Dimensionality Analysis of Students' Performance in 2013 BGCSE Agricultural Examination: Implications for Differential Item Functioning

Sello Editor Moyo

(201308048)

Faculty of Education

Department of Educational Foundations

A research essay submitted to the university in partial fulfilment for the degree of Master of Education in Research and Evaluation

Supervisor: Prof H. J. Nenty

17th August 2015

APPROVAL PAGE

This research has been examined and is approved as meeting the required standard of scholarship for partial fulfilment of the requirements for the Degree of Master of Education in Research and Evaluation in the department of Educational Foundations

Positions	**Names**	**Signatures**	**Dates**
Head of Department	Dr. M. R. Dambe	_________________	____________
Supervisor	Prof. H. J. Nenty	_________________	____________
Internal Examiner	Dr S. Koloi-Keailatse	_________________	____________
External Examiner	Rev. Pro. Imo Edet Umoinyang		17/08/2015

ABSTRACT

Equity and excellence in 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 dimensionality and fairness to all learners. The population for the study was all the 12784 students' responses who sat for the 2013 BGCSE agriculture examination. The students' responses were analysed using factor analysis and IRT (1PL, 2PL and 3PL) models to examine the psychometric parameter estimates of the forty test items; dimensionality analysis and the chi square test for each test item that fitted in the three IRT models. Differential Item Functioning (DIF) analysis for each item also was done according to gender and location type using logits test for t-test significance ($p < .05$). The findings revealed that examination was not unidimensional. None of the items fit the 1PL. Only one fit the 2PL and 8 items fitted the 3PL. 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 neither unidimensionality nor 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 DIF purpose among others.

DECLARATION

The work contained in this study was done by the author at the University of Botswana between August 2014 and May 2015. It is original work of Sello Editor Moyo except where due reference is made. No part of this study should be reproduced without authorized permission of the author.

DEDICATION

I sincerely dedicated to my beloved wife Goitsemodimo, and my sons for their unwavering support, guidance, appreciation and encouragement they offered throughout my entire period of research study. I also extend the dedication of my work to my late mother Kathiku Monica Moyo for the wonderful up-bring which shaped my life.

ACKNOWLEDGEMENT

In many ways, the completion of this research paper is the first of hopefully many milestones in a career of educational research, measurement and evaluation. Before I came to University of Botswana I knew very little about reading statistics and research in education. The application of Item Response Theory (IRT) captured my soul and I am looking to the future to pursuit further in educational measurement. This is grateful and largely thanks to my Prof H.J. Nenty for his humility, expertise, remarkable guidance, support, mentorship and countless patience for setting path of my study and provided an excellent and innovative insight in my study. He is indeed the father of IRT in Africa because of his highly informed level of understanding of his field and I have benefited lot from his school of thought ranging from classroom, out-side classroom, consultancy and generosity to social life.

Sincere thanks and appreciation is extended to the BIHL for selecting me as one of the first recipients of the inaugural BIHL Thomas Tlou Scholarship. Through this scholarship you have provided all my financial support and I had made to realise my dream. I would like also to thank our eminent historian Professor Thomas Tlou for his rich legacy he left us in which the BIHL recognised him through this scholarship and hence his benevolence my dream is a reality.

Credit is accorded to O. O Adedoyin for helping me with her IRT software I used to analyse my results. The Ministry of Education and Skill Development is appreciated for providing me with the study leave. I thank my friends and colleagues for their moral support. My gratitude to a colleague Mike Ochu for proof-reading my final draft.

Special thanks are extended to my queen of my heart, my wife Goitsemodimo for the encouragement and prayers she offered throughout my study period, she made me to realise my vision. I thank you all my fellowship of Old Apostolic Church for provide the prayer for my success and the Lord, All Mighty bless you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

The Problem

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 (Nenty, 1985). 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.

Very often national examination council are set up in each country and mandated to organize testing and examination of learners at different levels of education. The process of testing entails the translation of the course content into tasks which when presented to a learner elicits from him or her the type of behaviour called for in the course objectives, such that the result of a learner's responses to the items reflects the level knowledge has acquired (Nenty, 1985).

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 is 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 assessment development 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. According to Hunyepa (2014) "it is unreasonable for the nation to feign shock every year when national examination results are announced but do not do anything about it! Chances are examination results ...disappoint next year just like years before unless something drastic is done!" (p. 12). 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.

The declining results in student performance are even a major concern at the legislative arm of the government. Thus, the former Minister of Education and Skills Development, Honourable Pelonomi Vesnsion Moitoi made a statement following a decline at all levels of national examination results which include Primary School Leaving Examination (PSLE), Junior Certificate Examination (JCE) and Botswana General Certificate of Secondary Education (BGCSE). The Minister alluded that the disturbing trend has been occurring since 2005 (Moeng, 2013). This is a worrying trend and need to be reversed urgently.

The honourable Member of Parliament from the opposition, in response asked the Minister whether it was not appropriate to inform the nation that an educated and informed nation was no longer an attainable pillar for Botswana given this decline. Despite this, the issue of examination results is a national issue as it affects all. It is therefore demands an improvement measure in the quality of education and this is what most people hold dear to their hearts.

It is of interest also to note that the government of Botswana, for many years has made concerted efforts to increase its budgeting allocation to education. For instance, in the last two years, the Ministry of Education and Skills Development (MOESD) was allocated the largest amount of 26 percent in 2013 and P9.26 billion or 27.8 percent in 2014 of the

Ministerial recurrent budget (Republic of Botswana, 2013; 2014). Despite such a large fund allocation on education, the expected maximum performance is not recorded for education.

Another 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 (Utlwang, 2006). 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 (Utlwang, 2006).

Nevertheless, one may be tempted to speculate that the alarming decline in public examination results over the last couples of years started immediately after 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 (Thobega and Masole, 2008) 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.

Table 1 is an example of large scale achievement assessment results in which the examinations over years may have some set of items which are often maintained and secured for repeated use. If the precision and difficulty level of the trend items remain unstable over repeated administration, it also has some impact on the examination quality. The phenomenon is commonly known as item parameters drift (IPD). According to Well, Subkoviak and Serlin (2002) IPD poses a threat to measurement applications that require a stable scale.

In practice, IPD may occur due to a change in curriculum coverage, implementation of the state reforms, to mention a few. The shift transpired in Botswana education from COSC to BGCSE reform remains a challenge regarding the threat of IPD. The threat is not known on the basis of proper investigation and interpretation of student's achievement in national examinations. Again, apart from the information of pass or fail, Grade A or B, Grade C or D as shown in Table1, examination data carries a lot more in educational measurement. Instead of discarding data after grading as meaningless; it should be put into further inquiry for diagnostics and prognostics purposes. Thus, such data can produce useful insight into issues that have an impact on education of Botswana, only if critically analysed.

In the public view, the score a learner makes in an examination is the main focus of interest deemed as useful output of examination. However, the collection response by each learner to an item in the public examination is hardly considering useful way to transform the declining students' achievement. Nevertheless, valid decision to enhance quality of examination could be extracted from analysis of learners' response to an item in the examination data (Table 1).

The other observation made by the researcher is that the educational standards for the evaluation instruments for public examination purposes in Botswana like other Africa countries has for years been dominated by Classical Test Theory (CTT) despite its weaknesses (Nworgu & Agah, 2012). Amongst the various challenges confronting examination bodies in Africa is the demand for the development of new instrument. That is the equivalent quality of test items for the candidates taking the yearly examinations which calibrated on the basis of CTT. The application of CTT for a long time has been identified to have some limitations like, sample dependency, poor precision and undue focus on aggregate (sum) scores that deny one the opportunity of determining how the examinees performed on a specific item (Ojerinde, Popoola & Onyeneho, 2012). Critics mainly originate from comparative view of the student's attainments in public final examination on the basis of the tests which are developed as parallel form tests administered separately to candidates on different years and times. The problem with item statistics of CTT is that test analyses are sample dependent but characterizing items in IRT has taken care of this problem. With IRT, items can be calibrated without reference to the items by a test of fit of the model. Once items have been shown to fit the model, such items are chosen for test construction. Item calibration can be sample-free through controlling the influence of the ability level on the sample-bound item scores (Umobong, 2004).

Table 1

Comparison of 2009, 2010, 2011, 2012 and 2013 Agriculture BGCSE Grades

	Grades Awarded	Grade A*		Grade A		Grade B		Grade C		Grade D		Grade E		Grade F	
		No	Cum %	No	Cum %	No	Cum %	No	Cum %	No	Cum %	No	Cum %	No	Cum %
2013	13015	60	0.46	172	1.78	1261	11.47	3320	36.98	2854	58.91	2873	80.93	1815	100
2012	13269	48	0.36	148	1.48	1165	10.26	3347	35.5	2698	55.81	2898	77.65	2159	100
2011	11731	65	0.55	158	1.9	1136	11.58	3394	40.52	2591	62.6	2390	82.98	1484	100
2010	11680	77	0.66	226	2.59	1455	15.05	3771	47.34	2571	69.35	2037	86.79	1130	100
2009	13299	99	0.74	247	2.6	2087	18.29	4343	50.95	2865	72.49	2123	88.46	1167	100

According to Adedoyin and Mokobi (2013), the quality of test items in any public examinations is always examined through item analysis of examinees' responses. As such a psychological instrument cannot be assumed to provide accurate information without proper psychometric evidence to support claims of what the instrument purports to measure (Brian, Daniel & William, 2007). 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 (Lee, 2002; US Department of Education, 2004). 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.

The evaluation for assessment of dimensionality evidence of any given examination is standard practices in educational measurement. It is also very important to teachers, examiners and the public in general for decision making. According to Siamisang and Nenty (2012), a test is never better than the sum of its items, so to identify problematic items is only possible through item analysis. Assessing the dimensionality is very imperative to inform policy makers on any need of policy review, of which public examinations in Botswana as well should be more often is scrutinized through review. Many scholars (Nenty, 2004; Nworgu, 2011; Umoinyang, 2011) have recommended item response theory (IRT) approach be applied to analyse differential item functioning (DIF) as mean of contributing to test fairness. 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 interpreting the public examination, like assessing for unidimensionality and fairness of the examinations.

Background of the Study

As one seeks a solution to the recurring problem of declining performance in BGCSE examination, with particular emphasis on the multiple choices items, it is important to acquire an understanding of the background of this problem. In Botswana, the attainment of ten- year basic education programme is through completion of seven years of primary education and three years of junior certificate. The aim of the ten year basic education programme is to provide quality education that develops fully productive citizens for the 21st century. As part of the process of implementation of Revised National Policy on Education and Vision 2016 ideals, a broad, practically oriented curriculum has been developed. The curriculum provides opportunities for learners to develop technological skills that are related to the world of work (UNESCO, 2006).

The senior secondary programme is post-basic education with a practical orientation aimed at preparing learners for the world of work and community involvement. The two year broad based curriculum for senior secondary schools in Botswana takes cognizance of the nature of knowledge, the contribution that different subjects offer and the infusion of sensitive emerging issues (UNESCO, 2006). For 1994 and 1998, the transition rate from junior secondary to senior secondary education at 22 % and 37.6 % while 200/2001 was 49.1%, (UNESCO, 2005). Notably, the rate has increased year after year.

According to Brian et al. (2007), the introduction of a clearly defined testing process would enable an instructor to assess objectively, fairly and consistently the skills of each examination seeking employment. Testing has been fully accepted in most modern societies as the most objective method of decision making in schools, industries and government establishments. It is now used for admission, recruitment, promotion, placement, evaluation and guidance, research and teaching purpose among others (Emaikwu, 2012). In Botswana, education system uses high stake assessment of students through examinations. Thus, Botswana has four external examinations namely the Standard IV Attainment Test for diagnosis purpose, Primary Schools Leaving Examination (PSLE) for diagnosis purpose, Junior Certificate Examination (JCE) and Botswana General Certificate of Secondary Education (BGCSE) for selection purposes.

These examinations are developed in partnership with classroom teachers and curriculum developers, and previously administered by the Examination Research and Testing Division (ERTD) under the Ministry of Education before the establishment of BEC (Republic of Botswana, 2001). According to UNESCO (2006), the examination in Botswana used to be norm-referenced but has moved to towards criterion-referenced with effect from year 2012. The blue-print and assessment procedures appropriate for reporting the performance in subject components as well as in subject and overall grades have been developed.

Examination is assesses mostly through written papers and are designed to determine the achievement levels of the candidates. Every subject takes at least one paper. Project work and practical examinations are used for assessing aspects of the syllabus that cannot be assessed by pen and paper. The format of questions includes multiple choice, short-answer and structured questions (Republic of Botswana, 2001).

The Report of the National Commission on Education (1993) in Botswana proposed several educational reforms. One of the major reforms proposed is the localization of senior secondary school examinations. Recommendations tied to examination localization reform were: the formation of a National Examination Council; training of examiners and markers for all subjects; preparation, re-writing and adaptation of syllabi to local relevance and context; setting of examination papers; and grading system of the scripts with Cambridge as a moderating body (Revised National Policy on Education (RNPE), 1994). As a result of the recommendations, BEC was established with a general mandate to conduct school examinations and any other examinations for the Ministry of Education and to issue certificates in respect of such examinations (Republic of Botswana, 2002).

The BGCSE is the localization of the COSC. Thus BGCSE replaced COSC as a result the operationalization of the Revised National Policy on Education while Cambridge remains as a moderating body (RNPE, 1994). Such operationalization gave birth to BEC to run examination on the behalf of the Ministry of Education for the Republic of Botswana (Republic of Botswana, 2002).

BEC administers examinations of equivalent demands every year. The examinations are developed within BEC by the Directorates of Products Development and Standards with the help of the examination committees comprising of stakeholders drawn from different sectors of the MOESD. The process is executed by following the guide of test blue prints which are drawn from the current school curriculum. Scoring of the candidates scripts are done by trained examiners who all guided by the standardised marking schemes. The quality assurance of scoring is ensured through team leaders and principal examiners who sample examiners' work at regular intervals for verification (Republic of Botswana, 2002).

The move for localization of examinations was a transformative shift in the history of education system of Botswana to ensure that syllabi remain relevant to local issues and cultural aspect of Batswana as well as integrations of global village concepts. According to Utlwang's (2003) study on teachers' perception on examination localization, there is teachers' beliefs that locally made syllabi relate better to the local content, culminate in setting more valid and reliable examinations.

Agricultural science is one of the subjects that are examined at senior secondary level. In effect agricultural education is part of BGCSE programmes. The assessment syllabus is designed to assess candidates who have completed two-year course based on the senior secondary school and aims to assess achievement at all level of ability. Candidates are assessed in ways that encourage them to show what they know, what they understand and what they can do (Republic of Botswana, 2001). The structure of examination consists of three components which are Paper 1(Multiple Choice Items), Paper 2 (Structured Items) and Paper 3 (Course Work) (Thobega & Masole, 2008).

There are three main assessment objectives which include (i) knowledge with understanding, (ii) handling information, application and problem solving and (iii) practical and investigative skills. Republic of Botswana (2001) described each of the objective follows:

1. Knowledge with Understanding

Candidates should be able to demonstrate agricultural knowledge and understanding in relation to:

1.1 correct uses of terms, symbols, quantities and units of measurement;

1.2 correct references to facts, concepts, laws and principles;

1.3 safe agriculture practices that prepare students for a productive life;

Questions assessing these objectives will often begin with one of the following word: *define, describe, outline, state, etc.*

2. Handling Information, Application and Problem solving

Candidates should able to use oral, written, symbolic, graphical, tabular, diagrammatic and numerical presentations to:

2.1 locate, select, organise and present information from a variety of sources;

2.2 translate information from one form to another;

2.3 use information to identify patterns, report trends draw inferences, make predictions and propose hypothesis.

2.4 present reasoned explanations for phenomena, patterns and relationships;

2.5 solve problems of a quantitative and qualitative nature.

3. Practical and Investigative Skills

3.1 Practical skills and techniques

Candidates should be able to:

3.1.1 understand and follow instructions

3.1.2 choose and use suitable techniques, equipment and materials safely and correctly;

3.1.3 record observations, measurements and estimates

3.2 Practical investigations carried out by students candidates should be able to

3.2.1 identify problems and plan an investigation;

3.2.2 organise and carry out an investigation;

3.2.3 interpret and evaluate observations and experimental data;

3.2.4 draw conclusion and make recommendations

Weighting of Assessment Objectives

For the overall assessment, the approximate weight of the assessment objectives should be as reflected in Figure 1

1. Knowledge with Understanding	30%
2. Handling information, application and problem solving	40%
3 Practical and investigative skills	30%

*Figure 1.*Weighting of the assessment objectives

The written papers are Components 1 and 2 while Component 3 is course work. Component 1 consists of 40 multiple choice items, each with four possible responses whereas Component 2 is compulsory short-answer questions which is marked out of total 60. Both Components 1 and 2 are set from assessment objectives 1 (weighting; 30%) and 2 (weighting; 40%) as shown in Figure 1. The course paper continuous assessment of candidates' practical work which consists of practical tests and project work. This component is marked out of 75 and its assessment is based on objective 2 and 3. The weighting of the papers are; Component 1 (40 %), Component 2 (40 %) and Component 3 (20 %), (Republic of Botswana, 2002). It is through these examinations that the achievements of students in every subject are provided. It is therefore imperative to unpack the underlying factors or explain variations in many test scores to enhance measurement specialist for the decision making about the testee's performance. The examination data determine what psychological traits, ability, construct or attribute underlies testees' performance on the test and the quality of the examination (Nenty, 2008). So an analysis of the test dimensionality enables examiners to reveal whether the test instrument has assessed the testees on one and only one ability or skill.

Theoretical Foundation

Item Response Theory (IRT)

Item Response Theory (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 (Nenty, 2004). The latent variable, expressed as theta (θ), is a continuous unidimensional construct that explain the covariance among item responses (Steinberg & Thissen, 1995). Individual persons at higher level of θ have a higher probability of responding correctly an item.

IRT model uses item responses to obtain scaled estimates of θ, as well as to calibrate items and examine their properties (Mellenbergh, 1994). 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) (Nenty, 2004).

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 (Lord, 1980). 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 (Pido, 2012). Usually, an ICC has one, two or three parameters that are called parameters.

The most popular mathematical approach to ICC is the logistic form whose graph is an S-shaped curve (Duong, 2004). An example of the logistic ICC is presented in the Figure 2. The horizontal axis represents levels of ability denoted by symbol θ and vertical axis indicates the probability of correct response to the item denoted by P(θ). Embretson and Reise (2000) ascribe several ways of interpreting the P(θ). It can be considered as a probability that a particular examinee at the ability level θ answers the item correctly. An alternative way is to interpret P(θ) as probability of an examinee’s giving a correct answer to a randomly chosen item from a population of item supposed to measure the same ability. Another way, probably the most useful is that P(θ) can be viewed as probability associated with a randomly selected examinee at ability level θ answers the item correctly.

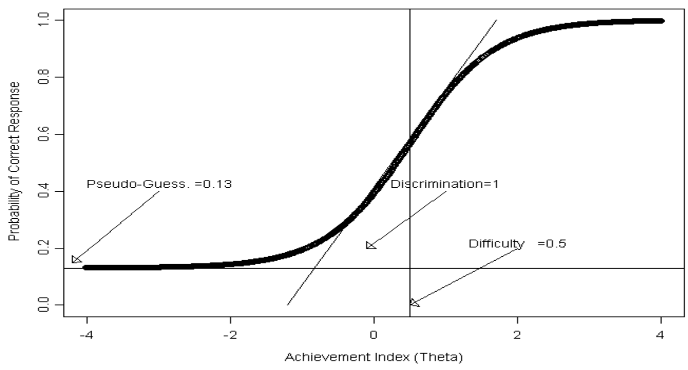


Figure 2. Item characteristic curve (ICC)

In a cognitive task, the a-parameter indicates the degrees to which examinees' response to an item varies with, or relates to their traits level or ability and it's the slope of an ICC at the point of inflection. Theoretically, a-parameter values range from negative infinity to positive infinity, and the typical a-parameter values are less than or equal 2 for MCI. A high a -value indicates that the item sharply discriminates between the examinees at the point of inflection but over a small range of θ, and a low a-value indicates that the item poorly discriminates over a wide range of the θ. Negative a-values show that examinees with low ability have higher probability of answering the item correctly than examinees with higher ability, hence such items are bad items and should be revised or discarded (Pido, 2012).

Item difficult refers to the amount of trait inherent in the item and item location parameter. It is the point of inflection on the ability scale θ at which has 50% probability of correctly answering an item. Operationally, the b-parameter represents the cognitive resistance of the item which is the amount of the trait under measurement which is needed just to overcome the task (Nenty, 2000). It corresponds to a point on the ICC where the slope of the curve is steepest. Item difficulty, in theory term, it varies negative to positive infinity. Items with b-values around the midpoint zero are of medium difficulty. Items with positive b

-value to the right of zero are difficult items and are only be answered correctly by those with high ability. Items with negative b-values to the left of zero are easy items and are answered by examinees with low ability and as well as those with medium and high ability (Pido, 2012; Nenty, 2004). While the c-parameter shows the likelihood that a person who is lacking in the trait will overcome or answer the item correctly with guessing index (Nenty, 2000; Nenty, 2004).

Assumptions of IRT

The IRT model is based on the assumption that the items are measuring a single continuous latent variable θ ranging from negative to positive infinity. IRT assumes that all items in a test must be developed to measure one and only one trait. Secondly, the IRT model assumes local independent of item responses. Thus, the performance across items in the same instrument should not be related except as a result of the influence of the trait level that they all are designed to measure (Nenty, 2004). The assumptions of unidimensionality and local independence are technically related in that items local dependence implies a separate dimension in a factor analysis. These assumptions are very useful for the determination of test dimensionality.

Briefs on IRT Models

The 3- Parameter Model

The 3- Parameter Model assumes that the three parameters are necessary for an estimate of a valid relationship between the probability of a correct response to an item and the trait level of an individual.

$$\mathbf{P(\theta) = c_i + [1- c_i]/[\ 1+ e^{-1.7a_i(\theta-b_i)}]} \qquad \mathbf{(1)}$$

This gives the probability of an individual with ability **θ** responding desirably or correctly to an item **i** with a difficulty of $\mathbf{b_i}$, a discrimination index of $\mathbf{a_i}$, and a vulnerable-to-guessing index of $\mathbf{c_i}$. The letter **e** that appears in the formula is the base of natural logarithm and is approximately equal to 2.718. The constant value **1.7** , which appears in the formula, is a scaling factor, applied to scale the logistic function [L(0,16679)] to approximate the normal function (0, 1.0) as closely as possible. It allows the logistic of frequency function to approximate the normal frequency function as closely as possible (Nenty, 2004).

The 2- Parameter Model

The 2- Parameter Model assumes zero vulnerability-to-guessing parameter for items; hence the parameter c_i is not necessary for the estimation of a valid relationship between the probability of a correct response to an item and the trait level of an individual. In that case, the logistic function is:

$$P(\theta) = 1/[\,1+ e^{-1.7a_i(\theta-b_i)}] \qquad (2)$$

This give the probability of an individual with ability θ responding correctly to item with a difficulty of b_i and a discrimination index of a_i

The 1- Parameter or Rasch Model

The development of the Rasch Model was independent of the development of IRT models in which one parameter logistic model is one. It places more demands on test item construction effort than the other two models. Items that fit the model should differ only in difficulty. This implies that the a_i and c_i are not necessary for the estimation of a valid relationship between the probability of a correct response to an item and the trait level of an individual.

$$P(\theta) = 1/[\,1+ e^{-1.7(\theta-b_i)}] \qquad (3)$$

This gives the probability of any person with θ responding correctly to an item **i** with a difficulty index of b_i.

Comparison of the Models

The three PL models assume a single trait underlying examinee performance. Each model uses a logistic function to link examinee ability θ to probability of correctly responding to the item P(θ). For each item, ICC is constructed by plotting the probability of getting the correct answer over the ability scale. This relation is non-linear because the variables are unbounded but probability is bounded (Duong, 2004).

Although all three models use the logistic functions, they differ in the number of item parameter they allow to vary. For instance for the 3PL model, ICCs differ with other two PL (1PL and 2PL) as it assesses the slope, location and lower asymptote. With IPL and 2PL models, the probability ranges from 0 to 1. In 3PL models, the lower asymptote might be greater than 0 and the range of probability is not from 0 to 1, but from c to 1 (Duong, 2004). The 3 parameter model is the most general model, and the other two IRT models (two- and one-parameter models) can be considered as models subsumed under three-parameter model (Lord, 1980).

Assessment of the Model Fit

According Duong (2004), the judgment of the suitability of a model for solving particular measurement problems can be based on three kinds of evidence: the appropriate of the model assumptions; the invariance of the obtained model parameters, and the accuracy of the model predictions. For unidimensional IRT models, the checking of the model assumptions should focus on four fundamental assumptions: unidimensionality, equal discriminating power, minimal guessing, and nonspeeded test administrations (Hambleton, 1989). Evidence on the appropriateness of model assumption can be used to select IRT models.

To check the invariance of ability, one method is administering examinees more than one set of items where items in each set have various level of difficulty. The score of examinees in all the tests should be correlated because the expected ability score for each examinee does not depend on the choice of item if the model fits the test data. One of the useful methods in checking model prediction is to analyse item residuals. In this method, after a model is chosen and parameters are estimated, predictions about performance of various groups of examinees are made. Predicted results are then compared with actual results. If the residuals are small, it is reasonable to accept the accuracy of the model predictions (Duong, 2004). The benefit of IRT models is to improve the development of test items and guide the examiner to balance items not just by testing higher order thinking. Rather, if an item misfit is diagnosed, it is due to poor item quality. For example confusing distractor in the multiple choice paper and hence such item is removed from test form or replaced (Mellenbergh, 1994).

According to Royal and Puffer (2013), evaluation of dimensionality of the multiple choice examination involves tests of fit, principal components analysis (PCA) of standardized residual correlations and data-to-model fit both overall and by individual item analysis. Generally, chi-square fit statistics are required to be nonsignificant (Bonferroni adjusted). Residual fit statistic are expected to be within a given range ± 2.5 for individual items and with a mean fit residual value close to 0.0 and standard deviation approaching 1.0 (usually < 1.4) for summary statistics (Velde, Beaton, Hogg-Johnston, Hurwitz & Tennant, 2009). This is helpful to discern if multiple dimensions are present and exactly where these dimensions might be in the dataset.

In addition, in setting a guide to determine model-data fit, Baker (2001) asserts that, "if the value of the obtained chi-square (or index) is greater than a criterion value, the item characteristic curve specified by the values of the item parameter estimates does not fit the data" (p.51). Gruijter and Kamp (2002) suggested that item(s) that does not fit a chosen model should be dropped from a given instrument or revise for subsequent use. Nworgu and Agah (2012), Adedoyin (2010) and Ene (2005) used chi-square test with probability greater than the alpha level of 0.05 significant level to selected items that fit 3PLM, 1-, 2- and 3PLM and Rasch model respectively. It is therefore desirable to employ some statistical methods like chi-square goodness-of-fit test to evaluate the degree of IRT model-data fit/misfit. That is in this study detailed analysis and interpretation of the 2013 BGCSE agriculture examination results will go a long way in performing an exploratory diagnosis.

Assessment of Test Dimensionality using Factor Analysis

Factor analysis (FA) is another analytic tool used to provide evidence regarding fit and unidimensionality, such as scree plots and eigenvalue-based indices (Reckase, 1979). FA used to determine the underlying structure of a measuring instrument and at times it is used to investigate the nature of the underlying factors in an existing scale.

Principle component analysis (PCA) together with eigenvalue plots is a common way to evaluate test dimensionality and has been used for decades (Lord & Novick, 1968; Hattie, 1985). The percentage of total variance explained by the first principle component is often regarded as an index of unidimensionality. The higher percentage of total variance of the first principle component accounts for, the closer the test is to unidimensionality. One downside with eigenvalue plots is that there is no statistical index available to decide the number of underlying dimensions. Various criteria have been proposed to solve the problem. Reckase (1979) recommended that a percentage of 20 or more of the total variance explained by the first principle component is necessary for the data to be viewed as unidimensional. Similarly Lord (1980) suggested checking the ratio of the first to the second eigenvalue, and compares that with the ratio of the second to any of the other eigenvalues. Kaiser (1970) suggested retaining any components with eigenvalues larger than 1.

In addition, PCA provides analytic tools for exploring model-data fit used to explore hypotheses regarding invariant measurement. However single model-data fit index can detect all of the possible sources of misfit (Reckase, 1979). Model-data fit is sample-dependent, and the key question in judging fit is: How good is good enough? There is no definitive statistical answer to this question, but various indices (including FA) can provide evidence to support inferences regarding invariance within a particular context (Randall &Engelhard, 2010).

Differential Item Functioning (DIF)

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" (Nenty, 2008, p.53). These are the most likely sources of differential item functioning. According to Nenty (2010), a measurement is effective to the extent that it sustains generation of valid scores, hence, an accurate estimation of the ability or trait under measurement. This is attained only and only if the item of the instrument developed evoke from each testee the only ability under measurement. Scores from a test which fail to represent the truth on what was being measured and hence cannot sustain fair and valid decision making, for educational purposes. Nenty, moreover, emphasises that decision made and programme or policies developed based on evaluation and research that resulted from the analysis of such score fail to provide solution to education programme because to begin with. Such scores do not represent the truth.

DIF is apparent, subsequent item bias analyses (e.g., content analysis, empirical evaluation) are needed to provide evidence to declare item bias. According to Holland and Wainer (1993; cited in Monahan, McHorney, Stump & Perkins, 2007), in DIF analyses after adjusting groups for overall performance with regard to measured trait, they are compared on item performance. In other words, in assessing test-takers response patterns to specific test items, or doing DIF, the comparison groups (e.g., males vs. females) are initially matched on the underlying construct under consideration like in this study is agriculture achievement. This helps researchers or test developers to determine whether item responses are equally valid for distinct groups of test takers (Zumbo, 1999). According to Nenty (2010) a test item is said to be item bias if: There is differential performance for individuals of the same ability but from different groups; it lowers the average score of a particular group; it contains language or content that is differentially familiar for different subgroups of the examinees and it contains sources of difficulty that are irrelevant or extraneous to the construct being tested; the test item, item stem, test instruction or distractor is not good enough or/and can be understood in more than one way by the examinees.

By extension Nenty (2010) attested that a test is biased if it contains clues that would increase the performance of one group over another. There are no equal learning opportunities so much that one group is more exposed to the information being tested than the other group; there are no equal access to relevant textbooks equipment, instruments, laboratories and workshops; there is no equal scoring format for the test takers; it contains offensive elements that would insult any group of examinees on the basis of their personal characteristics. Some of the sources of item bias are inadequate item formulation, language, test wiseness, poor item translation, when an item is invoking additional traits or abilities, when the topic of the test items are not in the curriculum of one of the subgroup, etc. (Zumbo, 1999).

When misinterpretation of test is brought to bear fruit in the education system, there is the possibility that some examinees will be unfairly treated. Hence, wherever this situation occurs, the test bias is presumed to exit in the measurement of ability. DIF, which is among them is test bias in measurement has become a heated, complex and pronounced issue in the western countries and most developing countries including African states are also becoming conscious of the concepts even though there is low use of psychological test in those developing nation (Joshua, 2005; Oche, 2012). DIF items are seriously a threat to the validity of the instruments to measure the trait levels of members from different populations or groups. In national examination instruments for Botswana, too are not exclusive to DIF mainly because some items may have reduced validity for between-group comparisons due to DIF. This is because their scores may be indicative of a variety of attributes other than those the scale is intended to measure.

Measurement 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 (Hambleton & Slater, 1997). IRT offers a powerful method of investigating item bias – also referred to as differential item functioning (DIF). The ICC of an item plots the probability of the “correct” response against the magnitude or level of the underlying (latent) trait being measured (Adedoyin, 2010). Osterlind (1983) describes ICCs as the most elegant of all the models to tease out item bias. IRT models assume unidimensionality, local independence of items and the fact that the probability that an examinee will respond correctly to a particular item depends upon the shape of the curve and the individual’s level with regard to the underlying construct being measured. However, it is not dependent upon the individual’s performance relative to any particular group (Osterlind, 1983).

The use of ICCs for DIF detection concerns the comparison of differences in the ICCs for different subgroups. Only two groups can be compared at a time, but a particular sample can be divided into various subgroups for such comparisons. “The area between the equated ICCs is an indication of the degree of bias present in a considered test item” (Osterlind, 1983, p. 61). Although both groups are on essentially the same scale, they need to be equated by means of a linear transformation. The difference in scales is caused by the fact that theta is arbitrarily defined as having a mean of zero (0) and a standard deviation of one (1) in each separate group (Owen, 1992b, Royal &Puffer, 2013, Royal & Puffer, 2012, Linacre, 1994).

Once the theta scales have been equated, meaningful comparison of ICCs of the two groups is possible. Procedures for decision making may include simply inspecting the graphs visually or calculating the actual differences. Limits or cut off criteria are arbitrary because no specific significance test is available to test difference between estimates of area (Osterlind, 1983). According to Eng and Hoe (n.d.) and DeBeer (2004) an item is discarded for DIF on the basis of following criteria:

IRT: c-values: $c > 0.3$,

IRT: a-values: $a < 0.80$,

CTT: $rit < 0.3$ unless IRT $a > 1.0$

DIF: area between the ICCs of any of the two or more DIF comparison groups > 0.5481

Visual inspection of the form of DIF, together with the magnitude of the area between the graphs of the two groups compared, is usually combined to determine whether an item should be flagged as biased (DeBeer, 2004). A distinction is made between uniform DIF and non-uniform DIF. In uniform DIF, the probability of answering an item correctly for one group is consistently lower than that of the other group. That is in the case of uniform DIF, the curve of the one group consistently falls higher (in terms of the Y axis) than that of the other, suggesting that the probability of passing or endorsing the item is uniformly higher for one group than for the other. In the case of non-uniform DIF, the curves cross at some point, implying that the item is more discriminating for one group than for the other (Hambleton, Swaminathon & Rogers, 1991; Adedoyin, 2010).

The other modified method of testing DIF is through group difference for ability 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 (Nenty, 2013). The t values are also used to detect DIF and hence were useful in assessing the dimensionality of agriculture examinations.

Conceptual Framework

In this study, conceptual framework was developed to establish the test of fit with IRT Models (1-Parameter 2-Parameter and 3-Parameter models) to identify the quality of items and hence means to diagnosis if the examination was fairly to every learner. This framework in Figure 2 was a typical and ideally tests development process which ensures that items reflects the same thing for all examinees or else modification is required.

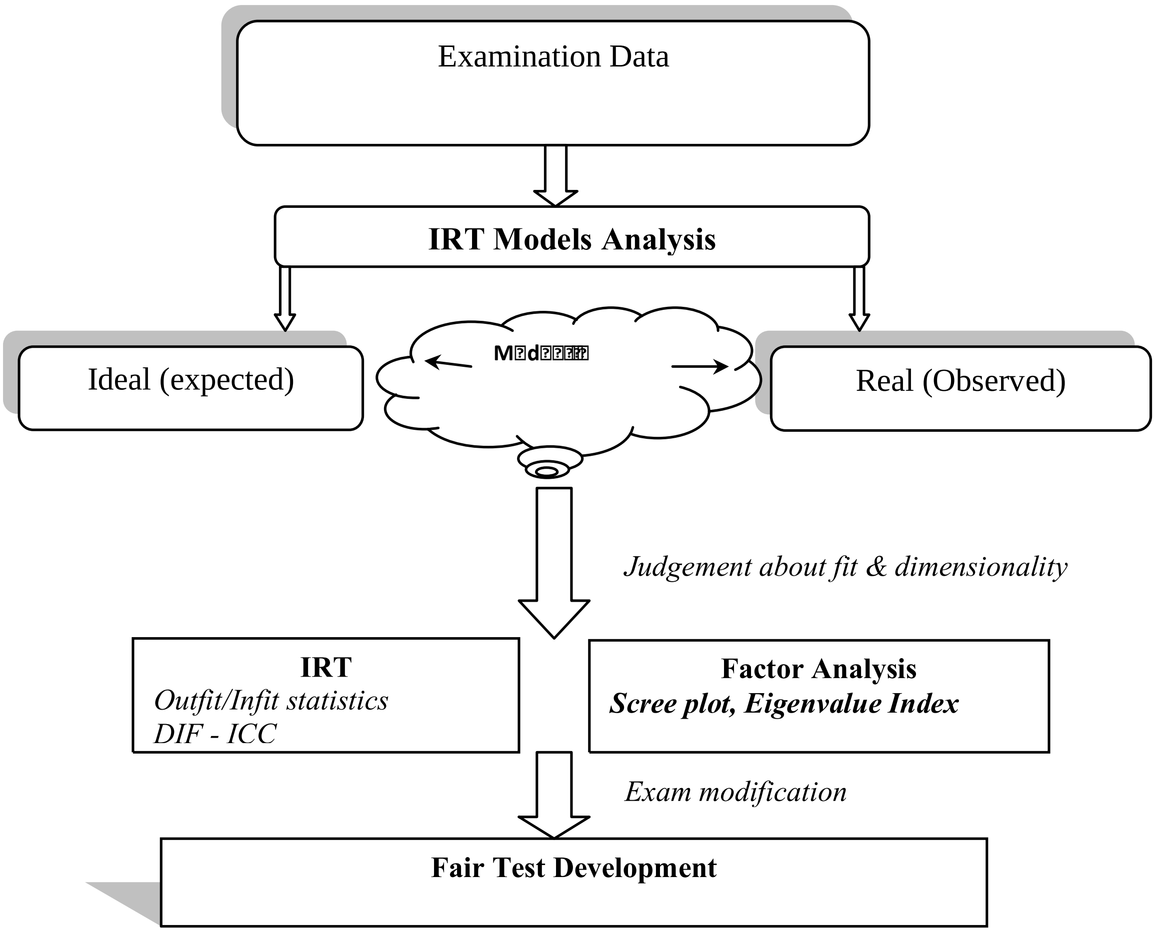


Figure 3. Conceptual framework for this study

Statement of the Problem

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 (Ministry of Education, 2013). Decline of high grades appears to be indicative of fall in standard of education. This was attested by (Thobega and Masole, 2008) who revealed that Botswana Examination Council (BEC) only review structured items and ignoring other components. 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 %) (Republic of Botswana, 2002).

For at least five years now, Botswana secondary schools students are labelled failures in their results from the national examinations. Many candidates did not pass the examinations and this has brought about fresh criticism of the Botswana educational system as irrelevant to the current generation. This issue is even worsened by making it political and the society seems to behave as if there is no solution to this problem. Some quarters of the society opt to label the students as failures. Policy makers and measurement specialist seem helpless. They seem to be bereft of corrective measure to the situation. Labelling young school-leavers failures may be a wrong approach. For one thing, some items quality in examination components are neither determined by item review nor empirically differentially item functioning. As observed by Thobega and Masole (2008), BEC only reviews the structured questions for agriculture but multiple choice items together with other components are ignored. This 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.

The painful amount is heavily felt by students who are passed into the society. If the quality of life a student or person cannot rise beyond the fairness by its educational system through valid measurement testing and examination, then begin of general downward trend in education quality of students which predicts a corresponding lowering of their life. Given that quality of life has a direct relationship with level of education, it would be maximized if the potential of every person in society is identified and developed to a maximum, only when testing and examination are fair to all learners (Nenty, 1985).

It is a cumbersome situation given the fact that psychometric evidence is hardly discussed in the public due to the technical challenge of the technocrats, even if the examination may have some items which are unreliable, invalid, unfit and highly bias. Corollary to this assertion is also the emphasis made by Nenty, Odili and Munene-Kabanya, (2009) that is existing absence of the regenerative feedback through large scale assessment is a handicap to attaining and maintaining high standard in education. Obviously, the decision made about examinations of that nature may mislead the stakeholders in education. It is also inappropriate to discuss students' achievement across different administration when education standard for assessment are partially executed merely on the basis of raw scores.

However, the use of modern way of interpreting of examination results is necessary as a means to extract meaningful information. Empirical literature reveals evidence that influences during examinee-by-item encounter distorts test result and renders its biased, hence limiting its validity, generalizability and comparability (Nenty, 2000). Thus, "the degree to which these extrinsic factors influence the testing process and hence its results differs across examinees, schools, classrooms, sex of examinee, school location (urban/rural) or districts" (Nenty, 2000, p. 48). Decisions based on test results can only be valid if the test and the testing process and interpretation are valid.

Differential item functioning is one of the most important considerations when the examination scores are to be used for selection of examinees. Hence, use of any test is that it must not be biased, rather should be fair to all candidates. Schumacker (2005; as cited in Perrone, 2006) explained DIF as a collection of statistical methods used to determine the fairness and appropriateness of examination items with regard to different groups (e.g., male and female, location, region etc) of test takers, hence aiding in the identification of biased test items. This technique (DIF) has been mostly used in research as a rather new standard in psychometric bias analysis. DIF procedures are in fact a response to the legal and ethical need to ascertain that comparable test applicants are treated equally (Jodin & Gierl, 1999).

Purpose and Objectives of the Study

In this study, the IRT models were used to assess the dimensionality for the 2013 BGSCE agriculture multiple choice items to detect differential item functioning. The purpose of this study was 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 objectives of the study are:

1. To determine the dimensionality of 2013 BGCSE agriculture multiple choice items.
2. To determine the extent to which items for 2013 BGCSE agriculture multiple choice items fit the one-parameter, two-parameter and three -parameter logistic models.
3. To establish the extent to which do gender and location influence DIF among students in the 2013 BGCSE agriculture multiple choice items.
4. To determine the consequences of dimensionality on item functioning.

Research Questions

1. What is the dimensionality of 2013 BGCSE agriculture multiple choice items?

2. What are the items of 2013 BGCSE agriculture multiple choice items that fit the 1PL, 2PL and 3PL model?
3. To what extent do gender influence DIF among students responses in the 2013 BGCSE agriculture multiple choice items?
4. To what extent does location influence DIF among students' responses in the 2013 BGCSE agriculture multiple choice items?
5. To what extent do the loading of gender-based DIF and non-DIF items influences the first factor extracted through factor analysis of 2013 BGCSE Agriculture Examination?
6. To what extent do the loading of location-based DIF and non-DIF items influence the first factor extracted through factor analysis of 2013 BGCSE Agriculture Examination?

Statement of the Hypotheses

$\mathbf{H_{A1}}$: There is significant influence between the loading of gender-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination.

$\mathbf{H_{A2}}$: There is significant influence between the loading of location-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination.

Significance of the Study

Examination results in Botswana are used as input in various decisions. DIF is one of the statistic analyses which can unearth if test question is either fair or unfair. DIF is only one of the means measurement specialists use to ensure that public tests are as fair as possible (Zieky, 2003).The significance of this study lies on the extent to which it can provide information on examination dimensionality which is linked to the fairness of the examination. The findings will contribute to the literature of educational and psychological testing in Botswana. Moreover, the researcher identified that Botswana has little literature on the empirical studies for psychometric testing regarding examination dimensions therefore this study will provide significant source of literature to the field of the educational research particularly in analysing the dimension of examination and implications for differential item functioning. The study will shed light on the fairness of examinations as well as the validity of the items.

The measurement specialists, classroom teachers, policy makers and BEC through this study would be provoked to know that test construction and examination data analysis are very important in ensuring that the public tests are as fair as possible which accord an equal treatment of test takers, diverse external input to test content, fairness review, promotion of proper test use and research on fairness issues. The findings of this study will also be of immense importance to BEC in the evaluation of the items for agricultural examinations and improving the quality of achievement examinations. In like manner, the students to whom this study is targeted would either directly or indirectly benefit. This is because if the orientation of their achievement examinations towards the fairness and they stand the chance, that is student of being benefited maximally.

Assumptions

The basic assumption underlying this study is that students who took the 2013 BGCSE Agriculture multiple choice items reflect various background and levels of ability. The researcher assumes that examination items are constructed by measurement specialists and analysed to ensure a reasonable level of validity and reliability. The researcher also assumes agricultural examination items are of quality and fair. That the relevant authority in the MOESD and BEC will offer maximum support required for assisting the researcher to carry out the study.

Limitations

Owing to the possibility of measurement error, the study was limited to the multiple choice items examinations for 2013 BGCSE. The study solely is quantitative which will apply one-parameter, two-parameter and three-parameter logistic fits for model and DIF analysis. The stringent traits of the one parameter model may affect the analysis of extreme values of the study. However, examination data is data used for this study. These are for national examination which goes through a careful validation before administration. In effect, one parameter model technique is also appropriate for the analysis.

Delimitation

The scope of the study is delimited on the basis of resource and the period of study available to the researcher as outlined in the sponsorship contract agreement. This had influenced the researcher to embark on a study in which involved secondary data. Agriculture is chosen because the researcher, by profession, is an agriculture teacher and teaches in a government school. In effect, the researcher has the expertise in the area and access to government schools examination data. Only Paper 1, multiple choice items will be considered for this analysis and it carries an equal weight like the constructed items (Paper 2).

CHAPTER TWO

Literature Review

Introduction

Several studies have been carried out on psychometric properties in educational measurement. In this study, literature review is focused on studies related to the differential item functioning of multiple choice items.

Gender and Differential Item Functioning (DIF)

DIF is a test bias analysis in measurement. It has become one of a heated, complex and pronounced issue in the western countries and most of the developing countries. Some African states are also becoming conscious of the concept, even though there is low use of psychological test in developing nation (Joshua, 2005; Oche, 2012).

On a study carried by Kalaycioglu and Berberoglu (2011) aimed 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. 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, Kalaycioglu and Berberoglu (2011) 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. This study shed light on DIF but did not spell out the DIF analysis and sample sizes used.

Hauser, Kingsbury, and Northwest Evaluation Association (2004) understudied the differential item functioning and test functioning. This study was designed to answer two key questions about the fixed portions of the fourth, eighth, and tenth grade in spring 2003 Idaho Standards Achievement Tests (ISAT) in Reading, Language Usage and Mathematics. Study was investigated on questions which follows (1) to what extent do test items exhibit substantive differences in functioning across student groups formed based on ethnicity and based on gender? and (2) what is the collective impact on test characteristics of any substantive differentially functioning items? Hauser et al. (2004) worth mentioned that DIF is not quite the same as what has commonly been called "bias." differential item functioning is an item characteristic that occurs whenever groups differ noticeably in their correct answer frequency, when achievement level is matched. It may be used as part of the process to identify biased items, but is almost always viewed as secondary to a sensitivity review. Because of its statistical nature, the researchers find out that many items that are identified as having DIF were not identified as having any difficulty in the sensitivity review (Kingsbury, and Northwest Evaluation Association,2004).

In another study, Ubi, Joshua and Umoinyang (2012) sampled 800 candidates' scripts from a pool of examination scripts of candidates who sat for the Joint Admissions and Matriculation Board's University Matriculation Examination (JAMB-UME) in Cross River State, Nigeria for the years 2002 and 2003. The purpose of the study was to assess the dimensionality of mathematics items using factor analysis. Results showed that JAMB-UME test revealed five significant dimensions and they concluded that examinations designed for selection of candidates might not be purely unidimensional, especially when items are fielded from a wide syllabus. They recommended, among others things that, since it might not be possible to set tests, particularly mathematics, that are purely unidimensional, test practitioners especially those in charge of selection examinations should endeavour to meet the principles of item construction like ideal item difficulty, high discrimination and high option distraction indices to compensate for violating unidimensionality requirement. The analysis for unidimensionality in this study would have been better if the analysis used Rasch model. Despite that the study paved the way for insightful view on obstacle to attain unidimensionality in some examination like mathematics.

Robin, Zenisky and Hambleton (2003) study was (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.

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 (Robin, Zenisky & Hambleton (2003).

On comparative method analysis for DIF, Magno (2009) 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. Magno (2009) 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.

In a study in Botswana, Adedoyin (2010) carried out a study using 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.

In a more recent study in Botswana, Motshabi and Nenty (2012) took an ethnicity-related differential item functioning (DIF) of English language examination items in primary schools in Botswana. Using three differential item detection methods the study analysed responses by primary school learners to items in the 2008 Primary School Leaving Examination (PSLE) in English language in Botswana. To ensure appropriate ethnic-mix of the subjects for the study, purposive sampling was used to select 2,587 out of a total population of 41,471 pupils who took the examination in 2008.

Both primary and secondary data collected from and on these subjects were analysed using the Mantel-Haenszel, standardization and maximum likelihood item analysis procedures to determine convergence in item functioning across three ethnic groups. Among examinees of the same ability in English language about 48% of the 60 items in the examination were found to function differentially across learners from the different ethnic groups. This was seen to have significant implications for fairness in cross-cultural assessment, teaching and learning in Botswana (Motshabi & Nenty, 2012). Relatively, the sample size was small.

Differential Item Functioning Analysis by Location

Zumbo and Gelin (2005) worked on a new methodology for item and test bias studies: "moderated differential item functioning (DIF). This technique expands the DIF methodology to incorporate contextual and sociological variables as moderating or mediating effects on the DIF. The study explored differential domain functioning (DDF). The focus of interpretation for this test is on the "domain" rather than the item. This moderated DDF effect is demonstrated on a multiple-choice and constructed-response provincial assessment test that was designed to match a specified mathematics curriculum. Participants were 45,728 grade four students, 45,022 Grade seven students, and 43,525 Grade Ten students in British Columbia, Canada. The data from these participants was narrowed down to create four contrast groups of communities that reflect differences in contextual variables: rural low-income, rural affluent, urban low-income and urban affluent.

Gender DDF was explored using a general linear statistical model. After statistically matching males and females on their mathematical ability, gender DDF was moderated by the contextual variables. Thus, those "moderation" approaches allowed one to investigate the effect of sociological, community-based contextual variables that may help one understand the complex functioning of DIF in large scale testing. In other words, what the authors are advocating is to take a more "sociological" and "ecological" approach to help educators understand differences in item and test performance (Zumbo & Gelin, 2005). Zumbo and Gelin also made an effort to introduce new concept of DIF which include sociological and community based contextual variables, and it applied large sample to enhance its external validity.

Amuche and Fan (2014) carried out a study for 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.

Amuche and Fan (2014) 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 (Amuche & Fan, 2014). 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.

Uruemu and Adams (2013) investigated items bias using differential item functioning approach in relation to school type (private and public schools), school location (urban and rural schools) using National Examinations Council (NECO) economics questions for 2010. The research design employed in this study was a comparative. The study sample comprised students in Delta State, Nigeria. Four hundred and forty seven (447) students were used. The test contained 60 items which was administered to the candidates. Logistic regression was used to analysis the data. The research findings showed that out of sixty items in NECO economics questions 10 items were biased in relation to school type and 8 items in relation to school location. The implication of the findings was that NECO economics examinations questions have presences of differential item functioning (DIF) hence it was not fair other candidates. From the findings, it was then recommended that test experts and developers should explore the use of DIF approach to detect biased items. The sample size is small, when taking in account the number of students seat for examination in Nigeria.

Eng and Hoe (n.d.) used the three-parameter logistic model to identify DIF items in Mathematics Paper 1 of ‘Sijil Pelajaran 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. 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. This study shed some light that location has an impact on item functioning, though the sample size of study was not indicated.

Wen (2014) applied multilevel structured in educational testing as students were nested within teachers who were nested within schools, and which further nested within districts to assess DIF on the achievement test. DIF analyses in multilevel data were investigated via a simulation study with an emphasis on studying DIF at the teacher-level only and at both student and teacher levels, followed by the impacts of DIF on ability estimation. The multilevel Rasch models were used to detect DIF at different locations in both exploratory and confirmatory manners. The finding showed that the power was larger when conducting confirmatory analyses than for exploratory. The magnitude of DIF at both levels and the proportion of manifest groups at both levels were two most influential factors on the power of detecting of DIF. However, no influential factors found had impacts on ability estimates (Wen, 2014). It is unfortunate that Wen did not show the sample size used.

Mokobi and Adedoyin (2014) also carried out 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. 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 (Mokobi & Adedoyin, 2014).The study shed some light that location has significant impact on DIF, hence it is of useful foundation to the current study.

Summary of Literature Review

One of the important considerations in selection and use of any test which has emerged from literature is that test must not be biased, that is test must be fair to all candidates. If measurement specialists want to use the results of tests and measures to make decisions, then, we have to conduct research to ensure that our measure is not biased. Literature further revealed that DIF explores if there is any biased in items of any kind to favour a particular group. In almost all the studies reviewed some item were bias. For instance, Adedoyin (2010) used IRT approach to detect gender biased item in public, the study found that some item were biased.

The interrogation of literature has assisted the researcher in the current study to identify the consistency among and the gaps left by other researchers. A reasonable number of studies on psychometric analysis used varied populations and some used only single method for analysis. Most studies reviewed appear to have used secondary education level related examinations and in just small sample single year, mostly in Nigeria. It is of a need to carry out such study in the context of another country like Botswana. In Botswana, only two studies so far, were carried out to identify DIF using IRT approach to detect bias. As Adedoyin (2010) put it in her words " this study only tried to identify the test item bias for 2004, Botswana mathematics Paper 1, there is need to detect gender bias test item other subjects in any public examination, through use of item response approach (ICC curves)" (pp. 397 – 398).

The current study is in line with some the reviewed studies. It will focus on the national examination for senior secondary school to assess its dimensionality. That is assessing the dimensionality of 2013 BGCSE agriculture multiple choice items using IRT models analysis.

IRT models are perceived as an elegant means to determine the dimensionality of the assessment tool. It is for this reason the IRT models was chosen to diagnose the 2013 agriculture examination for unidimensional investigation. The current study is not yet explored in Botswana context. Therefore, this distinguishes it from other literature studies which were reviewed.

CHAPTER THREE

Research Design and Methodology

Introduction

This chapter describes the method that was used to assess the dimensionality of 2013 BGCSE Agriculture examination items. The chapter presents and justifies, the methodological approach of the study as well as the procedures that were used to contribute a solution to the problem. The chapter constitutes of the research design, population of study, sampling procedures, sample, instrument, data collection and data preparation. The chapter also includes sections which address the study's ethical considerations and procedures for testing hypothesis.

Resea rch Methodology

This study sought to assess the dimensionality of 2013 BGCSE agriculture examination items. It was guided by the philosophy of the positivist paradigm. It is the approach which commands all inquiries which are value free. In the positivist paradigm, the researchers only use scientific method of gathering data in order to achieve objectivity and neutrality during the inquiry (Chilisa & Preece, 2005). The paradigm emphasises on generalization and opens to replication by other investigators (Ary, Jacob & Sorensen, 2010). The positivist approach underpins its techniques on quantitative orientation because it involves "selecting and studying samples from the population to discover the relative incidence, distribution and interrelations of sociological and psychological variables" (Kerlinger & Lee, 2000, p. 599) and generating findings, through statistical analysis for inference to the population. For the reason that the current study attempted to examine the dimension property of multiples choice items for agricultural education in BGCSE using an existing BEC test data, hence the positivist paradigm was appropriate for the study.

Research Design

The research design serves as a path finder for truth or valid findings. It is a means of controlling extraneous variables. Basically research design is the science (and art) of planning procedures for conducting studies so as to get the most valid findings (Kelinger & Lee, 2000). 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.

The main aim of the exploratory design is to identify the boundaries of the environment in which the problems, opportunities or situations of interest are likely to reside and to identify the salient factors or variables that might be found there and be of relevance to the research. Even though, the technique does not provide final solution to the problem deterioration of education quality including the high grades decline, at least it can establish a base for solution. Thus, it was hoped that exploratory discovering significant variable in educational measurement among variables and to lay the background for later, more systematic and rigorous testing of hypotheses (Kelinger & Lee, 2000). Exploratory technique attempted to lay groundwork for deeper understanding of the phenomenon that will lead to future studies.

Using exploratory technique to detect items which were DIF across examinees was a great strength to diagnose the validity and underlying dimension of 2013 BGCSE Agriculture Examination and this assumed to provide the researcher's direction on ways to improve and to monitor the instrument from one test administration to the next (Boone & Scantlebury (2005). Consequently results from this study would provide insight of the dimensionality of Botswana examination (in this case 2013 BGCSE Agriculture multiple choice item). That is to say the outcome of the current study would be of immense contribution to sensitizing fairness of the examination and ultimately sought a solution improves the quality of the next examination. The foregoing was attainable only after the exploring the situation, therefore an exploratory design was appropriate for this study.

Population of the Study

The population of a study is the target group with characteristics about which the research is interested in studying. Such members are group of individuals that have one or more characteristics in common that are of interest to the researcher (Nenty, 2013; Kerlinger & Lee, 2000). The population of this study consisted of students responses who attempted the 2013 BGCSE Agriculture Examination. The items were taken by those students who successfully underwent lessons in agriculture for two years at senior secondary school. 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. Agriculture is an optional subject in senior secondary school. Yet, it is one of universally applied science subjects. 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.

Sampling Procedures for Student Responses

In this study, every student's responses to multiple choice items in BGCSE Agriculture were given equal chance to be selected and this enhanced the external validity of the study. In effect, students' academic records in agriculture examinations for 2013 were available. The researcher 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.

Location was also another distinguishing parameter in which student responses to an item could have influenced the examination dimension. With effect from April 2010, the Botswana introduced ten (10) educational regions. From these educational regions, schools are classified according to location as rural, peri-urban and urban. 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. According to Kerlinger & Lee (2000) the maximum of the variance is through widening out the difference, if any, among levels of independence variables. In the current study, rural and urban locations were used to maximize the variance in the analysis of detecting DIF among the items in the examination.

Instrument

The current study was based on the use of secondary data on students. Student records have information on performance of every student in different subjects. Among the records was the agriculture multiple choice item examination which has all the responses of students to every item. From the researcher's experience, the agriculture examinations consist of three Components tested in as separate papers: Component 1 consists of 40 multiple choice items, each with four possible responses, Component 2 is compulsory short-answer questions which is marked out of total 60 and Component 3 practical paper (70 marks) and hence current study was focusing on Component 1only.

There was therefore no need to develop instruments for this study. 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.

Data Collection

Before embarking on the collection of data for conducting research, a letter was obtained from the Department of Educational Foundation, University of Botswana indicating the purpose and importance of the proposed research study. The researcher submitted that letter to government official in MOESD, Educational Planning and Research Services Department to seek permission to conduct the research project in the country. The importance and relevance of the study to Botswana including risks to the participants was explained. The research plan, researcher’s curriculum vitae and time schedule for collection of data were also attached to the letter.

BGCSE Agriculture Examination results were obtained from BEC. Student-by-item responses for year 2013 were collected as text data files on the Microsoft Excel platform using a re-writable compact disc. The biographical data at the senior secondary schools namely location and gender were represented the independent variables for individual student. The dependent or response variable is the individual score for BGCSE Agriculture items.

Ethical Considerations

The research uses the secondary data. It therefore did not require a direct participation of students. In effect, there exists no need for provision of consent form. However, as stated earlier the permission to use students' academic records for research was acquired from the MOESD. That permission was also a gatekeeper which introduced the researcher to BEC authority to access students' data. Such data contains confidential information about students and schools. As a way of maximising confidentiality, the data were coded and names of students or schools did not appear on the data. Test scores were kept safely and confidential to adhere to recognition of the protection of human rights (American Educational Research Association, 2005).

Data Preparation

The data collected for the study were prepared before the final processing. The data for this study was purely quantitative. The data include, BGCSE agriculture examination for Paper 1 consists of 40 multiple choice items from two major domains taught in the BGCSE syllabus. The data were imported from BEC. The codes for the school and test items were retained as used by the BEC while each student was given a unique individual code in five digit 00001, 00002 etc. A correct response to an item represented as 1, while an incorrect response represented by 0. Thus, the highest number of 1 was representative score of this paper. The maximum an individual examinee can score on this paper is 40. The independent variable of gender was coded as male = 1 and female = 2. The other independent variable of location type of students was coded as urban = 1, peri-urban = 2 and rural = 3.

Operational Definition of Research Variables

The variables of this study were gender, location and students' performance to an item. Student performance represents students' responses to items in which a score was awarded one to correct item and zero to the incorrect respond. Gender was classified into two independent levels: male and female. Similarly, location had two categories: only data from urban and rural levels.

Data analysis procedures

The process for the analyses of the data collected for the study involved the use of both descriptive and inferential statistics. Factor analysis using Statistical Package for Social Sciences (SPSS) version 21 was done. The exploratory factor analysis extracted the dimensions underlying performance on examination. Model fit for 1-parameter, 2-parameter and 3-parameter was done to establish the extent to which the 40 multiple choice items fit the models using BILOG-MG V3.0. The results of the analysis was tested for significance using chi-squares test to determine the level of item fit for each of the model. IRT 1-parameter logistic method using t-value was 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.

Further analysis was done through exploratory factor analysis on those items which were identified as DIF and non-DIF to extract their first factor and the loadings of the items on it. Independent samples Mann-Whitney t- test Statistical Package for Social Sciences (SPSS) with the latest version was applied to establish the significant different between gender based DIF and non-DIF items on the first factor exacted. The use of independent t-test was preferred for these variables because each of them contained one independent variable (gender based DIF and non-DIF items) and one continuous dependent variable, (first factor score) for non-parametric statistic.

An independent samples Mann-Whitney t- test was also applied to test the influence of location on factor scores of both biased and non biased items. This was justified by student location made of one independent variable (urban and rural) and one continuous dependent variable (loadings of items in the first factor) therefore it was a non-parametric statistic. The data analyses procedures are shown on Table 2. The significance level for testing the hypotheses was set at .05 for all statistical tests.

Table 2

Procedure for Testing the Research Question and Hypothesis for 2013 BGCSE Agriculture Examination

Questions/Hypothesis	**Dependent Variable**	**Independent Variable**	**Type of Statistical Analysis**
What is the dimensionality of 2013 BGCSE agriculture multiple choice items?	Item that load on each extracted factor	Extracted factors or components	Descriptive analysis

What are the items of 2013 BGCSE agriculture multiple choice items that fit the one-parameter, two-parameter and three-parameter logistic model?	Test items		Chi-square to test for item fit
To what extent do gender influence DIF among students in the 2013 BGCSE agriculture multiple choice items?	Test items	Gender-type for examinees	IRT- 1Parameter Logistical Method using t-values
To what extent do location influence DIF among students in the 2013 BGCSE agriculture multiple choice items?	Test items	Location-type for examinees	IRT- 1 Parameter Logistical Method using t-values
H_{01}: There is no significant difference between the loading of gender-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination.	Test items	Extracted first factor	Independent Samples Mann-Whitney t-test
H_{02}: There is no significant difference between the loading of location-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination.	Test items	Extracted first factor	Independent Samples Mann-Whitney t-test

CHAPTER FOUR

Data Analysis and Interpretation of Results

Introduction

In this chapter, it is important to restate right from the onset the procedure through which the research data were analysed. In fact, the study is purely quantitative, which sought to assess the dimensionality for the agriculture multiple choice items and to detect differential item functioning. That is examining the dimensionality of 2013 BGCSE Agriculture Examination as means of generating information with which contribution could be made to the improvement of test development. Statistical data analyses were carried out to answer on some questions and test some hypotheses. The study was guided by the research questions below:

1. What is the dimensionality of 2013 BGCSE Agriculture multiple choice items?
2. What are the items of 2013 BGCSE Agriculture multiple choice items that fit the 1PL, 2PL and 3PL model?
3. To what extent does gender influence DIF among students' responses in the 2013 BGCSE Agriculture multiple choice items?
4. To what extent does location influence DIF among students' responses in the 2013 BGCSE Agriculture multiple choice items?
5. To what extent do the loading of gender-based DIF and non-DIF items influence the first factor extracted through factor analysis of 2013 BGCSE Agriculture Examination?
6. To what extent do the loading of location-based DIF and non-DIF items influence the first factor extracted through factor analysis of 2013 BGCSE Agriculture Examination?

Data characteristics

In this study the researcher used a total of 12734 students' responses in the 40 multiple items for agriculture examination. As described in the Table 3, 47% of the total respondents were males and 53 % were girls. The students' responses were also classified into school location where the attended namely, urban (24.1%) peri-urban (44.9%) and rural (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).

Table 3

Students Response Characteristics

Parameter		**Frequency**	**Percent**
Sex			
Male		5995	47.10
Female		6739	52.90
Total		12734	100.00
Student by location			
Urban		3067	24.10
Peri-urban		5722	44.90
Rural		3945	31.00
Total		12734	100.00
Statistics		**Sex of Students**	**Total Scores**
N	Valid		12734
	Missing		0
Std. Deviation			5.13
Variance			26.35
Skewness			.326
Std. Error of Skewness			.022
Kurtosis			-.17
Std. Error of Kurtosis			.043
Minimum			5.00
Maximum			39.00
Mean			21.93

Q1: What is the dimensionality of 2013 BGCSE agriculture multiple choice items?

To answer this question, the responses of the students on the 40 multiple-choice items of BGCSE Agriculture Examination were subjected to factor analysis. This is a very important step prior to performing DIF analysis. Factor analysis was performed to determine whether or not a dominant factor existed among all items as it was expected that the BGCSE agriculture examination would come up with one dominant factor. This factor would represent the construct underlining the agriculture learning domains measured by the examination. A Principal Component Factor Analysis (PCFA) was conducted to determine the underlying structure of the data. The initial eigenvalues were greater than 1, which are considered significant.

Table 4 shows the percentage variance accounted for by each of the variables. Nine factors had eigenvalues over Kaiser's criterion of 1 and in combination explained 10.05 %. Thus, the first eigenvalue was 3.82 greater than the next eight eigenvalues (1.358, 1.146, 1.115, 1.102, 1.069, 1.040, 1.025, and 1.018) respectively. The first factor explained only 10.05 % of the variance in the data set. The second factor explained 3.40 % of the remaining variance. The rest of the variance was explained by the other 38 factors with 7 factors each having an percentage of variance between 2.80 and 2.50, then 30 factors each having a percentage of variance of between 2.49 and 1.80. These last 30 factors were eliminated because they did not contribute to a simple factor structure and failed to meet a minimum criteria of having a primary factor loading of greater than 1 eigenvalues.

A scree plot was produced to determine whether unidimensionality could be inferred. The scree plot should provide a convenient way of visualising a dominant factor from principal component analysis. An inspection of the scree plot of Figure 4 showed a high visual representation of relatively the first factor, but which accounts for only 10.05% of the total items variability.

In effect the overall analyses indicated that nine distinct factors with eigenvalues bigger than 1.0 underlay in 2013 BGCSE Agriculture Examination (see Table 5) and they accounted only 32.34 % cumulative variance (see Table 4).

Table 4

Total Variance explained of the 2013 BGCSE Agriculture Examination

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.021	10.053	10.053	4.021	10.053	10.053
2	1.358	3.395	13.447	1.358	3.395	13.447
3	1.146	2.866	16.313	1.146	2.866	16.313
4	1.115	2.787	19.099	1.115	2.787	19.099
5	1.102	2.755	21.855	1.102	2.755	21.855
6	1.069	2.674	24.528	1.069	2.674	24.528
7	1.040	2.599	27.127	1.040	2.599	27.127
8	1.025	2.563	29.691	1.025	2.563	29.691
9	1.018	2.545	32.236	1.018	2.545	32.236
10	1.000	2.499	34.735			
11	.995	2.486	37.221			
12	.984	2.459	39.681			
13	.974	2.434	42.114			
14	.962	2.405	44.520			
15	.957	2.393	46.912			
16	.942	2.356	49.268			
17	.939	2.347	51.616			
18	.934	2.336	53.951			
19	.924	2.311	56.262			
20	.919	2.298	58.560			
21	.909	2.274	60.834			
22	.905	2.262	63.096			
23	.897	2.242	65.337			
24	.887	2.217	67.555			
25	.882	2.204	69.759			
26	.876	2.190	71.949			
27	.870	2.174	74.123			
28	.856	2.140	76.263			
29	.847	2.118	78.381			
30	.830	2.075	80.456			

31	.826	2.066	82.522
32	.820	2.049	84.571
33	.815	2.037	86.607
34	.808	2.020	88.627
35	.795	1.989	90.616
36	.778	1.945	92.561
37	.773	1.931	94.492
38	.754	1.885	96.378
39	.730	1.824	98.202
40	.719	1.798	100.000

Extraction Method: Principal Component Analysis

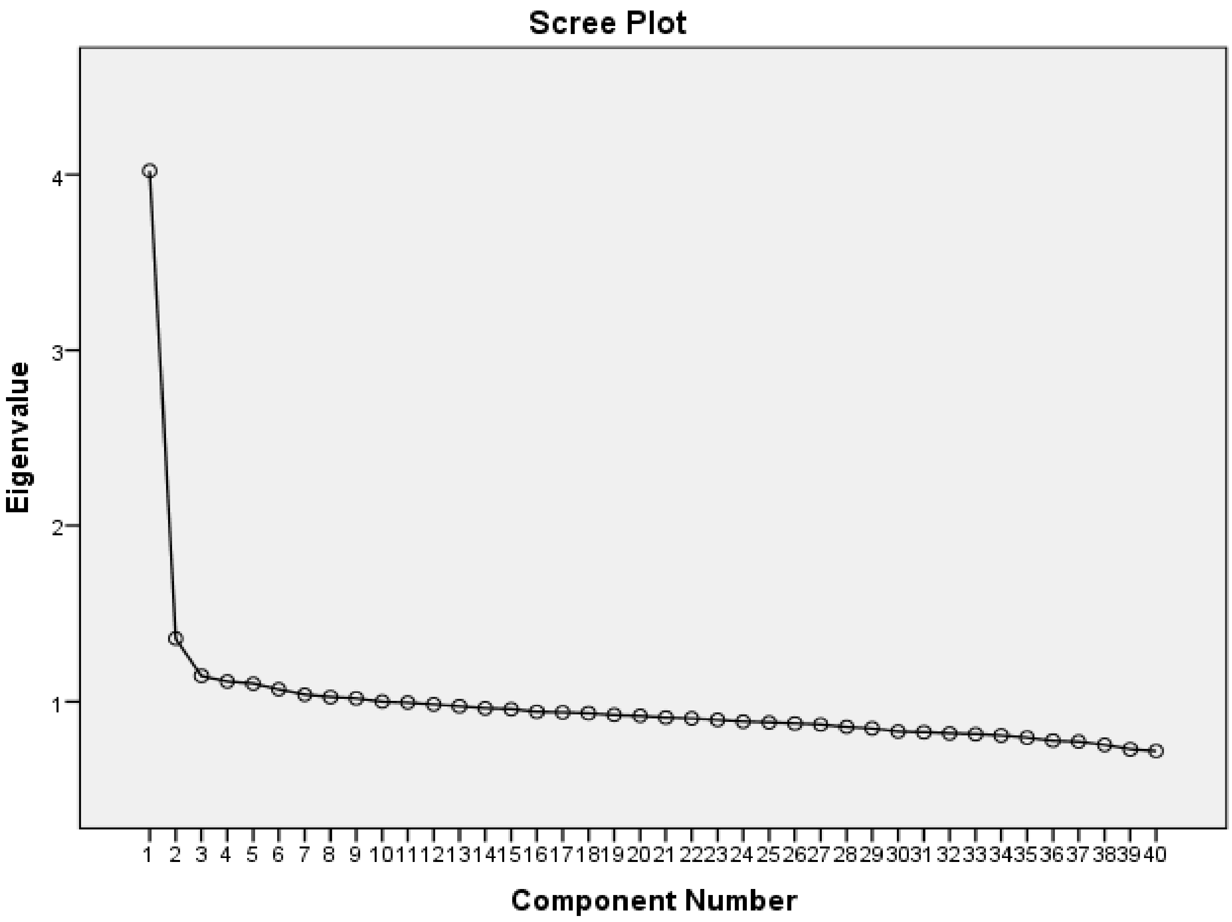


Figure 4. Scree plot showing dimensionality of 2013 BGCSE agriculture items

Table 5

2013 BGCSE Agriculture Examination Test Items Loading in Each Factor

Factor /Name/Item		Loading	Factor /Name/Item		loading
1	Farming Management Practice		**6**	Identification	
1	Q1	.238	1	Q2	-.349
2	Q3	.364	2	Q11	-.239
3	Q4	.384	3	Q13	.227
4	Q5	.311	4	Q22	.469
5	Q6	.453	5	Q23	.390
6	Q7	.460	6	Q34	-.218
7	Q8	.445			
8	Q9	.392			
9	Q10	.284			
10	Q11	.313			
11	Q12	.252			
12	Q14	.335			
13	Q15	.481			
14	Q17	.378			
15	Q19	.321			
16	Q20	.287			
17	Q21	.329			
18	Q23	.227			
19	Q24	.323			
20	Q25	.309			
21	Q26	.358			
22	Q27	.321			
23	Q28	.325			
24	Q30	.418			
25	Q31	.290			
26	Q32	.420			
27	Q36	.311			
28	Q37	.221			
29	Q38	.431			
30	Q39	.454			
31	Q40	.346			
2	Breeding/ Genetic Principles		**7**	Maintenance	
1	Q4	-.332	1	Q18	.345
2	Q10	-.219	2	Q22	-.207

	3	Q12	.215
	4	Q13	.242
	5	Q17	-.382
	6	Q19	-.257
	7	Q25	.282
	8	Q27	.220
	9	Q28	.400
	10	Q29	.359
	11	Q31	.285
3	Classification		
	1	Q1	.399
	2	Q2	.249
	3	Q10	-.355
	4	Q11	-.367
	5	Q16	.309
	6	Q19	.300
	7	Q21	.360
	8	Q22	-.251
4	Environmental/ Land management		
		Q1	.329
		Q10	.347
		Q22	.231
		Q25	-.201
		Q28	.310
		Q29	.509
		Q32	-.217
5	Marketing		
	1	Q2	.398
	2	Q12	.205
	3	Q18	.214
	4	Q28	-.345
	5	Q29	-.425
	6	Q33	.335
	7	Q34	.292
	8	Q35	.292
	3	Q23	-.401
	4	Q27	.225
	5	Q33	-.345
	6	Q27	.225
	7	Q33	-.345
	8	Q35	.377
	9	Q37	.282
8	Gross margin/ Profit		
	1	Q2	.241
	2	Q13	.296
	3	Q18	.237
	4	Q22	.255
	5	Q27	.215
	6	Q33	-.415
	7	Q34	.415
	8	Q35	-.370
9	Precaution in enterprises		
	1	Q11	-.213
	2	Q16	-.594
	3	Q18	-.255
	4	Q33	.383
	5	Q34	.275
	6	Q35	.262
	7		

Q2. What is the items of 2013 BGCSE Agriculture multiple choice items that fit the one-parameter, two-parameter and three-parameter logistic model?

To answer the question of whether 2013 BGCSE Agriculture Examination items do fit IRT models as a means to assessing the evidence of unidimensionality. The utility of the IRT model is dependent upon the extent to which the given responses fit this model. To determine whether the test item fitted the model, a chi-square test was run on the data set using BILOG-MG V3.0 item analysis computer programme to establish whether the items fitted the 1PL, 2PL and 3PL models. Table 6 showed the results of the chi-square statistics. The chi-square goodness of fit analysis showed that none of the items fitted the 1PL model because their residuals variances were statistically significant. With a 2PL model, only 1 item fitted model that is Item 27 in which, its residual variance was not statistically significant.

The chi-square values from the 3PL model, it is evidence that thirty-two items representing 80% of the total items in the test were statistically significant through their residuals variances and hence do not fit 3PL because. The table also indicated that 8 items representing 20% of the total test were not statistically significant and this means that they fit the 3PL because their residuals variances were not statistically significant. For the 3PL model Item 18 and 32 were omitted from the calibration as its initial slope was less than - 0.15.

Table 6

Chi-Square Test of Fit in 2013 BGCSE Agriculture Examination using 1PL, 2PL and 3PL IRT Models

ITEMS	1PL			2PL			3PL		
	Chi-square	Df	*P*	Chi-square	df	*p*	Chi-square	Df	*p*
1	151.30	9	.000	42.20	9	.000	25.10	9	.003
2	543.30	9	.000	117.90	9	.000	46.10	9	.000
3	70.20	9	.000	35.80	9	.000	45.70	9	.000
4	394.60	9	.000	130.00	9	.000	263.80	9	.000
5	20.50	9	.015	39.50	9	.000	38.10	9	.000
6	114.70	9	.000	76.50	9	.000	14.40	9	.109**
7	207.50	9	.000	122.80	9	.000	14.90	9	.094**
8	150.50	9	.000	80.50	9	.000	117.80	9	.000

9	440.60	9	.000	55.20	9	.000	60.30	9	.000
10	459.30	9	.000	60.70	9	.000	33.90	9	.000
11	204.10	9	.000	18.80	9	.027	21.90	9	.009
12	69.300	9	.000	43.60	9	.000	53.00	9	.000
13	281.90	9	.000	74.40	9	.000	59.40	9	.000
14	83.40	9	.000	25.60	9	.002	12.00	9	.211**
15	74.80	9	.000	76.40	9	.000	17.20	9	.046
16	292.00	9	.000	86.90	9	.000	92.60	9	.000
17	341.80	9	.000	24.30	9	.004	23.90	9	.005
18	581.10	9	.000	158.20	9	.000			
19	25.50	9	.003	22.50	9	.007	37.10	9	.000
20	172.30	8	.000	29.30	8	.000	55.40	8	.000
21	256.70	9	.000	80.00	9	.000	109.40	9	.000
22	30.00	9	.000	32.60	9	.000	10.00	9	.353**
23	114.40	9	.000	56.40	9	.000	23.00	9	.006
24	25.00	9	.003	19.80	9	.020	6.70	9	.664**
25	191.30	9	.000	67.80	9	.000	99.70	9	.000
26	112.40	9	.000	81.70	9	.000	113.10	9	.000
27	68.20	9	.000	8.00	9	.538**	3.90	9	.918**
28	354.70	9	.000	60.90	9	.000	53.60	9	.000
29	146.90	9	.000	55.30	9	.000	46.10	9	.000
30	83.30	9	.000	100.20	9	.000	12.70	9	.174**
31	682.60	9	.000	148.20	9	.000	117.30	9	.000
32	809.30	9	.000	92.50	9	.000			
33	154.30	9	.000	51.90	9	.000	56.70	9	.000
34	122.70	9	.000	67.20	9	.000	12.80	9	.171**
35	137.20	9	.000	34.60	9	.001	36.00	9	.000
36	189.60	8	.000	111.40	9	.000	18.80	9	.027
37	349.50	9	.000	58.10	9	.000	32.20	9	.000
38	90.20	9	.000	36.50	9	.000	39.60	9	.000
39	795.50	9	.000	228.30	8	.000	235.40	9	.000
40	552.10	9	.000	167.80	9	.000	70.90	9	.000

** The item selected with probability great than the alpha level of .05 significant level

Q3. 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) (see Table 7).

Table 7 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 Items 3, 4, 11, 12, 14, 16, 21, 24, 29, 36, 31 and 38. While 17 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 (see Table 8). 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.

Table 7

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 (b-value)	SE	Logit (b-value)	SE	Logit	SE	
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 8

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

Q4.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 also used 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 (see Table 9). 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 10). 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 9

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 (b-value)	SE	Logit (b-value)	SE	Logit	SE	
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*
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 10

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

H_{01}: There is no significant difference between the loading of gender-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination.

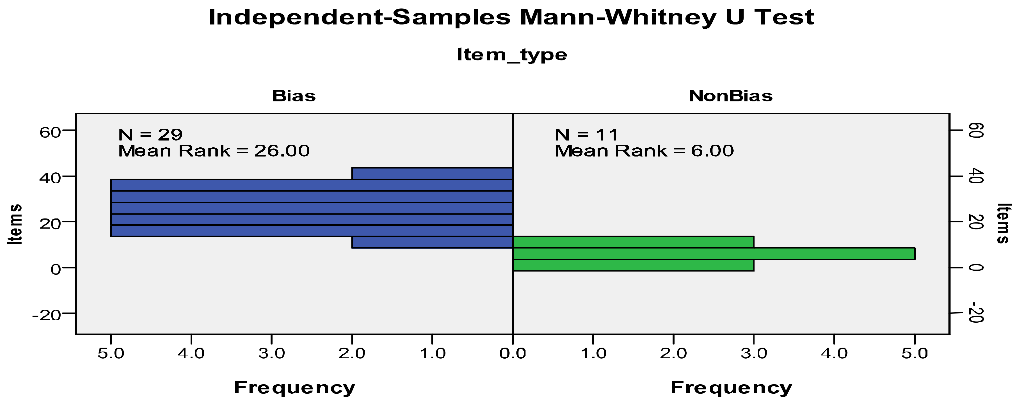
To test this hypothesis, 11 of non-DIF items were subjected to factor analysis in order to observe the loading of items on the first factor (see Table 11).The first factor analysis had an eigenvalue of 1.63 and this factor explained 23.27 % of the variance in the data set. In case of the DIF items, the first factor extracted an eigenvalue of 3.82 hence, it accounted for 9.55 % of the data set. The loadings on the first factor by DIF and non-DIF were tested for significant difference ($\alpha = 0.05$). This hypothesis was tested by performing independent sample Mann-Whitney U test. The test indicated that there was significant difference between the loading of gender-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination, (U = 319.000, z = 4.831, p = .000). This means that the first factor for gender based non-DIF items explained 23.27% of the variance of all the items, and this is greater than that of the first factor for DIF items which accounted for 9.55% of the variance. The null hypothesis was rejected and the alternative hypothesis retained.

Table 11

Analysis for Loadings of Gender based DIF and non-DIF Items on the First Factor Exacted of 2013 BGCSE Agriculture Examination

Non-biased items	Factor 1	Biased Items	Factor 1
Item_10	.461	Item_31	.477
Item_2	.457	Item_40	.444
Item_28	.431	Item_39	.443
Item_37	.420	Item_16	.417
Item_13	.382	Item_21	.369
Item_1	.366	Item_4	.347
Item_35	.355	Item_29	.333
Item_34	.319	Item_36	.327
Item_19	.210	Item_23	.326
Item_27	.139	Item_7	.323
Item_17	-.030	Item_14	.318
		Item_6	.311
		Item_33	.306
		Item_12	.299
		Item_38	.294
		Item_25	.288
		Item_15	.284
		Item_26	.266
		IteM_3	.263
		Item_22	.246

		Item_5	.211
		Item_30	.167
		Item_24	.166
		Item_20	.146
		Item_8	.086
		Item_11	.065
		Item_9	-.037
		Item_32	-.086
		Item_18	-.117
Eigenvalues	**1.629**		**3.822**
% of Variance	**23.273**		**9.554**



Total N	40
Mann-Whitney U	319.000
Wilcoxon W	754.000
Test Statistic	319.000
Standard Error	33.014
Standardized Test Statistic	4.831
Asymptotic Sig. (2-sided test)	.000
Exact Sig. (2-sided test)	.000

Figure 5. Independent samples Mann-Whitney U test of gender-based DIF and non-DIF items on the first factor exacted of 2013 BGCSE Agriculture Examination

H$_{02}$: There is no significant difference between the loading of location-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination

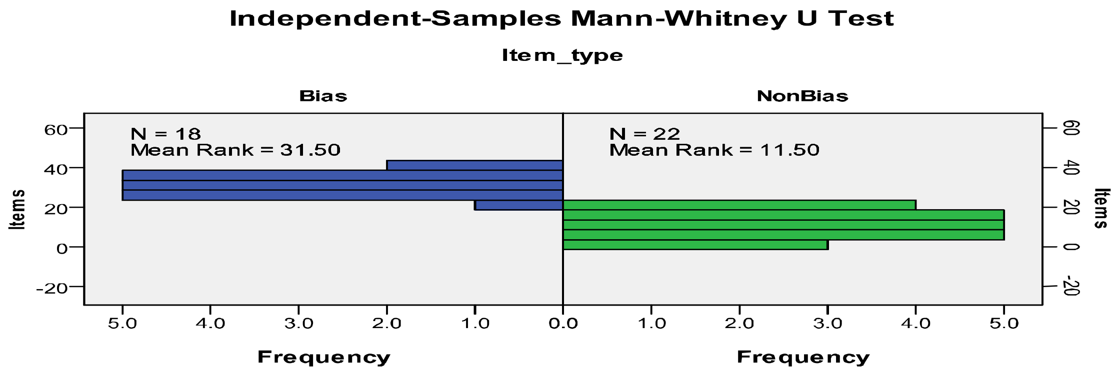
Before testing this second hypothesis, 22 non-DIF items were subjected to factor analysis to extract the loading of items on the first factor (see Table 12).The first factor for the analysis has an eigenvalue of 2.294. Hence this explained 10.429 % of the variance in the data set. For the DIF items, the first factor extracted the eigenvalue of 2.427 and hence accounted for 13.484 % of the data set. The two first factor loadings of the DIF and non-DIF were then tested for significant difference (α = .05). They were tested by performing independent sample Mann-Whitney U test. The result indicated that there was significant difference between the loading of location-based DIF and non-DIF items on the first factor extracted through factor analysis of BGCSE Agriculture Examination, (U = 396.000, z = 5.383, p = .000) (see Figure 6). This means that the first factor for location non-DIF items which accounted 13.48 % variance of all the items, and this is greater than that of the first factor for DIF items accounted 10.43 % variance. The null hypothesis was rejected and the alternative hypothesis retained.

Table 12

Analysis for Loadings of Location based DIF and non-DIF Items on the First Factor Exacted of 2013 BGCSE Agriculture Examination

non-biased items	Factor 1	Biased Items	Factor 1
Item_2	.520	Item_1	.415
Item_7	.375	Item_3	.240
Item_8	.163	Item_4	.378
Item_12	.324	Item_5	.237
Item_13	.423	Item_6	.365
Item_14	.373	Item_9	-.088
Item_18	-.124	Item_10	.483
Item_19.	.211	Item_11	.070
Item_20	.116	Item_15	.341
Item_22	.279	Item_16	.479
Item_23	.388	Item_17	-.010
Item_24	.184	Item_25	.333
Item_26	.303	Item_28	.468
Item_27	.135	Item_33	.353
Item_29	.359	Item_36	.361
Item_30	.192	Item_37	.479

Item_31	.503	Item_39	.472
Item_32	-.103	Item_40	.487
Item_34	.372		
Item_35	.403		
Item_38	.328		
Item_21	.387		
Eigenvalues	**2.294**		**2.427**
% of Variance	**10.429**		**13.484**



Total N	40
Mann-Whitney U	396.000
Wilcoxon W	567.000
Test Statistic	396.000
Standard Error	36.783
Standardized Test Statistic	5.383
Asymptotic Sig. (2-sided test)	.000

Figure 6. Independent samples Mann-Whitney U test of location based DIF and non-DIF items on the First Factor Exacted of 2013 BGCSE Agriculture Examination

CHAPTER FIVE

Discussion, Implications, Conclusions, and Recommendations

Introduction

This chapter discusses the findings of the study on dimensionality analysis of student's performance in 2013 BGCSE Agricultural Examination as implications for DIF based on the outcome already interpreted in the previously chapter. An overview of the study was presented in relation to research questions. The discussion was carried out by bringing together the findings of this study, findings of the other related studies reviewed in chapter two and the related theories presented in the first chapter. Having discussed the findings, the implications of the findings were presented and conclusion made. Finally, in this chapter the recommendations were presented and suggestions for further study were made.

Discussion

Findings of the study focused on the application of the IRT model to assess the dimensionality for the 2013 BGCSE Agriculture multiple choice items and to detect DIF. Forty items were explored by factor analysis and thereafter, the 1PL, 2PL, and 3PL were applied using BILOG-MG V3.0 yielded item parameter estimates that are very useful to test developers, users and researchers.

Dimensionality of 2013 BGCSE Agriculture Examination

In assessing dimensionality of a set of 2013 BGCSE Agriculture items, it was discovered that the examination is not unidimensional. The first factor did not meet Reckase (1979) recommendation that a percentage of 20 or more of the total variance explained by the first principle component is necessary for the data to be viewed as unidimensional. That is, a factor analysis on the inter-item correlation matrix should show that the first factor accounts for at least 20 of the variance of the unrotated factor matrix or second the eigenvalue of the first factor should clearly exceed that of the second factor (Reckase, 1979). The answer to the research Question 1 as revealed in Tables 5 and subsequently Figure 4 showed that there was no evidence of unidimensionality. It appears that the agriculture examination was multidimensional rather than unidimensional.

Even though the total variance was very small, it confirms the findings of Ubi, Joshua and Umoinyang (2012) who stressed that examination design for selections of candidates might not be purely unidimensional, especially when items are fielded from a wide syllabus. Like the aforementioned researchers alluded, agriculture in general is an applied science subject. It has a wide breadth of syllabus. For instance the BGCSE Agriculture multiple choice items are constructed on the contents which ranges from mathematics, chemistry, physics and biological concepts as reflected in the assessment objectives for the syllabus (Republic of Botswana, 2001). These have contributed to the items to measuring different things and not only one thing. In addition, since the 2013 BGCSE agriculture was multidimensional and hence it appeared not appropriately to be analysed using IRT models. However, it was further subjected to IRT analysis to see if the findings did corroborate the exploratory factor analysis already made.

Mode Fit for One-Parameter, Two-Parameter and Three-Parameter Logistic Model

The result presented in Table 5 for Question 2 showed how the items of agriculture fit the 1PL, 2PL and 3PL models. The chi-square goodness of fit statistics revealed that none of the items fitted the 1PL model and only one fit the 2PL model. This implied for 1PL and 2PL models, all the 2013 BGCSE Agriculture items were not invariant measurement items except only Item 27 for 2PL model, for which the was a fit. In other words, items for agriculture examination were neither unidimensional nor locally independence as confirmed earlier through exploratory factor analysis stage. Thus, it appeared that through 1PL and 2PL models analysis items in the agriculture national examination multiple choice items were not measuring one and only one trait. This can also be explained through the infringement of local independence because the items performances across the examination were related. Therefore, trait level was not the only influence being measure by the agriculture examination (Nenty, 2004).

Despite the unfit of items on the 1PL and 2PL models, other scholars do consider to opt for other models which are less stringent when exploring model fit of items regarding unidimensional. This also corroborated by (Reckase 1979) who attested that no single model-data fit index can detect all of the possible sources of fit or misfit. To respond to that the same 40-items were subjected to 3PL model analysis and this revealed that 20 % of the total items fitted and 80 % did not fit the model. From Gruijter and Kamp (2000) suggested that, item(s) that do not fit a chosen model should be dropped from a given instrument or revise for subsequent use. With the 3PL model, only 8 items fit the model and hence given this were appropriate items in measuring student ability items in agriculture. Even with use of less stringent model, the 2013 BGCSE Agriculture, the fit analysis results remained unsatisfactory, hence one is tempted speculate that the 2013 BGCSE Agriculture assessment instrument held lot to be desired as far as IRT scrutiny is concerned.

In effect, the remaining 32 items represented by 80% of the total items were required to be dropped or revised from the agriculture examination. The unfit items were indicative of bad items and hence not suitable for national examinations unless revised critically to correct their fault. This finding was in line with that of Nworgu and Agah (2012); Adedoyin (2010); and Ene (2005) who applied chi-square test with probability greater than the alpha level of .05 significant level to selected items fit models they used in their studies respectively.

Differential Item Functioning by Gender -Type

The findings of this study showed that 29 items out of the total of 40 items 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 (Emaikwu, 2012). 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 of Kalaycioglu and Berberoglu (2011) and Robin, Zenisky and Hambleton (2003) who detected DIF on some items favoured males' students over females in some items. 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 attested by Nenty (2010), a 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. Thus boys might have had a sound command of language related to agriculture which enabled them to outperform their females' counterparts. Culturally appears boys to be better than girls in agriculture, not as subject of study in school, but as farming practices. This empowers boys at 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 Thobega and Masole (2008) who observed that BEC only reviewed the structured questions for agriculture and ignored the multiple choice items. 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 Nenty, Odili and Munene-Kabanya (2009) stressed that the existing absence of the regenerative feedback through large scale assessment is a handicap to attaining and maintaining high standard in education.

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 that by Amuche and Fan (2014). 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 (Amuche & Fan, 2014).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 students 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 some likely 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 by Nenty (2008) 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. One of these factors mentioned above might have had an influence of students in rural and urban perform differently in some items despite their same ability level.

In the contrary Mokobi and Adedoyin (2014) in their study revealed the existence of location biasness in mathematics examination was shifted towards the students who attended schools rural areas. 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 Eng and Hoe (n.d.) in their study DIF items favoured urban group.

Exploratory Factor Analysis for the Loading of Gender and Location-Based DIF and non-DIF Items

The study had found out that there was significant influence on the loading of gender-based DIF and non-DIF items on the first factor extracted through factor analysis of 2013 BGCSE agriculture examination. Factor one is assumed as a combination of items loadings which assessed achievement in agriculture examination. The result, in fact, reflected that the dominant factor accounted for fairness of the items was greater dormant factor for unfair items. It was vital to have non-DIF items first factor had greater share of the item variance accounted for examination because it was through that factor assumed as an indicative measuring agriculture achievement.

The significant influence was also found on the loading of location-based DIF and non-DIF items on the first factor extracted through factor analysis of 2013 BGCSE Agriculture Examination. This finding was quite interesting and yet surprising to a certain extent, that the location-based DIF significantly explained the difference of dominant factor than non-DIF items. This had forced the researcher to speculate that agriculture multiple choice items for the examination was explained significantly by the systematic error variance other than the construct validity of the agriculture achievement.

Obviously, the decision based on the findings like this is dangerous not only to students alone but also to the nation. If one attempts to interpret student's performance by school location as students in rural performed better than student attended in urban school, it is a wrong interpretation. The first factor which accounted for total variance did not measure the desirable construct validity rather it measured systematic error variance described through DIF and such interpretation is prone to mislead the nation. This is consistent with Nenty (2010) who justified that score from a test which fail to represent the truth on what was being measured and cannot sustain fair and valid decision making for educational purposes. The findings of the study here could not be related to other studies because there has been very little literature review on loading of location-based DIF and non-DIF Items on the first factor extracted through factor analysis.

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. Through the dimensional analysis of the examination, it was found that the agriculture examination was not unidimensional and very few items fitted IRT Models. This meant that during the ability-by-task interaction during test taking by the student, there were some demands by some items that that provoked behaviour or trait other than that under measurement (agriculture achievement) hence those were a source of multidimensionality. The findings have also revealed that gender and location were significant sources of DIF in the agriculture examination.

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. In the analysis of loadings of the first factor extracted gender-based DIF items was significant and location-based DIF items also was statistically significant. However, the location-based DIF items loading of the first factor was explained systematic error. Lastly, the results of this study, as it explored the national assessment tool, showed that 2013 BGCSE Agriculture Examination was neither unidimensionality nor fair to all students.

Recommendations

The findings and observation of the study showed that 2013 BGCSE Agriculture as a national examination was not assessing student knowledge of agriculture validly and fairly through the assessment tool used. On the basis of the above conclusion and implications, the following recommendations are made:

- Botswana Examination Council's (BEC) department of the Directorates of Product Development and Standards should commit itself to constructing items that fit objective measurement models like Rasch, 2-parameter or 3-parameter models and reducing any form of DIF in items including gender and location bias in national examination.
- 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.

- BEC should conduct frequency fresher courses for teachers which emphasis on fairness and other qualities of test items development.
- Using test scores to enhance quality of items. BEC should take it a yearly task to analyse responses to test items and develop test item bank that are derived from the quality assurance practices.
- BEC should take a deliberate decision to intensify reviewing of items including multiple choice items to determine the extent to which each item meets the assumptions of the IRT model under consideration. This would enable them to produce quality items for criterion-referenced decision which is what BEC is currently using in grading students. The concerns of the reviewers should:
  (i) Given the table specification, are the items in the test well spread with regards to the nature and type of behaviour they are measuring?
  (ii) How well do the items cover the domains of contents or indicators of behaviour under measurement?
  (iii) Does the item construction, setting language, idea or interest portrayed, picture/diagram used, relevance and illustration tend to undue advantages or disadvantage to a particular group of testees over the other group (s)?
- Teacher training institutions should expose pre-service teachers into the test development which meet IRT assumptions particularly unidimensional and local independent. It should be stressed to trainee teachers that the syllabus is central to all assessment and advocate IRT test analysis which will also detect DIF items.
- The Botswana government through the Ministry of Education and Skills Development should solicit for donor funding that would be especially to teachers and examiners on test construction or item writing and modern test analysis. This

would accord them to be in position to produce valid, reliable and fair assessment tools which are bias free.

Recommendation for Further studies

- Further study should be directed towards analysis of Parametric Comparability of items in 2004–2013 BGCSE Agriculture Examinations.
- Another should focus on analysis of item distraction to assess DIF, because distribution of distract has the potential for DIF which can influence the performance of a student.

References

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.

Adedoyin, O.O., & 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.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999*). Standard for educational and psychological testing.* Washington, D.C: American Education Research Association.

American Educational Research Association, (2005). AREA code of ethics. Washington

Ary, D., Jacobs, L. C., & Sorensen, C. (2010). *Introduction to research in education*. Belmont, USA: Wadswoth cengage learning.

Amuche, C. I., & 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

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

Bhakta, B., Tennant, A., Horton, M., Lawson, G., & Andrich, D. (2005).Using item response theory to explore the psychometric properties of extended matching questions examination in undergraduate medical education. *BMC Medical Education.*

Boone, J. W., & Scantlebury, K. (2005). The role of Rasch analysis when conducting science educational research utilizing multiple choices test. Wiley InterScience. www.interscience.wiley.com.

Brian, F.F., Daniel, H.B., & 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.

Chilisa, B., & Preece, J. (2005). *Research methods for adult educators in Africa*. Paris ,UNESCO Pearson.

Comfrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Duong, M. (2004). *Introduction to item response theory and its applications*, research development paper. https://www.msu.edu/~dwong/StudentWorkArchive/CEP900F04-RDP/Mihn-ItemResponseTheory.htm.

Edelen, M.O., & Reeve, B.B. (2007). *Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement*. Springer Science and Business Media.

Ene, C. U. (2005) Application of Rasch model in assessing the attitude of students towards biology in senior secondary schools in Enugu Education Zone. Unpublished M.ED Project. Department of Science Education, University of Nigeria, Nsukka.

Eng, L. S., & Hoe, L. S. (n.d). 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.

Emaikwu, S. O. (2012). Issues in test item bias in public examinations in Nigeria and implications for Testing. *International Journal of Academic Research in Progressive Education and Development*, 1(1), 2226-6348.

Embretson, S.E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwal, NJ:Erlbaum.

Gruijter, D, N. & Kemp J. T. C. (2002). *Statistical test theory for education and psychology*. Boston:Kluwer.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9(2), 139-164.

Hambleton, R.K. (1989). Item response theory. Three parameter-logistic model. *Centre for Study of Evaluation*. Retrieved from http://www.cse.ucla.edu/products/reports/R220.pdf

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

Hauser, C., Kingsbury, G., & Northwest Evaluation Association, (2004). Differential item functioning and differential test functioning in the "Idaho Standards Achievement Tests" for Spring 2003. Northwest Evaluation Association.

Hunyepa, J. (2014). Khama must be petitioned over problems in education. The Botswana Gazette, News, Botswana (pty) Ltd, Gaborone, Botswana

Igwue, D.O., & Adikwu, O. (2012).Measurement of intellectual functioning of Nigerian youth: The predictive validity of JAMB/UME in relation to students' performance in University. *Educational Research*, 3(8), 639-644.

Jodin, M. G., & Gierl, M. J.(1999). Evaluating type I error and power using an effect size measure with the logistic regression procedure for DIF detection.University of Alberta.

Joshua, M. T. (2005). *Fundamentals of tests and measurement in education*. Calabar, Nigeria: University of Calabar Press.

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika*, 35, 401-415.

Kalaycioglu, D. B., & Berberoglu, G. (2011). Differential item functioning analysis of the science and mathematics items in the university entrance examinations in Turkey. *Journal of Psychoeducational Assessment, 29*(5), 467-478.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research*. Belmont, USA: Wadswoth Cengage learning.

Kobrin, J. L., Kim, Y., & Sackett, P. R. (2012). Modeling the predictive validity of SAT mathematics items using item characteristics. *Educational and Psychological Measurement, 72*(1), 99-119.

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

Lee, H. S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24* (2), 115 - 136.

Linacre, J. M., & Wright, B. D. (1994). Dichotomous mean-square infit and outfit chi-square fit statistics. *Rasch Meas Trans*, 8(2), 360.

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

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading. MA: Addison-Wesley.

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education, 37*(8), 919-931.

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.

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

Moeng, G. (2013). Issues in education, Gaborone, Botswana. Mmegi.

Mokobi, T. & 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.

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.

Motshabi, K. W., & Nenty, H. J.(2012) Analysis of ethnicity related DIF in PSLE 2008 English language: implications for fairness in cross-cultural assessment, teaching and learning. *Journal of Educational Assessment in Africa*, 7, 55-65.

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

Nenty, J. H. (In press). *Fundamentals of quantitative research education*. Book under preparation, University of Botswana, Gaborone.

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

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

Nenty, H. J. (1985). Fundamental of measurement and evaluation in education. Unpublished monograph, University of Calabar, Nigeria.

Nenty, H.J. (2000). Some factors that influence students' patterns of response to mathematics examination items. *Boleswa Educational Research Journal*, 17, 47-58.

Nenty, H. J., Odili, J. N., & Munene-Kabanya A. N. (2009). 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.

Nworgu, B. G., & Agah, J.J. (2012). Application of three-parameter logistic model in the calibration of a mathematics achievement test. *Journal of Educational Assessment in Africa*, 7, 162 - 172.

Nworgu, B. G. (2011). Differential item functioning: A critical issue in regional quality assurance. *Journal of Educational Assessment in Africa*, 6, 112-123.

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.

Ojerinde, D., Popoola, K., & Onyeneho, P. (2012). A comparison between classical test theory and item response theory: Experience from 2011 pre-test in the use of English language paper of the unified tertiary matriculation examination (UTME). *Journal of Educational Assessment in Africa*, 7, 173 - 191.

Oluseyi, A. E., & Olufemi, A. T. (2012). The analysis of multiple choice item of the test of an introductory course in chemistry in a Nigerian University. *International Journal of Learning, 18*(4), 237.

Perrone, M. (2006). Differential item functioning and item bias: Critical considerations in test fairness. *Columbia University Working Papers in TESOL & Applied Linguistics. 6* (2), 1-3.

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.

Posrie, R. (2010). Appropriate difficulty and discrimination indices of a criterion - referenced test item used in a classroom learning research and development. *International Employment Relations Review, 16*(2), 60.

Randall, J., & Engelhard, G. (2010). Using confirmatory factor analysis and Rasch measurement theory to assess measurement invariance in a high stakes reading assessment. *Applied Measurement in Education,* 23, 286-306.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.

Republic of Botswana. (1994). *Report of the revised national policy on education*. Gaborone. Botswana.

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

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.

Republic of Botswana. Ministry of Education (1994). *The revised national policy on Education -1994. Gaborone*: Government Printer.

Republic of Botswana. Ministry of Education. (2001).*Botswana Examinations Council, revised assessment procedures for junior certificate examination programme*. *Gaborone*: Government Printer.

Republic of Botswana. Ministry of Finance and Development Planning (2013). Budget speech. Gaborone: Government Printer.

Republic of Botswana. Ministry of Finance and Development Planning (2014). Budget speech. Gaborone: Government Printer.

Robin, F, Zenisky, A.L & 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.

Royal, K. D., & Puffer, J. C. (2012).Understanding the "sum of subtest to overall score discrepancy" on the maintenance of certification for family physicians examination. *J Am Board Fam Med*, 25(2), 260-261.

Royal, K. D., & Puffer, J.C. (2013).Dimensionality of the maintenance of certification for family physicians examination: Evidence of construct validity. *Am Board Fam Med*, 11(3), 286-288.

Shaibah, H. S., & Van der Vleuten, C. P. M. (2013). The validity of multiple choice practical examinations as an alternative to traditional free response examination formats in gross anatomy. *Anatomical Sciences Education, 6* (3), 49-156.

Siamisang, F. T., & Nenty, H. J. (2012). Analysis of gender-based differential item functioning (dif) in 2007 TIMSS examination among students from Botswana, Singapore and USA. *Journal of Educational Assessment in Africa,* 7, 043-054

Steinberg, L., & Thissen, D. (1995). Item response theory in personality research. In P.E. Shrout & S. T. Fiske (Eds.), Personality research, methods, and theory: a festschrift honoring Donald W. Fiske (pp. 161-181). Hilldale, NJ: Erlbaum.

Suen, H.K., & McClellan, S. (2003). Item construction principle and techniques. Chapter in N. Huang (ed). *Encyclopedia of Vocational and Technological Education*, 1, 777-798.

Thobega, M., & Masole, T. M. (2008). Predicting students performance on agricultural science examination from forecast grades, *US-China Education Review*, 5 (10), 45-52.

Ubi, I. O., Joshua, M. T., & Umoinyang, I. E. (2012) Assessment of dimensionality of mathematics tests of university matriculation examination in Nigeria: Implications for regional development. *Journal of Educational Assessment in Africa,* 7, 122- 130

Umoinyang, I.E. (2011). The challenge of removing consistent errors in achieving tests using differential item functioning (DIF) detection methods. *Journal of Educational Assessment in Africa,* 6, 120-132

UNESCO (2006). World data on Education (6thed). http://www.ibe.unesco.org/

Uruemu, O., & Adams O. (2013) Differential item functioning method as an item bias indicator. *Educational Research*, 4(4), 367-373.

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

Utlwang, A. (2003). The localization of Cambridge School Examinations as a quality assurance measure*. AEAA Conference*, Cape Town, South Africa.

Vedle, G. V., Beaton, D., Hogg-Johnston, S., Hurwitz, E & Tennant, A. (2009) Rasch analysis provides new insights into the measurement properties of the neck disability index*, American College of Rheumatogy,* 61 (4), 544 - 551.

Well, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*, 77-87.

Wen, Y. (2014). DIF analyses in multilevel data: Identification and effects on ability estimates. The University of Wisconsin Milwaukee.

William, L. K., & Mark, G.S. (2003). How well do multiple choice tests evaluate student understanding in computer programming classes? *Journal of Information Systems Education*, Vol. 14(4), 389 - 399.

Wilson, S., & Maclean, R. (2011). *Research methods and data analysis for psychology.* New York: Mc Graw Hill.

Zieky, M. (2003). DIF Primer: Center for education in assessment. *Educational Testing Service.*

Zumbo, B. D., & Gelin, M, N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated (or mediated) test and item bias. *Journal of Educational Research & Policy Studies,* 5(1)

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

BOTSWANA EXAMINATIONS COUNCIL
Botswana General Certificate of Secondary Education

AGRICULTURE **0599/01**

Paper 1 Multiple Choice **October/November 2013**

45 Minutes

Additional Materials: Multiple Choice Answer Sheet
Soft Clean Eraser
Soft Pencil (type B or HB)

READ THESE INSTRUCTIONS FIRST

Do not open this booklet until you are told to do so.
Read the instructions on the separate answer sheet very carefully.
Write your name, Centre number and candidate number in the spaces provided on the Answer Sheet.
DO NOT write your name if this has already been done for you.
Sign your name in the space provided on the Answer Sheet.

There are **forty** questions on this paper. Answer **all** questions. For each question, there are four possible answers **A**, **B**, **C** and **D**. Choose the **one** you consider correct and record your choice using a **soft pencil** on the separate answer sheet.
Each correct answer will score one mark. A mark will not be deducted for a wrong answer.
Any rough working should be done in this booklet.
Do not use staples, paper clips, highlighters, glue or correction fluid.

This document consists of **11** printed pages and **1** blank page.

[Turn over

1 Which of the following represents a correct naming of USDA Land capability classification?

- **A** 8
- **B** eight
- **C** VIII
- **D** 卌 III

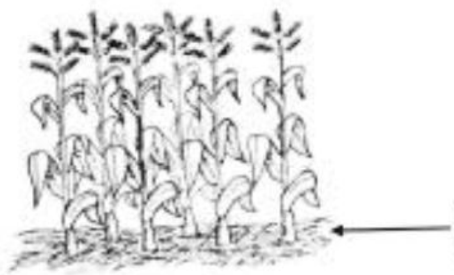
2 Which pair of the environmental factors will encourage the occurrence of pests and diseases in crops?

- **A** high humidity and high light intensity
- **B** high humidity and high temperature
- **C** heavy rains and high light intensity
- **D** light rains and high temperature

3 What percentage of land in Botswana is freehold?

- **A** 6%
- **B** 23%
- **C** 29%
- **D** 71%

The drawing shows a management practice. Use it to answer question 4.



4 How does the management practice conserve water for plant use?

- **A** by encouraging run-off
- **B** by lowering transpiration
- **C** by reducing evaporation
- **D** by speeding up infiltration

5 Which pair of statement is correct for chemical weed control?

	Advantage	Disadvantage
A	effective in weed control	low labour requirements
B	low labour requirements	pollutes the environment
C	eradicates soil microbes	low labour requirements
D	eradicates soil microbes	toxic to man and livestock

6 Which statement correctly describes a systemic pesticide?

A pests are killed when they suck the plant sap

B pests are killed when they touch the pesticide

C pests are killed when they inhale the pesticide

D pests are killed when they swallow poisoned leaves

7 Which of the following is controlled through the use of herbicides?

A birds

B insects

C rodents

D weeds

8 Which component of a pesticide will **not** kill the pest?

A active ingredient

B aerosol

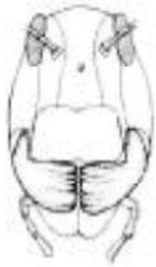
C fumigant

D inert material

9 Which set of description correctly matched the pesticides?

	Fumigants	Dusts	Aerosols	Sprays
A	dry powder	liquid under pressure	pesticide dissolve in water	releases gas
B	releases gas	dry powder	liquid under pressure	pesticide dissolve in water
C	liquid under pressure	pesticide dissolve in water	releases gas	dry powder
D	pesticide dissolve in water	releases gas	dry powder	liquid under pressure

The diagram shows a mouthpart of a pest. Use it to answer questions 10 and 11.



10 In which class of pests does it belong?

- **A** biting and chewing
- **B** boring
- **C** cutting
- **D** piercing and sucking

11 Which group of pesticides will be appropriate for controlling the pest?

- **A** contact and systemic
- **B** contact and stomach
- **C** stomach and systemic
- **D** stomach and fumigant

12 Which type of plants produce flowers when day length is less than a critical number of hours?

- **A** short day and light insensitive
- **B** long day and light insensitive
- **C** day neutral and short day
- **D** long day and day neutral

13 What is **not** a possible danger of using chemical pesticides?

- **A** beneficial organisms affected
- **B** chemicals persist in the soil
- **C** increased pesticide resistance
- **D** increased soil water retention

Use the information to answer question 14.

A 30 *l* knapsack sprayer was filled with water. A sample area of 10 m^2 was sprayed at a steady speed. 5 *l* of water remained after spraying.

14 What is the rate of spray application for this knapsack sprayer?

A 0.5 l/m^2

B 0.5 m^2/l

C 2.5 l/m^2

D 2.5 m^2/l

The photograph shows a lawn management practice. Use it to answer question 15.

15 What lawn management practice is being carried out?

A cultivation

B edging

C scarification

D spiking

The information below outlines the steps followed when using a vegetative material to plant a lawn. Use it to answer question 16.

1. Place the stolons into shallow channels
2. Temporary cover the stolons with plastic
3. Firm the soil
4. Cut stolons with at least two nodes

16 What is the correct order of carrying out the steps?

A	1	2	3	4
B	1	3	4	2
C	4	2	1	3
D	4	1	2	3

17 What is the advantage of rotational grazing?

A it allows recovery of pasture plants

B it allows quality pasture to be reserved for young stock

C animals use less energy to find pastures

D it encourages selective grazing of pastures

18 Which of the following will cause pasture deterioration?

A controlling pasture weeds

B debushing the veld

C indiscriminate burning of pastures

D reseeding the pasture

Use the information to answer question 19.

A farmer notices that two chickens in the poultry house show some signs of disease

19 What immediate action should the farmer take to avoid further spread of the disease?

A call the veterinary officer

B give the chickens some vaccines

C isolate the diseased chickens

D kill the whole flock

20 What is the main reason for having a rough floor in livestock houses?

A to ease cleaning

B to avoid parasites

C to prevent slipping

D to improve durability

Use the information to answer question 21.

Mr Kitso keeps chickens for sale. He kept the following information: number of chickens dead, number of bags of feeds used, amount of money spent on labour, amount of water used per month.

21 Which information can form part of his financial record?

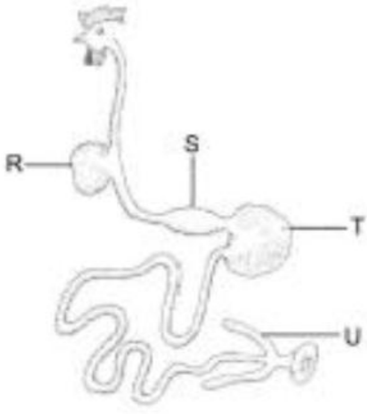
A number of chickens dead

B number of bags of feeds used

C amount of water used per month

D amount of money used on labour

The diagram shows the digestive system of a chicken.
Use it to answer questions 22 and 23.



22 In which part is grit found?

A R

B S

C T

D U

23 What is the function of part **U**?

A digest cellulose

B digest fats

C digest proteins

D digest starch

24 Which fodder is preserved under air-tight conditions?

- **A** hay
- **B** herbage
- **C** forage
- **D** silage

Use the information about KK investments to answer questions 25 and 26.

<u>KK investments</u>

Total area of farm = 50 hectares

Livestock kept =	cows	10
	donkeys	20
	bulls	2
	goats	6
	sucking calves	5

0.75×6

Given that: Livestock units equivalent

1 cow / donkey	=	1 LSU
1 bull	=	1.5 LSU
1 goat	=	0.75 LSU
1 sucking calf	=	0.0 LSU

10+20+3+4.5

25 What is the stocking rate of the farm?

- **A** 0.75 LSU/ha
- **B** 0.86 LSU/ha
- **C** 1.16 LSU/ha
- **D** 1.33 LSU/ha

50 hectares = 37.5
1 hec = x
x = 0.75

26 Given that the carrying capacity of the above farm is 2.5 LSU/ha, what can be concluded ab the stocking condition of the farm?

- **A** correctly stocked
- **B** fairly stocked
- **C** overstocked
- **D** understocked

27 Which of the following is an environmental factor that influences variation among organisms the same species?

- **A** crossing over
- **B** meiosis
- **C** mutation
- **D** nutrition

Use the information to answer questions 28 and 29.

In cattle, a lack of horns (polled) is dominant to the presence of horns. A heterozygous polled bull is crossed with a homozygous recessive cow.

28 What will be the genotype ratio of the cross?

A 1:1

B 2:1

C 3:1

D 4:0

29 If 50 calves are born from this cross, how many of them can be expected to be polled?

A 25 calves

B 33 calves

C 38 calves

D 50 calves

30 Which term best describes the amount of a product offered for sale at a given price within a given period of time?

A demand

B marketing

C selling

D supply

31 Which of the following is a feature of planned economy?

A decisions made are based on profit

B government allocate resources for production

C prices are controlled by market forces

D there is a lot of wastage of resources

32 Why is it important to carry out a market research for a farm product?

A to promote sales of a product

B to determine demand for a product

C to create awareness about a product

D to provide information about a product

33 Which **one** is an example of a facilitating function of marketing?

A financing

B selling

C storage

D transportation

Use the information to answer questions 34 and 35.

A farmer raised 500 broilers. At 8 weeks he sold them at P35.00 each. He paid rent of P600.00, bought feeds for P2 500.00 and hired casual labour for P1 200.00.

34 If these were the only expenses incurred, what is the gross margin of the farm?

A P26.40

B P27.60

C P30.00

D P35.00

600 + 2500 + 1200

− 15 100

4300

17 500

17 500

4 300

35 Given that the total fixed costs of the farm were P6 000.00. What profit will be made by the farmer?

A P7 200

B P7 800

C P9 000

D P11 500

60 000

− 4 300

36 Which material is appropriate for making a cattle crush?

A brick

B hedge

C wire

D wood

37 An earth road has potholes and over grown road side bushes.
Which pair of maintenance activities must be carried out on this road?

A compacting and clearing

B gravelling and clearing

C gravelling and paving

D cleaning and paving

38 Why is it recommended that the wall of an earth dam be made of clay soil?

- **A** to increase storage capacity
- **B** to reduce evaporation
- **C** to reduce seepage
- **D** to purify water

39 What are the positions of the valves during the power stroke of a diesel engine?

	inlet valve	outlet valve
A	closed	closed
B	closed	open
C	open	closed
D	open	open

40 Why is it important to lubricate movable parts of an engine?

- **A** to increase heat and reduce friction
- **B** to increase friction and reduce noise made by parts
- **C** to reduce friction and prevent over heating
- **D** to reduce speed of parts and increase noise made by parts