parameter Logistic Model. University of Technology, Mara Sarawak.

26. Mokobi T. and Adedoyin O.O. (2014). Identifying location biased items in the 2010 Botswana junior certificate examination mathematics paper one using the item response characteristics curves. *International Review of Social Sciences and Humanities*, 7(2), 63-82.

27. Boone J.W. and Scantlebury K. (2006). The role of Rasch analysis when conducting science educational research utilizing multiple choices tests. Wiley InterScience. www.interscience.wiley.com.

28. Nenty H.J., Odili J.N. and Munene-Kabanya A.N. (2008). Assessment training among secondary school teachers in Delta State of Nigeria: Implication for sustaining standards in educational assessment. *Journal of Educational Assessment in Africa,* 3, 110-123.