UNIVERSITY OF CAPE COAST

ANALYSING THE PSYCHOMETRIC PROPERTIES OF THE BOTSWANA JUNIOR CERTIFICATE EXAMINATION AGRICULTURE MULTIPLE-CHOICE ITEMS USING ITEM RESPONSE THEORY

TAPELA BULALA

2023

UNIVERSITY OF CAPE COAST

ANALYSING THE PSYCHOMETRIC PROPERTIES OF THE BOTSWANA JUNIOR CERTIFICATE EXAMINATION AGRICULTURE MULTIPLE-CHOICE ITEMS USING ITEM RESPONSE THEORY

BY

TAPELA BULALA

Thesis submitted to the Department of Education and Psychology of the Faculty of Educational Foundations, College of Education Studies, University of Cape Coast, in partial fulfilment of the requirements for the award of Doctor of Philosophy degree in Measurement and Evaluation.

AUGUST 2023

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original research and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature: ……………………….. Date: ……………………….

Name:……………………………………………………………………..

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature: ……………………… Date: 10/5/24

Name: PROF. ERIC ANANE

Co-Supervisor's Signature: ………………………. Date: ……..……………

Name:……………………………………………………………………...

ABSTRACT

The main purpose of this study was to analyse the psychometric properties of the Junior Certificate Examination (JCE) agriculture multiple-choice items using item response theory. In addition, the study examined Differential Item Functioning (DIF) in relation to sex, location, and school type. The study adopted the descriptive research design. Utilising the census approach, 123,218 candidates' responses with 4,928,720 cases of data points from the 2018-2020 examination period was used for the study. Secondary data on candidates' responses to each item were obtained from Botswana Examination Council (BEC). The 3-parameter logistic model of IRT was utilised for purposes of data analysis. Raju's procedure was employed for the DIF determination using *DifR* package within the R-studio environment. The study revealed that items had varied levels of difficulty, 15% to 20% of the items exhibited poor discrimination, and some items had high guessing indices. It was also revealed that agriculture science items had both uniform and nonuniform DIF across demographic variables. The study concluded that the contribution of items to student achievement in agriculture science was good with few items having validity and reliability concerns. It further concluded there exists DIF across all three demographic variables of gender, location, and school type. Resource disparities and sex-based cultural practices contribute to most DIF items. It was recommended that BEC should consider validating their examination items using IR.

KEYWORDS

Agriculture Science

Core subjects

Examinees

Invariance

Item quality

Local independence

Logistic Regression (LR)

Raju's area measurement

Reliability

Unidimensional

Validity

ACKNOWLEDGEMENTS

Firstly, I wish to express my profound gratitude to all those who have contributed immensely to the success of this study. I'm grateful to the All-Powerful God for giving me the grace that kept me going through this study. Prof. Eric Anane and Dr. Andrews Cobbinah, who supervised the project, have been a great help and guide. I'm very thankful to Mr. Francis Ankomah for his helpful advice, corrections, and helpful criticisms, all of which made it possible for me to write this thesis well.

I appreciate the scholarship foundation of RUFORUM. Finally, my appreciation goes to Chawangwa Mudongo who helped to make this project a success.

DEDICATION

To my wife, Mrs Tapela; my children- Sarona, Mmilili, Junior, and Thabiso; my late mother, Ms. Tjatiwila Bulala; and my late sister, Gezephi Bulala.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

BEC	Botswana Examination Council
CTT	Classical Test Theory
DIF	Differential Item Functioning
ICC	Item Characteristic Curve
IIF	Item Information Function
IRB	Institutional Review Boar
IRT	Item Response Theory
RNPE	Revised National Policy on Education
UCC	University of Cape Coast

CHAPTER ONE

INTRODUCTION

The basic education level is the foundation on which all future academic advancements are built. It is for this reason that appropriate teaching pedagogy and assessment instruments must be used to accurately determine the true ability of a student. Testing is one critical component through which a society adjudicates the effectiveness and quality of an educational system (Medina, Poveda, & Ortz, 2019). Thus, in basic education, national high-stakes examinations are at the centre of measuring the quality of education. Scores emanating from these national public examinations form the basis for decision-making such as placement, repeating a class, progression, and even determining a career path (Osei-Mensah, 2012).

It is against this backdrop that countries and, at times, regions have set up examination bodies such as the Botswana Examination Council (BEC) and the West African Examinations Council (WAEC), respectively. BEC was established by an act of parliament in 2002. The BEC's mandate is to provide a credible and responsive assessment and examination system. To achieve its mandate, the Council must develop and deliver quality examinations that accurately and equitably measure learners' abilities. In this study, the researcher sought to examine the psychometric properties of the agriculture junior certificate examination and to understand the sources of performance disparities among examinees based on their demographic features.

Background to the Study

Education is thought to be the key to national development, which cannot be achieved without a good investment in people through teaching and learning. As a result, quality education has become increasingly important in today's competitive world. However, according to UNESCO (2004), the quality of education is declining worldwide. Of great worry, the observed decline goes against the Sustainable Development Goal 4 (SDG 4), which aims to give everyone a quality education that is fair and open to all. A similar decline in academic performance has been reported in Botswana (Mphale & Mhlauli, 2014; Bulala & Malema, 2019; Bulala & Mbisana, 2019). In particular, the researchers reported that students' academic performance has shown some decline in many school subjects, while in agriculture, the decline has become consistent at all levels of education (Hulela & Miller, 2003; Sibanda, Hulela, & Tselaesele, 2016; Makwinja, 2017; Oitsile & Oats, 2020).

In Botswana's education system, students sit for three public examinations: Primary School Leaving Examinations (PSLE) at the end of seven years of primary education, Junior Certificate Examination (JCE) at the end of three years of junior secondary school, and Botswana General Certificate of Education (BGCSE) at the end of two years of senior secondary school level. The role of public examinations in Botswana's education system is to assess student academic achievement at each level of study. These kinds of tests play a big role in figuring out a student's academic path and progress. PSLE and JCE examinations aim to select students for junior and senior secondary respectively. Notwithstanding, PSLE is rather a diagnostic assessment as all examinees progress to junior secondary regardless of their academic performance. According to the Revised National Policy on Education, RNPE (1994), the mass progression from primary school to JCE is meant to achieve 10 years of basic education. However, for progression from junior secondary to senior secondary students must obtain a pass credit in the core subjects.

At the junior secondary school level, which is the focus of this study there are six core subjects which are English, Mathematics, Integrated Science, Setswana, Social Studies, and Agriculture Science. Implied here is that at JCE, Agriculture Science is part of a high-stake examination. Thus, students' progression and choice of subjects in senior secondary school are based on how learners performed in junior high school thereby curving an examinee's future career path (Osei-Mensah, 2012).

The plan for testing is made for students who have finished a three-year course at junior secondary school. It is meant to test all ability levels of achievement. According to the BEC's 2013 assessment syllabus, the agricultural examination has three required parts that are each tested as a separate paper. Paper 1 has 40 multiple-choice questions, and each one has four possible options. It accounts for 35% of the total score. Paper 2 is a short-answer type question with a total of 60 marks and contributes 40% to the total overall score.

The third assessment component, Paper 3 is a centre-based assessment carried out by external examiners from BEC and it contributes 25% to the overall score. It is a culmination of students' practical work records from the first year to the third year. Students are expected to present two crops at the maturity stage during the external moderation time. The objective of this school-based practical assessment is to provide candidates with a greater opportunity to demonstrate their practical farming skills and techniques, including the application of knowledge and concepts. Paper 4 is a replacement for centre-based assessment (Paper 3) and carries the same weightage. It is intended for centres and candidates who are not able to complete expected school-based practical coursework activities. However, according to Thobega and Masole (2008) the mean mark for agriculture school-based assessment (paper 3) scores are biased towards the maximum attainable mark, whilst the mean mark for theoretical examinations was near the median.

Interestingly, among all six core subjects, Agriculture Science is the only practical subject. According to (RNPE, 1994), the rationale is to develop a human resource base capable of reviving agricultural production and stimulating economic growth in the country. This was based on the observed decline of agriculture's contribution to the Gross Domestic Product (GDP) from 40% at independence in 1966 to below 3.8% in 1994. It further posits that agriculture education play a very important role in the country's development as over 50% of Batswana live in rural areas, where agriculture plays a major economic role (Botswana Statistics , 2023).

In addition, Hulela, Mukuni, Abreh, Kasozi, and Kraybill (2021), posit that agriculture plays an important role in developing countries like Botswana and as such must be included in the school's curriculum. Rammolai (2009) shared the same view, saying that in the curriculum, agriculture is seen as a prerequisite for economic development; therefore, it's teaching at the secondary school level is to create the ability to farm and stimulate the growth of the economy. Therefore, agriculture at secondary school is considered pre-vocational to facilitate the orientation of students to the world of work.

Okoye and Udoudo (2015) define vocationalisation as a move away from traditional academic subjects and toward more practical ones like agriculture, business, design and technology, and home management. Besides combining practical and academic subjects, learning can also come from self-reliance activities and clubs and societies (Wrenn & Wrenn, 2009). In some cases, the whole school is seen as a productive business that gives students a wide range of administrative and technical production experiences to help them work better (Lu & Xu, 2022).

Several initiatives were launched by the Ministry of Education and Skills Development (MOESD) of Botswana in recognition of the vital role that agriculture education plays in equipping students with life skills. Among these are making agriculture science the only core practical subject at the junior secondary level, increasing the number of agriculture teachers to be trained, introducing agriculture science at the primary level in the year 2005, and including agriculture science in multiple pathways initiative. Multiple pathways were introduced in 2020 to provide learners with a transition from an academic to a vocational stream, where they can concentrate primarily on agriculturally based practical training at selected senior secondary schools. The introduction of agriculture science in primary school curricula in 2005 was intended to equip primary school students with agricultural knowledge and facilitate their transition into the junior secondary agriculture curriculum (Kgoboge, 2020).

In addition, despite government initiatives such as high public expenditures and free access to basic education, failure rates are still high (Suping, 2022). In fact, according to Statistics Botswana (2018), low pass rates and transition rates of below 50% from Junior Secondary School to Senior Secondary are still recorded. This shows that, overall, JCE has not done a good job of getting many students ready for Senior Secondary School and vocational schools, as it was meant to do. The Botswana Examination Council (BEC) examinations continues to record poor results in agriculture science. Table 1 shows students' performance in agriculture examinations at PSLE, JCE, and BGCSE respectively.

Table 1- *Students' Performance in Agriculture Multiple-choice Examinations at PSLE, JCE and BGCSE 2018 to 2020*

	Performance of Students in Agriculture		
Year	PSLE	JCE	BGCSE
2018	55.65	42.1	38.25
2019	61.04	46.05	29.34
2020	63.77	45.15	30.19
Average	60.15	44.43	32.59

Source: Botswana Examination Council (BEC) 2018, 2019, 2020

From Table 1, PSLE had the highest average pass rate of 60.15%. This was followed by JCE where 44.43% of candidates passed, and finally by BGCSE where 32.59% of students passed. The largest decline in Agriculture averaging a 15.72% pass rate is between PSLE and JCE, while between JCE and BGCSE is 11.84%.

Table 2 shows students' performance by sex and school type (public and private). The performance as shown in Table 2, the average pass rate for the years 2018, 2019, and 2020 is at 44.43%. For the three years under review, females have consistently performed better than their male counterparts on the credit pass (A-C %) with averages of 48.83 and 39.94 percent respectively. The results show that examinees from public schools recorded the highest proportion in grades C or better (36.97%) while private schools contributed a paltry (7.46) in agriculture science.

Table 2- *JCE Agriculture Multiple choice Percentages Pass Rate (A-C %) by Sex, Public and Private from 2018 to 2020*

Year	Total A-C %	Male	Female	Public Schools	Private Schools
2018	42.1	38.49	45.69	37.67	4.43
2019	46.05	42.67	49.33	37.51	8.54
2020	45.15	38.66	51.47	35.74	9.41
Average	**44.43**	**39.94**	**48.83**	**36.97**	**7.46**

Source: Botswana Examination Council (BEC) 2018, 2019, 2020

Figure 1 shows student performance by education administrative regions at JCE.

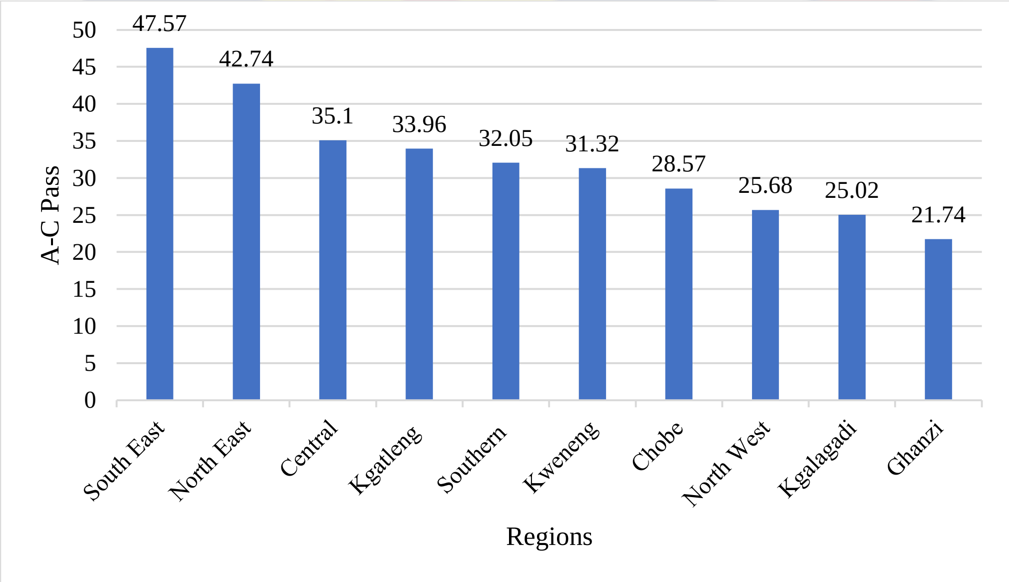


Figure 1- Performance by regions for candidates awarded grade C or better 2018-2020

From *Figure* 1, the South East and North East region which predominantly covers urban schools had the highest pass rate. Notably, Kgalagadi and Ghanzi which are two of the remotest rural areas recorded the lowest overall performances for the three years under review with 22.96 and 21.74 respectively.

Undoubtedly, parents, teachers, policymakers, test developers, and the government will all be concerned about the performance trend over the years. In addition, the glaring disparities between students from different demographic backgrounds raises issues about equity in assessment. The various factors, including teacher competencies, parental involvement, the availability of teaching resources, and the nature or quality of items, among others, can influence the performance of examinees in general examinations (Ndlovu, 2018; Amie-Ogan & Friday, 2020; Naite, 2021). However, Ashikhia (2010) asserts that the quality and nature of the test items and the examinees' characteristics are among the most influential of these factors in student academic performance. This assertion is supported by Ayanwale, Adeleke, and Mamadelo, (2018), who further state that the quality of assessment at all levels of education is determined by the quality of measuring instrument used. Thus, emphasises the significance of analysing items to determine their properties. Therefore, characteristics of the examination item leading to the observed performance of students on the agriculture science examination must be examined.

The JCE is high-stake testing that determines progression of students to senior secondary schools. The items used lead to the observed scores. The scores represent students' ability therefore Botswana Examination Council must determine psychometric properties of the items used to assess examinees. This helps to ensure that student's ability and skills are accurately measured. It also ensures accountability to all stakeholders in education system. To this end Tommy and Udo (2019), posit that item developers are expected to create high-quality items that can be used to assess the abilities of students in both homogeneous and heterogeneous settings, as the value of such a measure would be contingent upon its quality. The items by BEC must accurately reflect the ability and achievement of the students.

If BEC uses test items that are consistently of poor quality, they may introduce sources of error leading to wrong decision about student learning and growth. Quansah, Amoako and Ankomah, (2019) opine that often test developers such as teachers, (who in Botswana are members of the item development team) have limited skills in the construction of examinations leading to questionable reliability, and fairness of the assessment tasks.

As Nenty (2004) said, an important sign of an instrument's psychometric quality is how valid, fair, and reliable it is, and this problem can be solved by using modern psychometric methods. In addition, the assessment instrument must be well-calibrated and must pose tasks that correctly capture the proficiency of examinees (Mislevy, Behrens, Dicerbo, & Levy, 2012). The tasks should serve as hurdles for students to go through and exhibit their talents in the testing and examination environment. These assessment tasks in educational measurement are known as items and possess statistical indicators (psychometric properties) that define the quality of an item in the instrument used (Rust & Golombok, 2014). Thus, psychometric properties can be defined as characteristics of test items that estimate a trait or a construct of interest.

Therefore, test developers must evaluate the item quality to ensure that the instrument accurately measures and ranks examinees' knowledge and abilities. "Item analysis is essential to test design because it evaluates examinees' responses to individual test items to determine the quality of those items and the overall quality of the test" (Pande, Parate, Nikam, & Agrekar, 2013, p. 46). In addition, it provides a more accurate representation of the characteristics of test items and guides item review, selection, and banking (Obon & Rey, 2019). Thus, it examines the performance of an item concerning other factors to better comprehend its characteristics and efficacy.

Furthermore, the manual for educational and psychological testing referred to as *standards* specifies that "to ensure proper accountability, there is a need to conduct periodic checks of the stability of test items on which scores are reported" (American Educational Research Association-AERA; American Psychological Association-APA & National Council on Measurement in Education-NCME, 2014). Thus, the suitability of an instrument is established through the stability of psychometric indices (item difficulty, item discrimination, item bias, and guessing parameter) over time. Aborisade and Fajobi (2019) defined psychometric aspects of examinations as qualities present in tests that are used to evaluate examinees.

The indicators of quality test items are echoed by Rust and Golombok, (2014), who state that within psychometrics, there are four fundamental principles whereby the quality of assessment is judged, namely reliability, validity, standardization, and freedom from bias. In support of using items with known and stable psychometric properties, Anene and Ndubi (2003) also posit that all items making high-stake assessment examination must satisfy the criteria of reliability, validity, usability as well as fairness to maintain acceptable standards. In sum, a test should be valid, reliable, fair, and devoid of item bias for it to be deemed to of acceptable standards.

Reliability is defined as the consistency, or accuracy of the measurement tool (Abonyi, 2011). This means that if the agriculture science items are repeated over time, they must produce slightly different or similar responses at different times due to measurement error. Therefore, the smaller the error for a test, the more reliable it is. The larger the error, the lower the reliability. In general, if a test has good psychometric properties, it can effectively measure student learning. The score on a test can be viewed as a combination of the correct score and the error rate. The standard error of measurement gives the standard deviation of the measurements of the same person (APA, 2014). Notwithstanding, BEC reports on the composite validity of their tests, and it cannot be said individual items contributed equally to the observed reliability coefficient (Moyo, 2017).

Validity refers to how well the assessment tool measures the underlying outcome of interest. However, the latest edition of the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) states "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (p. 11). What is evident from the view of the definition is that validity is not a property of the tool itself, but rather of the interpretation or specific purpose of the assessment tool within settings and examinees' context (Sullivan, 2011). Applied to the context of this study, data on agriculture science items is highly valid when several pieces of evidence can be provided regarding students' response patterns and how the data were used. Thus, an item with high validity will measure accurately the construct it is supposed to measure. However, the BEC do not report on validity nor dimensionality of their tests (Moyo, 2015).

The third facet of item quality which is perhaps very critical in practical subjects like agriculture science is equitable treatment of all examinees during the testing process (fairness or freedom from bias). According to Baharloo (2013), fairness is the essence of good assessment; fair treatment of all test takers without negative evaluation in the assessment process, fair access to the standards of the scale and suitable for use in interpretation of test scores for intended purposes According to the RNPE (1994), all learners regardless of their socio-economic status must have equal opportunities to quality basic education. Since student performance in achievement tests is a measure of the quality of education, it is important to establish behaviour by examinees from different groups (location, gender, or school type).

According, Alordiah, (2015), because national examinations are taken by examinees from various demographic backgrounds across a country, they could be harder for a certain group of test-takers. while very easy for other groups. However, if something in an item causes examinees of the same skill level but from different groups to act differently, the item exhibits differential item functioning (DIF). Thus, DIF may can be an indicator of bias in testing. When measuring, 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). Agricultural science as a core subject is done by all students countrywide. These students are from different setting and likely to have varied backgrounds. Furthermore, if DIF is detected, subsequent analyses of item bias (e.g., content analysis, empirical, evaluation) is needed to identify sources of and provide evidence to declare item bias (Liao & Yao, 2021; Özkan, & Güvendir, 2021).

Thus, when used correctly, assessments provide an objective process and data that can reduce the effects of subjective or implicit bias, or more direct intentional discrimination thus becoming fair to all examinees. When tests aren't fair, different groups of test-takers get different scores every time indicating bias. So, test bias is a psychometric property of test scores; it is the numbers that back up claims that tests are unfair (Furr & Bacharach, 2008). Xi (2010), defines fairness "as comparable validity for all the identifiable and relevant groups across all stages of assessment, from assessment conceptualization to the use of assessment results" (p. 154).

According to Enunwah, Akwa, and Okon, (2014), researchers have advanced several variables that introduce bias and affect students' performance at secondary school level. These variables include gender, location, teaching methods/philosophy, school ownership, and other factors.

In the context of this study, JCE is written by examinees from towns and rural areas. The contribution of different demographic spread particularly in relation to assessment of agriculture science has not been fully interrogated. Interestingly results over the years show that examinees from urban-based administrative districts outperform those from rural areas (as can be seen in figure 1). Teaching and learning of Agriculture science unlike the other five core subjects is practical-oriented. The delivery and comprehension of content require hands-on activities (Mudekunye et.al, 2020). The practical activities are usually performed at the school farm or community farms around the school.

Several studies have reported performance disparities between rural and urban based examinees (Moyo, 2015; Ihechu, & Madu, 2016; Amao, et al., 2016**;** Annan-Brew, 2020). In Botswana, some schools particularly those in towns lack sufficient land for school farms which are required to reinforce students' understanding and acquisition of skills agriculture science. It is the view of the researcher that resource disparities may lead to inequality in assessment. This is borne out of the fact that some examination items may be maybe in pictorial form depicting crop and animal management husbandry practices. These types of items will most likely be better understood by examinees who participated in hands-on practical sessions. The other observation by the researcher is that due to large classes for a practical subject like agriculture some practical sessions are limited to a demonstration by teachers with less or no direct practical involvement by students. In fact, Weeks, (2002) posit that one of the major contributing factors to poor student performance in vocational subjects like agriculture is the lack of resources.

Agriculture science as a core subject at junior secondary school is done by both female and male students. However, literature by different scholars is inconclusive regarding gender and academic performance in agriculture science. Studies (Bulala & Malema, 2019; Otemuyiwa, 2017) reported no significant difference while (Amao, Adewuyi, Gbadamosi, Salami, & Ogunjinmi, 2016), reported that males perform better than females in agricultural science. The observed difference in academic performance must be investigated based on how each of the items function between female and male students. Furthermore, Otemuyiwa, (2017) posit that some subjects or disciplines such as engineering, arts, crafts and agricultural sciences considered as male domain whereas others like home economics, typing and nursing are seen as female domain. This is echoed by Schuster and Martiny, (2017) who posit that there are cultural and social context stereotypes that Science, Technology, Engineering, Mathematics are best performed by male students.

According to the assessment syllabus (Republic of Botswana, 2013), agriculture is an applied science subject based on animal husbandry (ruminants and non-ruminants), crop husbandry (field crops and horticulture), agricultural economics, agricultural engineering, and soil science. However, Ayeni (2022) points out that females are not given prominence in science-related activities. Unfortunately, for agriculture beside being science-based it is also culturally seen as masculine-inclined. This highlights the challenges female examinees may face in dealing with Agriculture science test items.

Furthermore, according to Food Agricultural Organisation [FAO], (2018), the cultural norm in Botswana is for men to engage in the more physically demanding agricultural tasks. These include animal care, land preparation, and cultivation. In contrast to males, females oversee tasks such as planting, weeding, crop management, horticulture, and poultry. These arrangements socialise boys and girls differently and expose them to agriculture in different ways, which may impact their academic performance. However, Animasahum (as cited in Otemuyiwa, 2017) argues that male and female students can learn and perform comparably well if exposed to comparable learning activities.

The Botswana school ownership pattern is such that some schools are owned by the government while others are privately owned. In a study by the UNESCO International Bureau of Education [IBE-UNESCO], (2010) on world education data, it was reported that in Botswana, students from private schools perform better than their counterparts in the public schools on average. Anigbo, (2006) reported that there was appreciable difference in academic performance in favour of private schools. Notwithstanding, Enunwah, Akwa and Okon, (2014) posit that students from public schools performed better than their counterparts in private schools due the fact that they are well resourced with better facilities.

However, in Botswana, the researcher has observed that though most private schools are generally well resourced, agriculture education seems to lack critical resources such as school farm. This observation is supported by the fact that most of the private schools are in towns where land availability is a challenge. In fact, with the option of school based practical assessment or written practical examination, most private schools have opted for the latter. By opting for written practical examination over direct practical assessment the schools cite lack of land for agriculture farm. The lack of resources for Agriculture science may result in students denied critical practical lessons which are needed to reinforce learning. The performance of examinees from private school in agriculture science is very poor when compared to public schools (as can be seen on Table 2). The observed disparities in performance between examinees from different demographic background could be pointing to items functioning differently among them.

However, Bundsgaard, (2019) posit that the presence of DIF poses a serious threat to test validity as it creates unequal opportunities among examinees taking the same test and renders education less inclusive. *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 2014) points out that DIF is a validity evidence based on internal structure of the instrument. Therefore, DIF presence threatens the internal structure as response to item is governed by something other than the construct that the instrument is intended to measure. DIF detection assists the test developer to be aware of situations where examinees of the same ability but from different subgroups have different probabilities of success on an item.

In view of the dangers posed by DIF to test validity, several techniques, or methods to detect it have been developed (Zumbo, 2003). Zumbo further posited that DIF detection techniques can be classified into either, Item Response Theory (IRT) and Classical Test Theory (CTT). However, (Bichi & Talib, 2018) argues that due to limitations of CTT like sample dependency of coefficient measures and poor estimates of measurement error, psychometricians should rather adopt IRT. Implied here is that under CTT parameters are test-dependent and examinee dependent. Consequently, group mean differences are based on total test scores across different examinee subgroups. This makes the parameters estimates to be unstable when taken from different samples. In view of the heightened CTT limitations, many researchers (Nenty, 2004; Nworgu, 2011; Umoinyang, 2011;

Anan-Brew, 2020) have suggested the adoption of IRT to detect DIF and ensure fairness in testing.

From the readings of (Lord, 1980; Rogers & Swaminathan, 1993; Raju, 1988; Cohen, Kim, & Wollack, 1996), there are several IRT-based methods that can determine the overall quality of test items particularly the detection of DIF. These methods include item characteristic curve (ICC), Item Response Theory likelihood ratio (IRT-LR), IRT-two parameter model, and IRT-three parameter model among others. However, IRT-based approaches are premised on some key assumptions. These assumptions are; unidimensionality, measurement invariance, and local independence.

Despite the availability of various IRT methods, there is consensus that multiple-choice items are generally prone to guessing, making 3-PL the most suitable method for such items (DeMars, 2008; Amad & Mokshein, 2016; Falani, Nisraeni, & Irdiyansyah, 2017). They posit that multiple-choice items are scored binary by treating one option choice as correct, while distractors are treated as equally wrong. From the explanation above, 3PL superiority is based on its ability to match the examinee's ability to the probability of getting an item right. A correct response that is accompanied by a mismatch between the two estimates is attributed to guessing. Thus, if an item misfit its diagnosed, it's because the item is not very good. From this point of view the current study on agriculture science multiple-choice items chose 3PL for determine the quality of the tests.

Statement of the Problem

The consistently low pass rates in Agriculture Science in the BEC examination are a big problem, and they make it harder to reach SDG 4's goal of providing a quality education to everyone. High-stake items that are faulty and of poor-quality lead to increased probability of poor discrimination, and implausible distractors (Cole, Bergin, & Summers, 2018; Rush, Rankin, & White, 2016). This results in inaccurate reflection of students' true ability. BEC, (2020) reported that candidates who obtained credit passes in agriculture during the years 2018, 2019, and 2020 examinations were 42.10%. 46.05% and 45.15% respectively. Agriculture Science multiple-choice items make up 35% of the total score, and their role in the drop needs to be investigated.

Although BEC evaluate the quality of its examination through CTT, it does not account for freedom from bias (Adedoyin, 2010). Downing (2004) and Ugodulunwa (2014) suggest that for there to be quality assurance in assessment, there should be a shift of emphasis from CTT to IRT as it provides solution to group dependency which is inherent with CTT. Additionally, IRT assesses reliability of individual items through Item Information Function curves which is a CTT limitation. The lack of analysis regarding test dimensionality is a critical issue related to the BEC reports. Moyo (2017) on a study about dimensionality of BEC examination reported that though BEC treat their examinations as unidimensional, they were found to be multidimensional. However, treating a multidimensional test as unidimensional can introduce uncontrolled bias into the evaluation of examinees (Walker & Beretvas, 2003; Moyo, 2015). Therefore, there is a need to examine the factor structure of JCE agriculture science multiple choice items to provide basic evidence of validity and reliability.

From 2018 to 2020, there were also differences in how students performed on tests based on their gender, where they lived, and who owned the school. This could mean that different items worked in different ways for different subgroups. Even so, DIF is a serious threat to the validity of a test, and it can lead to decisions that could hurt the educational outcomes of test takers (Latterell & Regal, 2003; Adedoyin & Mokobi, 2013). Previous studies indicate that when it comes to on multiple-choice items men do better than women as compared to constructed response (Robinson & Lubienski, 2011; Sohn, 2012; Reardon, Kalogrides, Fahle, Podolsky, & Zárate, 2018)). The United Nations Development Programme [UNDP], (2016) ascertained the existence socio-cultural stereotypes in gender roles. They reported that females in Botswana mostly deal crop farming while males predominantly manage livestock. This sex-based roles in farming activities gives different exposure and is likely to be carried over into student teaching and learning .

In many sub-Saharan African nations, the learning divide between urban and rural communities is an ongoing issue (Sumida & Kawata, 2021). According to earlier research, there is a learning gap between urban and rural areas because of the differences in student characteristics and educational resources between the two (Opoku-Asare, & Siaw, 201; Rodrigues, Costa, Silva, Mariano, & Jesus Filho, 2021). In Botswana, rural schools are usually the least ranked in academic performance when JCE results are released. Based on the literature on agriculture science the issue of how items function between the rural and urban examinees has not been explored. In the researcher's view performance gap between urban and rural schools is worrying and raise several questions fairness in assessment.

Several studies have reported disparities in academic performance between privately owned schools and public schools. In fact, the studies attributed the higher performance in private schools to low student-teacher ratio and better facilities. Notwithstanding, the performance of private schools in agriculture science in Botswana shows a different trend as private schools perform poorly compared to public schools. The researcher has encountered a situation where several private owned schools lack agriculture farms to carry out practical training of students. Thus students in private schools end up not acquiring necessary skills which their public schools counterparts get from practical lessons.

In addition, the agriculture science results cannot be said to satisfy the government's efforts to reverse the declining contribution of the agriculture sector to Gross Domestic Product. What is known extensively is that students, teachers, the environment, parents, and school-related factors are partially accountable for the decline in student performance (Oitsile & Oats, 2020; Sibanda, Hulela & Tselaesele, 2016; Baliyan & Nenty, 2015). Nevertheless, the studies on contribution of the test quality (reliability, validity, and fairness) to the observed decline in agriculture science in Botswana were not found. When the reliability of scores as accurate measures of student achievement is in question, these scores cannot be used to make future educational decisions (Solano-Flores, 2013). Given the significance of this examination's applications, the test items' quality must be evaluated. This observation amplifies the need to assess the quality of items to support and give credence to decisions emanating from student scores in Agriculture science examination.

Few IRT-based studies in Botswana have focused on multiple choice items for English, Mathematics, and Science (Adedoyin, 2010; Motshabi & Nenty, 2012; Siamisang & Nenty, 2012; Adedoyin & Mokobi, 2013). Unfortunately, these studies focused only on one aspect of test quality which is fairness. However, not all items which are DIF-free can be considered valid and reliable. Studies must assess most if not all measures of examination quality (Hagquist 2019; Liu & Bradley, 2021).

Thobega and Masole, (as cited in Moyo, 2017) have reported that BEC only review structured items, leaving other assessment components. This assertion is of great worry, considering that multiple choice items in agriculture science contribute 35% to overall student score. This scenario gives rise to irresistible inference that the observed poor performance may be emanating from multiple choice items.

Purpose of the Study

The overarching purpose of this study was to analyse the psychometric properties of the JCE agriculture multiple-choice items using item response theory.

Research Objectives

Specifically, the study determined the:

1. characteristics of the JCE agriculture multiple choice items organised by BEC based on the:
   a. difficulty parameter,
   b. discrimination parameter, and
   c. level of guessing

2. contribution of the JCE agriculture multiple choice items organised by BEC to the measure of the students' achievement ability in the subject.
3. differential item functioning (DIF) of the JCE agriculture multiple choice items administered by BEC in terms of sex.
4. differential item functioning of the JCE agriculture multiple choice items administered by BEC with reference to school location.
5. differential item functioning of the JCE agriculture multiple choice items administered by BEC with reference to private and public schools.

Research Questions

The study was guided by the following research questions:

1. What are the characteristics of the JCE agriculture multiple choice items organised by BEC based on the:
    a. difficulty parameter,
    b. discrimination parameter, and
    c. level of guessing?
2. What is the contribution of the JCE agriculture multiple choice items organised by BEC to the measure of the student's achievement ability in the subject?
3. What is the level of differential item functioning (DIF) of the JCE agriculture multiple choice items administrated by BEC, in terms of sex?
4. What is the level of DIF of the JCE agriculture multiple choice items administered by BEC with reference to school location?

5. What is the level of DIF of the JCE agriculture multiple choice items administered by BEC with reference to school type (public and private)?

Significance of the Study

The findings of this study contribute to the literature in the field of assessment and quality assurance at basic education level. The study adds to knowledge provided by previous studies concerning quality of BEC items. The previous studies mostly focused on aspects of dimensionality and DIF as measures of item quality and this cannot be said to adequately cover total test quality. The study, therefore, improves the depth of knowledge in quality of assessment, thereby serve as a foundation for emerging research using IRT.

The study has great significance to examination experts (item writers), public examination bodies, curriculum development experts, practicing teachers and government. It identified items that do not possess acceptable thresholds with respect to difficulty, discrimination and guessing index. Based on the findings of this study practising teachers, examination experts (item writers), and public examination bodies can bank items with known acceptable parameters. This means that test developers and users can use these results to decide if an item can be used as is, if it needs to be changed before it can be used again, or if it should be taken out of the active item pool. The test specifications should say what makes an item's performance acceptable in the context of what the test is for and how it will be used.

Furthermore, government through the department curriculum development may fund capacity building workshops for key stakeholders on item analysis using IRT. This study will help various stakeholders, especially public examination bodies, teachers, and policymakers to appreciate the benefits of using IRT as an alternative to CTT in estimating item psychometric qualities. The findings of the study will guide them to develop sound test items which are valid, reliable, and free from bias.

National examinations are high-stake and BEC must demonstrate beyond doubt, that decision on the examinee's progression to the next level is fair and devoid of any bias. Therefore, assessing DIF on the agricultural test will reveal if any items measured differently by gender, location, and type of school for examinees with similar abilities. DIF detection will demonstrate if the examinee's gender, location, and type of school introduce some bias. Furthermore, content analysis of items flagged for DIF revealed the sources of DIF. This is very important as it serves as red flag for teachers, curriculum specialists, and examination bodies to be alert on their instructional delivery, content development, and item generation respectively.

The results add to what is known about testing in education and psychology in Botswana. Also, the researcher found that there isn't much written about empirical studies on the psychometric properties of agriculture examination items in Botswana. It is envisaged that the study will provide a useful source of literature for educational research, especially in agriculture assessment.

For teachers, a good understanding of DIF in a core practical subject like agriculture can lead to deliberate and improved instructions during teaching and practical sessions to close the performance gap that may arise from different demographic backgrounds.

Delimitations

This study was limited to agriculture science for three major reasons: (1) researcher have personal experience about the subject matter under investigation and this became the starting point of this research, (2) it is the only practical subject at junior secondary level in Botswana which is compulsory to all students, (3) there is a lot variability in agriculture activities across the country which may impact on teaching and learning of agriculture in schools.

Although there are three student assessment pieces (multiple choice, short answer and school-based practical), this study primarily focused on multiple choice items between 2018 and 2020. The JCE Agriculture Science multiple-choice items form about 35% of the score which has implications on the score of the examinees when these items are poor. In addition, previous studies have reported that multiple-choice items are difficult to develop leading to poor discrimination indices in students with high scores (Iñarrairaegui, Fernández-Ros, Lucena, Landecho, García, Quiroga, & Herrero, 2022; Applegate, Sutherland, Becker, & Luo, 2019). The short-answer type and school-based continuous assessment papers did not form part this study. This is because the study used secondary data based on students' responses but the other two paper's score were captured as composite scores.

The study did not explore how certain key psychometric properties like distractor analysis and ceiling effect analysis. It focused on validity, reliability, and fairness through analysis of item difficulty, discrimination and guessing parameters as well as Differential Item Function (DIF) analysis.

Furthermore, although there are several measurement theories, this study only used IRT which was chosen based on its efficacy.

Limitations

Despite the significance of the study, there were three key limitations. First, the data used were only for a single BEC review cycle (2018, 2019 and 2020). This is despite the fact the observed decline in agriculture science span from the revised Junior Secondary Education curriculum implemented in January 2010. Data from earlier years has been summarised and as such did not show examinee's response patterns.

Secondly, the researcher did not have control over factors such as sample size, school, and gender distribution. For example, there was high sample size disparity of 95% and 5% examinees from public and private schools respectively. However, several studies (Kilmen, 2016; Acar, 2011; Herrera, & Gómez, 2008; Awuor, 2008), states that variabilities in sample size have a great effect on DIF detection in test items.

The study did not explore the potential intersectional effect or confounding DIF across categories of identity. The de-identification of examinee makes it impossible to form distinct intersectional groups. To accurately explain interactions which reflects the intersections of each examinee's identities, intersectional approach to DIF analyses must be explored (López, Erwin, Binder, & Chavez, 2018; Russell, Szendey, & Kaplan,2021).

Definition of Terms

Key terms used in this study are explained below.

Psychometric properties

Characteristics of a test that can estimate a latent trait.

Differential item functioning (DIF): DIF is the difference in the scores of a test item between two groups of test takers with similar abilities.

Focal group: The group of candidates whose performance in the test attracted the most attention. For example, vulnerable groups or minorities are often considered target.

Person ability: Estimate of an individual's ability based on his performance on items measuring a trait.

Reference group: A group of test takers that serves as a benchmark to compare their performance on the test with the group's performance. For example, majority group or high performing group.

Organization of the Study

This study's content was broken into five chapters. The first chapter exposes the reader to the study's background, statement of the problem the study's objective, research questions, the purpose of the study, delimitation, limitation of the study, and definition of terms.

The main aims and objectives of the study are further developed in Chapter Two. Based on research objectives the study is situated within a larger historical context of the psychometric features of the agricultural test items of the Botswana Examination Council junior secondary school examination utilizing the item response theory method. The final section of this chapter focused on significant empirical and theoretical literature review that helped with the study's design and implementation.

The approach for the study is described in the third chapter. The study design, population, sample and sampling technique, research instrument, data collection, and data analysis procedures are covered in this chapter. The background characteristics of respondents are discussed in this chapter.

The study's fourth chapter focused on data analysis and discussion of findings. The analysis was carried out in accordance with the study's research questions. The study's summary, conclusion, and suggestions will be presented in Chapter Five. This chapter will also suggest research topics for future study.

CHAPTER TWO

LITERATURE REVIEW

Introduction

The purpose of this section was to review the literature on issues related to quality of test items. These issues covered included theoretical framework, conceptual review, and empirical review. The review of related literature is organised under the following subheadings:

1. Theoretical framework
    a. Item Response Theory (IRT)
2. Conceptual review
    a. Validity
    b. Reliability
    c. Differential Item Functioning
3. Empirical review
    a. Characteristics of test items on Agriculture examinations based on 3-Parameter Model
    b. Contribution of test items in Agricultural examination to the measure of students' achievement in the subject.
    c. DIF of test items in terms of gender
    d. DIF of test items in terms school location
    e. DIF of test items based on type of school (government or private)
4. Conceptual Framework

Theoretical Framework

According to Adedoyin, Nenty, and Chilisa (2008), the measuring of behavioural change is prone to inaccuracies due to its indirect character. It is indirect and inferential in the sense that it uses what is observe throughout the indirect measurement procedure to forecast or measure what is desired. As a result, it requires theories or principles to provide guidance and a foundation for anticipating or gauging what is desire from what is perceived. CTT and IRT are two distinct but compatible measuring theories for psychometricians.

Item Response Theory (IRT)

According to Hambleton, Swaminathan and Rogers (1991), the concept of IRT is derived from the works of Thurstone on "mental development" by Frederic M. Lord and George Rasch in the 1950s. Lord (1980) noted that CTT focuses on issues related to group dependency, weak assumptions, and parallel forms of tests. Alternatively, IRT appears to provide a response to the shortcomings of CTT such as group dependency and test level statistics.

In latent Response Theory, which is another name for IRT, how likely it is to answer a question correctly or giving a specific answer depends on one's ability and the properties of the question (Aborisade & Fajobi, 2020). The purpose of IRT is to estimate the probability that a candidate of a given intelligence level will respond correctly to an item in the test. The candidate's response to test items is determined by their unobservable cognitive capabilities. These behaviours are referred to by psychometricians as latent traits or abilities. In its simplest form, the IRT recognizes that (1) the test taker's ability or disposition to pass the test is a particular ability or trait, and

(2) the test taker's probability of giving a correct answer is linked to the tester's ability and can be depicted moving by a monotonically growing curve (Zanon et al., 2016). Figure 2 shows the item-characteristic curve (ICC). The ICC is so important that it is still considered one of the assumptions of IRT.

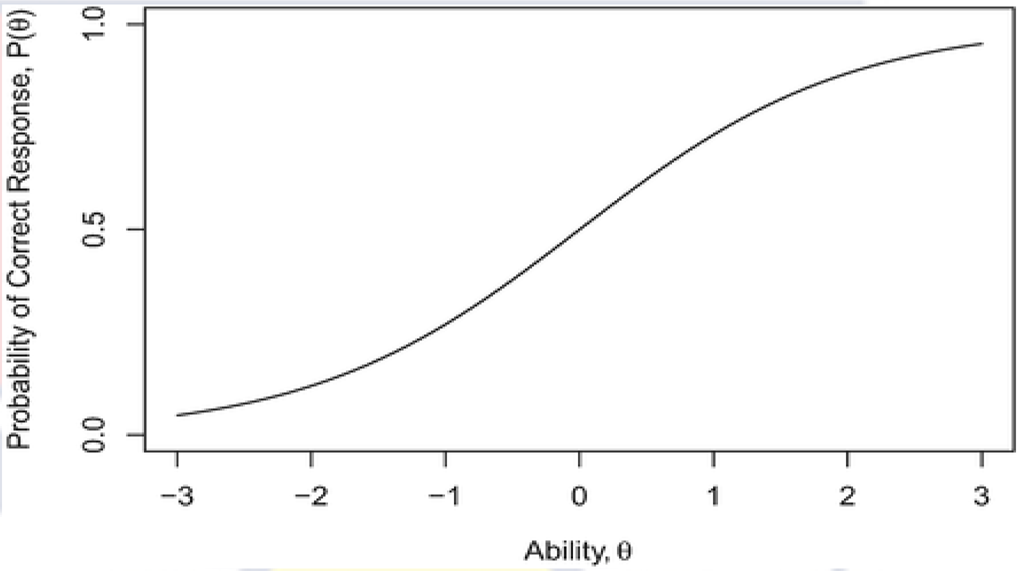


Figure 2- Item-characteristic curve

In Figure 2, the x-axis represents the test taker's ability, and the y-axis represents the probability of answering the question correctly. Thus, the "S-shaped" curve shows the probability that candidates with different abilities (theta) will answer correctly. Based on this, Effiom (2021) defined ICC as a way to show the link between a person's place on the spectrum of latent variables and the likelihood that they will give a certain response. In summary, the ICC identifies the link between the probability of giving a correct answer and a person's ability to answer correctly.

From this point of view, the chance of getting the correct answer from those with reduced abilities is less, especially more difficult ones (Effiom, 2021). Candidates who possess higher abilities are more likely to give correct answers or overcome difficult questions. For example, candidates with more knowledge about agriculture have a better chance of getting more accurate answers to questions about agriculture than candidates with less abilities.

Therefore, examination bodies like BEC should ensure ICC is determined for every item kept in their item bank. This enhances the chances of identifying items' flaws and taking remedial action since the quality of an assessment tool is key to determining if the desired change (learning) has taken place. Messick (1994), Oliveri and Lawless (2018), and Fechter et al. (2021), have said that the quality of the items has a direct effect on the validity of conclusions drawn from their results.

If the assessment tool (test) is composed of items that are quite challenging, one helpful aspect of the ICC is that it moves to the right, indicating that the test is more difficult. Conversely, if the assessment tasks are too easy, the ICC shifts to the left. According to An and Yung, (2014), it has been shown that items with flatter ICC are less discriminating because the likelihood of a right answer at less proficiency is roughly identical to the chance of an appropriate answer at strong levels of ability. The steeper the slope of the curve, the greater the item's ability to distinguish. In sum, ICCs are valuable since they show where and how well certain items perform in terms of discriminating across various ability levels. Because items and test examinees are evaluated on the same scale, it is possible to correlate the two for better measures (Osadebe & Agbure, 2020).

The Greek letter theta θ for ICC represents the standard deviation or skill level represented by the total test scores plotted on the horizontal x-axis. Similarly, the vertical axis pi (θ) represents the weight of an item and is represented by the percentage of test takers who passed the item. Embretson and Reise (2013) define P(θ) as the probability of the tester meeting the criteria.

P(θ) could be changed to the chance that a candidate will answer correctly a question picked at random from a set of questions meant to test the same skill (Nenty, 2004). Thus, P(θ) can also be seen as the associated random evaluation of the probability level to answer a question correctly. This may be the most useful way to look at it.

When assessing cognitive function, the discrimination index (*a*) shows the extent to which the test taker's responses to different items are related to their ability level. This is estimated at the inflection point of the slope of the ICC. In principle, the value of the discrimination index can range from positive infinity to negative infinity, but in practice, the index value for binary items is usually 2 or less. (Alagoz, 2005; Sabri, (2013).; Stemler & Naples, 2021; Veldkamp, 2013).

A high value indicates that the item is highly discriminating for the examinees at the turning point. However, *a* value of zero indicates that the items are not differentiating examinees with higher ability from those with lower ability. Negative indices mean that less ability candidates can answer a question better than high ability candidates, so these questions are bad and should be checked or thrown out (Pido, 2012).

Additionally, ICC (parameter *b*), provides information about cognitive demand of items. Like hurdles that that a runner must overcome to show his/her abilities, the *b-parameter* increases rather than decreases with increasing complexity of items. Ability levels are set or converted to zero mean and 1 standard deviation to allow the test designer to compare *b*-values in two directions. The converted *b*- values are usually between +3 and 3, but in practice it is usually between -2 and 2, where small values indicate simple features (Şahin & Colvin, 2015). The authors also suggest that the ICC can alert item writers if a question performs differently among different users. Items with a *b*-value around zero in the middle of the mean have moderate cognitive demand. Items with a positive *b*-value to the right of zero are difficult and only those with higher competencies can answer correctly. Items with *b*-values to the left of zero are easy and can be answered by test takers with low and moderate skills (Nenty, 2004; Pido, 2012).

However, if a question has different ICCs for different groups, it reflects different properties between the two groups. In this case, the item works better at least in one group. The c-parameter, on the other hand, represents the probability of a candidate who cannot understand the complexity of the question to overcome the object by guessing or to respond correctly to the object (Nenty, 2004).

IRT Models

In contrast to CTT, latent models have been praised for being able to produce accurate and consistent predictions (Tavakol & Dennick, 2012). IRT employs a variety of techniques to generate constructs that can identify not only the source of measurement error and uncertainty, but also how question difficulty interacts with assessments of an individual's ability. The item characteristic curve serves as a foundation for IRT models. In addition, IRT models can be used to develop, validate, and refine polytomous model metrics, in which different items measure distinct aspects of a continuous latent model (Kean, Bisson, Brodke, Biber, & Gross, 2018).

IRT was built around the concept of the normal ogive. Estimates on the ICC x-axis usually represent latent ability between -3 and +3. The y-axis represents the test taker's chance of being successful. This curve determines the examinee's ability to meet the demand for an item only as a function of the item demand in the test (Crocker & Algina, 1986). Since the probability is neither less than nor greater than one, the ICC usually has a long S-shape. For binary items, the probability of giving a correct answer, changes with the ability trait. "The main difference found in the most popular response models currently lies in the mathematical form of Pi (θ), ICC. It is up to the test item developers and users to select one of the many mathematical functions that will serve as the ICC method" (Hambelton, Swaminathan, & Rodgers, 1991, p.26).

One-parameter model

The one-parameter model is the simplest and is called the Rasch model. This model only estimates the difficulty index of each item and assumes that the discrimination parameter is constant in all items. For example, items are merely specified by a single characteristic, such as their location or their difficulty (*bi*). The Rasch model finds, separates, and estimates examinee and item measures to come up with these probabilities. The one-parameter model is ideal when the focus is to determine students' abilities only. This model is not ideal for the current study because the agriculture science JCE is used for selecting students (discriminating) who have achieved a predetermined level of competency.

In the Rasch analysis, the examinee ability is a natural logarithm of the success rate divided by how likely it is to fail, ln (p / 1-p). The higher the logits, the greater the proficiency and vice-versa (Tavakol, & Dennick, 2012). For example, if an examinee correctly answers 70% of the test items, the odds ratio for the whole test is ln (0.7/0.3) =0.84 logit, which is the examinees' ability. On the other hand, you can figure out the logit for item difficulty by switching the numerator and denominator in the above formula. So, both student ability and item ability can be shown on the same logits scale. Zero on the scale is in the middle of the range of skill and the range of difficulty.

Therefore, the findings of one-parameter frameworks possess the quality of specified objectivity, which entails that the ranking of item difficulty is identical for every test-taker regardless of ability, and the order of individual ability is also identical for all items regardless of difficulty for all test-takers (Abdullahi 2016). The one-parameter model is as follows:

$$P_1(\theta) = \frac{e^{\theta - b_i}}{1 + e^{e - b_i}} \quad i = 1, 2, 3, \ldots n$$

With

P_i (θ): probability answers correctly item I on ability θ

b_i: difficulty level parameter of item i

a_i: item discrimination

n: the number of items

D: 1.7

e: 2.718

The *b* parameter, which is also called "item difficulty" or "threshold," is an x-axis index that tells us how easy or hard an item is (see Figure 3). The point on the x-axis where the curve crosses the 0.5 probability value on the y-axis shows where the item index is. When the difficulty rating is negative, it means that the item is easy, and when it is positive, it means that the item is hard. A simple question works for low-ability test takers, while a hard question works for high-ability test takers (Baker, 2001; Thorpe & Favia, 2012). Items with a *b* value of more than 1 are very hard, while items with a *b* value of less than -1 are easy. When the value of *b* is between -0.5 and 0.5, the level of difficulty of the test questions is in the middle. Baker (2001) said that in theory, difficulty values can be anywhere from -0 to +0, but in practice, they are usually between -3 and +3. Ceniza and Cereno (2012) gave the following explanation of what *b*-values (item threshold or difficulty) mean: Very easy means less than -2, easy means between -0.50 and -2.00, average means between -0.49 and 0.49, difficult means between 0.50 and 2.00, and very hard means more than 2.00.

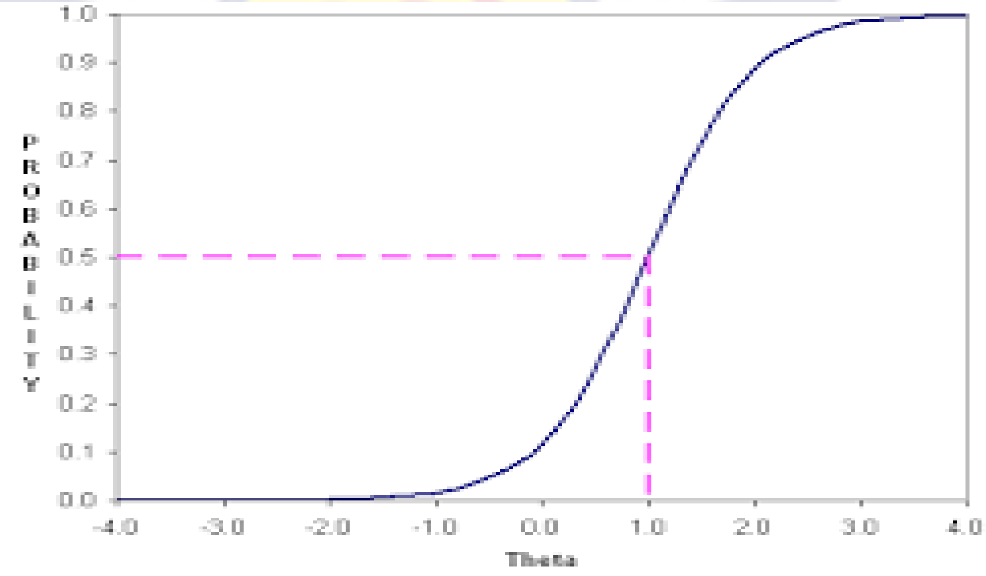


Figure 3- 1PL-Characteristic Curve

The index of an item's location is the point on the x-axis at which the curve crosses the 0.5 probability value on the y- axis. From figure 3 above the index of item location (difficulty) is 1.0 crossing the curve from the 0.5

probability on the y-axis. The discrimination parameter is assumed constant and guessing fixed at zero.

Two-parameter model

Unlike the one-parameter model, one more parameter which is discrimination is added to this model. However, guessing remains fixed at zero. This model like the one-parameter is not ideal for the current study because it falls to match examinees' abilities and guessing probability. Using it for the JCE may result in allowing students to progress to next level of study purely on luck than ability.

The two-parameter model is as follows:

$$P_1(\theta) = \frac{e^{Da^i(e-b^i)}}{1+e^{Da_i(e-b_i)}} \quad i = 1, 2, 3, \ldots n$$

P_i (θ): probability answers correctly item I on ability θ

b*i*: difficulty level parameter of item i

a*i*: item discrimination

n: the number of items

D: 1.7

e: 2.718

The (*a*) parameter, which is also called "item discrimination" or "slope," shows how well an item can separate people with low abilities to those with abilities to the right (Thorpe & Favia, 2012). Its steepness can be graphically expressed through the item characteristic curve (ICC). If *a* parameter value is positive, it means that students with higher skills are more likely to get a question right, while students with lower skills are less likely to get a question right. When *a* parameter value is negative, it means that students with high ability are less likely to get an answer right, while students with low ability are more likely to get an answer right.

Baker (2001) says the following about the range and meaning of values for item discrimination: Very low: between .01 and .34, Low: between .35 and .64, Moderate: between .65 and.134, High: between 1.35 and 1.69, and Very high: above 1.70. an item with discrimination less than 0.50 must revised or discarded (Baker, 2001).

However, Bichi and Talib, (2018) posit that discrimination value (*a*-value) greater than 1 is typically desirable for a good test, although a-values greater than 0.75 may also be acceptable. Bichi and colleague suggested a summary of the discrimination interpretation shown in in the Table 3.

Table 3- *Interpreting Discrimination Parameter Values*

Discrimination Value	Quality of an Item
$a \geq 1.70$	Item is functioning quite satisfactorily
$1.35 \leq a \leq 1.69$	Good item; little or no revision is required
$0.65 \leq a \leq 1.34$	Moderate: little or no revision is required
$0.35 \leq a \leq 0.64$	Item is marginal and need revision
$a \leq 0.34$	Poor item; should be eliminated or revised

Source: Bichi and Talib (2018).

Three-parameter model

Although all three models assume a single trait underlying examinee performance for dichotomous items and use the logistic functions, they differ in the number of item parameters they allow to vary. The 3-parameter model is all-encompassing of the three models as it adds a third parameter: pseudo-guessing (see Figure 4). According to Jia, He and Zhu (2020), this model is particularly useful for multiple-choice items and true-false testing. The model postulates that test-takers can achieve success on item by guessing the correct response.

The three-parameter model is expressed as:

$$P_1(\theta) = c_i + (1 - c_i)\frac{e^{Da^i\left(e - b^i\right)}}{1 + e^{Da_i(e - b_i)}} \quad \text{i = 1, 2, 3, …n}$$

P_i (θ): probability answers correctly item I on ability θ

b*i*: difficulty level parameter of item i

a*i*: item discrimination

c*i*: pseudo guess parameter

n: the number of items

D: 1.7

e: 2.718

The *c* parameter also known as a pseudo-guessing parameter (Thorpe & Favia, 2012) is the likelihood that an examinee with very low ability can guess the correct response to an item and therefore has a greater-than-zero probability of answering correctly. An examinee who randomly selects responses to items with four options can answer these items correctly about 1 out of 4 times, meaning that the probability of guessing correctly is about 0.25. Harris (1989) concluded that the items with 0.30 or greater c-values are considered not very good, rather c-values of 0.20 or lower are desirable. In like manner, Akindele (2003) also noted that items do not have perfect c-values because examinees do not guess randomly when they do not know the answer.

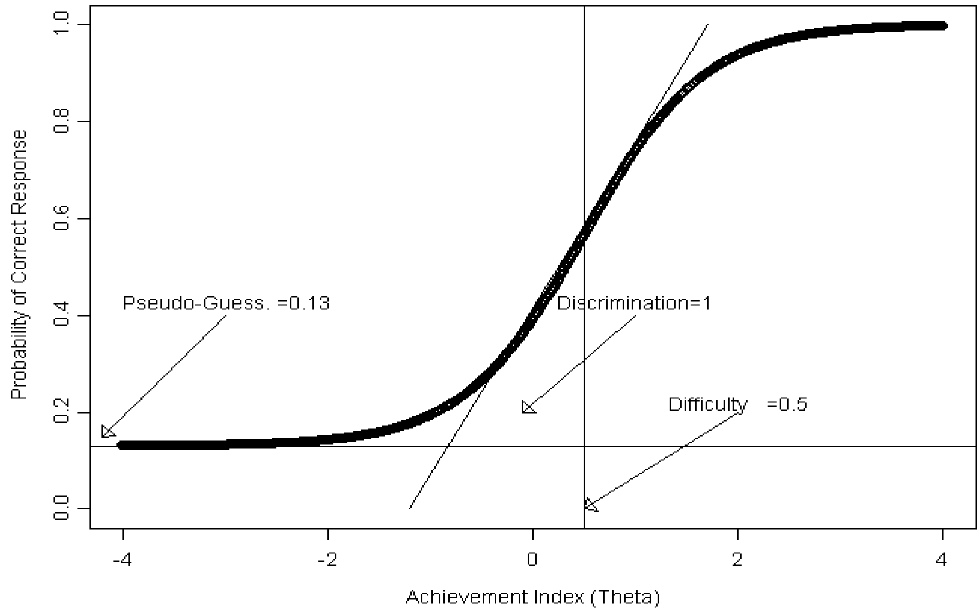


Figure 4- The 3PL-characteristic curve

According to Lord (1980), two- and one-parameter models are considered subsumed under the three-parameter model. From figure 4 both discrimination and difficulty indices are reflected together with the guessing parameter. In the 3PL model, Duong (2004) says that the lower asymptote range of probability is not from 0 to 1, but from *c* to 1 (for example, 0.13 in Figure 4). Since the JCE Agriculture Science multiple-choice items is an achievement test aimed at correct ranking and selection of students for progression to next level of study, the three-parameter model is the most suitable for this purpose. The 3PL model unlike the 1PL and 2PL models, accounts for guessing which is a common feature of multiple-choice items in estimating students' ability.

Assumptions of IRT

There several important assumptions are under the IRT framework, which includes unidimensionality, local independence, monotonicity, and item invariance as briefly explained below.

Unidimensionality

According to Tavakol and Dennick (2013), One of the assumptions of Rasch modeling particularly for dichotomous items is that a test optimally measures a single underlying construct; this is termed unidimensionality. Unidimensionality assumes that only one trait or dimension is required to explain the performance of examinees on a test. Unidimensionality in IRT models says that any set of items in the response model (i.e., items designed to measure a common trait) should be able to measure the test taker's ability on the same ability from any collection of items in the domain of items included in the model (Yang & Kao 2014).

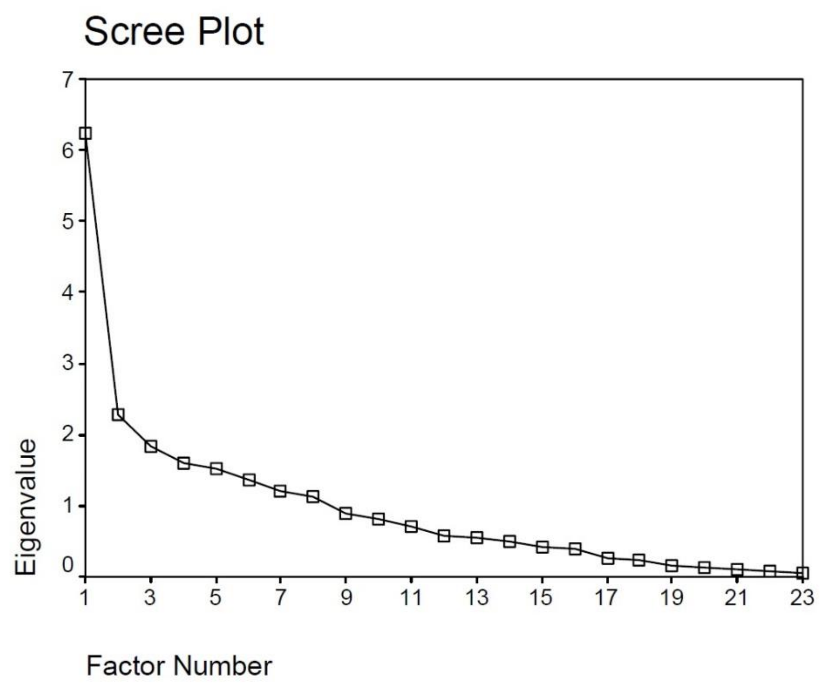
Therefore, JCE agriculture items must be from similar material/content such that its context leads to measuring a single skill. If the background of the content material is too varied, the items will have different meaning to subgroups of students responding to the items. If the item response model fits the database, the number of items (if they aren't very few) or the statistical properties of the items won't affect how well each test measures actual ability. Instead, the number of items and the statistical properties of the items will. So, this assumption is a key part of proving that the instrument is valid because it measures how much each item adds to the construct of interest.

In order to determine the unidimensionality of the tests, analysis was made on the scores of the tests. By using eigenvalues and variances, scree plots and common points, factors, features or basic patterns can be extracted or created (Field, 2005, p. 16). Factor analysis takes advantage of the interaction of factors to create a linear relationship between observed variables. Although IRT is used more frequently, it is used in both CTT and IRT paradigms (Kean, Bisson, Brodke, Biber, & Gross, 2018). Different items with interactions are considered multiple items, while items that are very different are considered a single item. According to Sheytanova (2015), all factors with an eigenvalue greater than 1 should be kept as the measure of the scale when using the Guttman-Kaiser rule. In addition, according to the Guttman-Kaiser rule, the underlying structure should explain at least 70% of the variance.

According to Hagell (2014), the Rasch model also uses principal component analysis (PCA) of residuals. The first or primary factors identified are often called Rasch factors or factors. Analysis of the remaining data will reveal more, such as the first, second, or third construct (Tavakol & Dennick 2013). Data support for measurement or structure favours one-dimensionality when the first observed variable gives an eigenvalue less than two. For example, if the eigenvalue is 4.7, which explains that about five items were measuring for other constructs, by rounding to the nearest whole number.

A scree plot is a type of graphic representation that is often used to figure out how many principal components to keep. A scree plot is a simple line graph that shows the eigenvalues for each principal component. It shows the number of factors on the left and the eigenvalues on the right. It always shows a curve going down. Most scree plots start high on the left, drop quickly, and then flatten out at some point. This is because the first component usually explains a large part of the variability, several other components explain a moderate amount, and the last components explain only a small part of the total variability. Jolliffe (2002) explained: "The scree plot shows how much each principal component explains of the total variance. Also, the decision about how many components to keep describing a dataset well can be made based on ad hoc rules like "components with a variance > 0.70 or where the cumulative proportion of variation is >80% or >90%." Figure 5 shows an example of a unidimensional scree plot.

Figure 5- Scree plot



Sources: wikepedia.com

In the Figure 5 the scree plot, the slope of the curve is clearly leveling off (the "elbow) at indicates at component 2. The number of factors above the 2 (elbow turn) that is generated by the analysis is one. At this point, an additional factor would add relatively little to the information already extracted. Consequently, in this example the underlining constructs is unidimensional.

Local Independence

IRT also works on the idea that each area is independent. Even though the latent trait is included in the equation or quantitatively fixed, it is assumed that the exam takers' answer patterns have nothing to do with each other statistically (Yang, 2014). Thus, response to each item in the test is statistically independent of the responses to all other items on same test if ability is held constant. However, violation of local independence by JCE agriculture items may lead to double award or double punishment if examinees an item right or wrong respectively.

Another way to test this idea is to examine discrimination a. If items are highly correlated or predictive, they will have very high scores (eg >4) relative to other items in the test (Hambleton, Swaminathan, & Rogers, 1991). According to Reeve and Fayers (2005), few variables analysed by the residual correlation matrix may indicate a violation of this assumption. Items with good correlation according to the Rasch model indicate that one of the two items is not valid for the test. Correlation between items greater than 0.50 is considered an indicator of response dependency and items should be evaluated.

However, according to Hambleton, (1993) in the latent space, the local independence and the unidimensionality assumptions are similar and therefore factor analysis methods can also be employed for the former because once unidimensionality is met; local independence is assumed to be met.

Item characteristic curve (ICC)

ICC is used in response to the development of logistic curves (Hachey, 2008). It depicts the relationship between an examinee's level of ability on an item located on the horizontal axis and the probability of displaying the ability on the vertical axis. The ICC is monotonic indicating that the probability of endorsing an item will continue to increase as the trait level increases (McCarty, 20050). For example, an examinee with a 1.7 ability level will have a higher probability of correctly answering an item than one with a 1.5 ability level, and an examinee with a 1.5 ability will have a higher probability of answering the same item than one with a 1.2 ability level. In sum, the higher the ability, the higher the chance examinee will respond correctly.

Item invariance

One of the assumptions made by IRT is that the item attributes and underlying features are invariant across multiple samples with varying features. According to Abedalaziz (2011), invariance means that the item parameters (such as difficulty, discrimination, and guessing) do not depend on the ability distribution of any particular group of examinees, and the examinee ability parameter (θ) does not depend on a particular set of test items. This means that, for a well-specified IRT model, the ICC for the two groups being studied will be the same, no matter how the abilities of the people in the groups are spread out (Humbleton, Swaminathan & Rogers, 1991). IRT is a good way to look into DIF because of this property, as the presence of different ICCs is a sign that an item works differently for two groups (Abedalaziz, 2011).

According to Embretson (1996), "item invariance allows one to obtain unbiased estimates of item parameters from unrepresentative samples (e.g., low-trait groups vs. high-trait groups) if the data fit the model" p. 341. The assumption of item invariance should theoretically hold in all cases, but in real life, the data does not always support this. This may be due to poorly written materials or materials that different samples interpret differently. When a subject behaves differently in subgroups after controlling for ability, it is considered to have DIF (Effiom, 2021). DIF can be observed in item discrimination and item difficulty. Items with DIF will display different sets of ICCs or CRCs in different groups. However, it is important to first assess the invariance of the basic property of these two groups.

IRT Model Fit Assessment

Any time an IRT model is used, it is important to figure out how well the model's assumptions fit the data and how well the test data fit the IRT model that was chosen for that situation. If the IRT model's assumptions aren't met or if the IRT model used doesn't match the test data, the IRT model parameters can be estimated in a way that is wrong or unstable.

In this study, the assessment of the fit of the IRT model will be carried out on the total agricultural science research population. The large sample size used in assessing IRT model assumptions and model fit should provide a stable a reliable result on assumption and model fit.

Statistical Procedures for Testing DIF

Some parts of the history of the development of DIF methods are like the history of IRT. It might not be enough to talk about DIF without talking about IRT-based ways to get DIF. Methods for diagnosing DIF can be put into two main groups: IRT-based methods and methods that don't use IRT. According to Wang and Su (2004), there are two methodological approaches to obtaining DIF: those that rely on the IRT model, and those that do not rely on the IRT. First, IRT model measurement is required, and a mathematical evaluation process is followed, based on statistically significant assets derived from measurement results.

The determination of the presence of DIF in items can be based on either IRT methods or non-IRT methods. However, (Hambleton, Swaminathan, & Rogers, 1991) use the terms parametric and nonparametric instead. The IRT methods makes a comparison of item parameters and the area between Item Characteristic Curves (ICC). The second group included Classical Test Theory (CTT) methods, Factor Analysis (FA) methods, and Categorical-Data-Analysis-based methods (Gomez-Benito & Navas-Ara, 2000).

Item Response Theory encompasses a family of mathematical models that specify the possible relationship or interaction of test-taker response, his/her latent trait, and the item demands. It is imperative to correctly identify DIF if present in items because the precise functionality of those items between groups will then be established (Hambleton, Swaminathan, & Rogers, 1991).

Furthermore, Murphy and Elliot (2006) point out that of all DIF detection methods, those based on IRT, especially the three-parameter IRT model, are considered theoretically sound. This is because IRT uses the Item Characteristic Curve (ICC) to express the interaction between the ability of the test-taker and the probability to respond to an item correctly.

Methods of Detecting DIF

There is no "best method" of DIF analysis that works well and can be used for everything (Lai, Teresi, & Gershon, 2005). Thus, procedures for detecting DIF should try to measure both uniform and non-uniform DIF, but not all methods can actually find non-uniform DIF. This is because DIF detection procedures for uniform or nonuniform ordinal DIF use some way to measure how different the two groups are in how well they do on the item. When DIF is not uniform, some of the differences will be good and some of them will be bad. So, most DIF detection methods that were made to find uniform DIF can't find non-uniform DIF, which is sometimes called crossing DIF.

McNamara and Roever (2006) say that there are four main ways to look for DIF: (1) analysis based on item difficulty (comparing estimates of item difficulty); (2) nonparametric approaches (using contingency tables, chi-square, and odd ratios); (3) item-response-theory-based approaches (including 1, 2, and 3-parameter analyses, which often compare how well statistical models fit); and (4) other approaches (including logistic regression, generalizability theory, and multifaceted measurement).

Gao (2019) also said that researchers have come up with two main types of methods for finding DIF. The first is the Mantel-Haenszel (MH) procedure (Holland and Thayer, 1988) and the Logistic Regression procedure, which use observed total scores as estimates of the trait being measured (Swaminathan & Rogers, 1990). The second type of method is based on latent variable models in which the trait is directly estimated. In recent years, researchers have become more interested in these methods, such as IRTLR tests (Thissen & Steinberg, 1988), Lord's 2 test (1977, 1980), the improved Wald test (Langer, 2008; Woods et al., 2013; Tay et al., 2015; Cao et al., 2017), and the Multiple Indicators Multiple Causes.

In the context of item response theory (IRT), an item functions differently between two groups when: (1) the item parameters estimated from two groups of examinees are significantly different (e.g., Draba, 1977; Lord, 1980; Wright & Stone, 1979); (2) the area between item response functions (IRF) estimated from two groups is significantly large (Kim & Cohen, 1991;

Wainer, 1993); or (3) the likelihood functions obtained using likelihood ratio from two groups of examinees are significantly different (e.g., Thissen et al., 1993).

Also, French and Miller (1996) say that the best way to find DIF in dichotomous items is to use IRT methods. Recently, research has been done to see if these methods can also be used for cases with more than one answer (Swaminathan & Rogers, 1990; Wainer, Sireci, & Thissen, 1991). IRT models are a cool and useful way to understand and model DIF, but they are not the most common way to find DIF. IRT's DIF happens when two or more groups of test-takers answer a question differently about a latent variable (Lai, Teresi, & Gershon, 2005).

Deciding on the Magnitude or Extent of DIF

Based on what Cohen (1992) said about small, medium, and large effects, Zumbo and Thomas (1996) came up with rules for figuring out if a DIF is small, medium, or large using R2. They said that R2 values below 0.13 meant there wasn't much DIF, between 0.13 and 0.26 meant there was some DIF, and above 0.26 meant there was a lot of DIF. A 2-df chi-square test was also used to decide if an item was statistically significant in both the medium and large categories.

Miceli, Marengo, Molinengo, and Settanni (2015) said that a nominal Type 1 error rate of 5% should be used to compare the percentage of DIF found over the total number of tests to figure out how big DIF is in achievement tests. They also say that 5% is the level below which DIF in the test can be expected to be due to chance alone. A percentage that is higher than the accepted error rate should be taken as proof that the test has a significant DIF.

Mahmoud (2021) says that IRT has different ways to find DIF that depend on the model of the theory. All of these ways are based on the principle that the estimated item parameters are different between the focus group and the reference group.

Likelihood-Ratio test (IRT-LR)

Thissen, Steinerg, and Gerrard (1986) and Thissen, Steinerg, and Wainer (1993) came up with the likelihood ratio test to figure out how important differences in how different groups answered questions were when using an IRT model (Cohen et al., 1996). To make the comparison, nested two-group item response models with different constraints are statistically compared to see if the item response function (IRF) for a certain item is different for the reference group and the focal group (Woods, 2009). The two models are called the compact model, which assumes that there are no group differences, and the augmented model, which tests one or more items for DIF (Thissen et al., 1993). For each augmented model, the IRT-LR method assumes that some items (an anchor set) do not have DIF. The likelihood-ratio test (IRT-LR) is based on figuring out the LR value, which is given by the following equation:

$$LR = -2logLC - (-2logL_A)$$
$$= -2logLC + 2logL_A)$$

Where LC is the compact model's log-likelihood and LA is the extended model's log-likelihood (Cohen et al., 1996). Under the null hypothesis, the degrees of freedom are equal to the number of estimated parameters, so the LR statistic has a 2 distribution. If the LR statistic is significant, you need to do follow-up tests to compare how well the two models fit, keeping all of the item parameters the same except for one (parameter an or b). The point of IRT-LR is to find out if letting the item's parameters vary between groups improves the fit of the model by a lot (Price, 2014).

Notwithstanding, Hambelton, (1989) emphasized that to obtain accurate estimates for the parameters of the item in IRT, a large sample of attest 1000 is required. However different studies as cited by Mahmoud (2021) propose different sample size as for example, Goldman and Raju, (1986) suggested that the minimum sample size for estimating accurate parameters in the (Rash) model is (250), Guyer and Thompson (2011) recommended 300, whereas Thissen and Wainer (1982) recommended 500. All the studies that were cited show that sample size is a key factor in how well DIF detection works. It is important, then, that researchers try to get the biggest sample sizes they can, considering things like time and resources. So, making the sample as big as possible is helpful because it affects how reliable the results are.

If there are two groups, the (LR) values follow a chi-square distribution with the same number of degrees of freedom as the number of parameters to be estimated (focal and reference group). Significance means that there is a difference between the estimated parameters of two models. This means that you need to do more tests to compare how well the two models fit the data. The main idea behind IRT-LR is to see if the parameters for an item are different between groups in a meaningful way (Price, 2014, as cited by Mahooud) (2021).

Lord's Chi-square test (Wald Test)

Lord (1980) came up with a test called the Wald test, which uses the 2 statistic to look for DIF under IRT. The Wald test compares vectors of IRT item parameters between groups to find DIF items. In the 2-parameter logistic model used by IRT, the parameters for item discrimination and item difficulty are compared. If the vectors of an item's parameters are different between groups by a lot, then the item works differently for these groups. Lord came up with a test to measure the significance of DIF for location parameters only in two-group studies.

Generally, IRT-related approaches, according to Lai et al. (2005), involve comparison of the Item Characteristic Curves (ICCs) or comparison of the item parameters, and they don't use observed scores like the MH and LR methods do. IRT-based procedures limitations are tha they are sensitive to sample size and model-data fit, take a long time, and don't have tests of significance for indexes like the area between item characteristic curves.

Raju's area measurement

This is one of the DIF detection methods. This procedure is an IRT-based technique developed by Raju to detect differences in the likelihood of respondents/examinees with same ability but due to group membership, they have different probability of getting an item correct (Raju, 1988). This technique is referred to as Raju's area measurement or measure. As an IRT procedure, Raju focuses on the region between the trace lines of the reference and focus groups (Wright & Oshima, 2015; Yörü & Atar, 2019). Figure 6 is a pictorial representation of Raju's method.

Figure 6 shows the trace lines for the focal and reference groups. It is evident that the probability getting an item correctly is higher for the focal group than for the reference group as most of the levels of ability. This indicates that that the item favours the focal group. Therefore, the DIF estimate from Figure 6 is the gap (d_i) between the two trace lines.

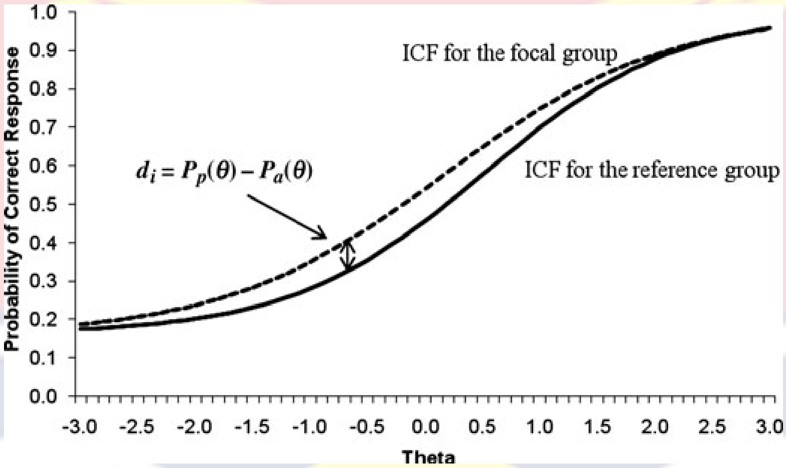


Figure 6- Raju’s measurement area
Source: Oshima and Morris (2008)

From this procedure, a non-zero value for the region between the trace lines is indicative of DIF. However, a zero figure suggests the presence of no DIF. In the case of the former, the Z-statistic is estimated using the estimates of the item parameters as well as their associated standard errors. The Raju’s Z-statistic results in a noncompensatory differential item functioning (NCDIF) value. The NCDIF values are classified as ‘A’, ‘B’, and ‘C’ levels, respectively, based on the following: less than .003; greater than or equal to .003 but less than .008; and greater than .008 (Wright & Oshima, 2015).

Further to the use of Raju’s procedure, scholars have admonished the determination of effect size to determine the magnitude of DIF present (Cohen, 1988; Monahan et al., 2007; Wright & Oshima, 2015). This in addition to the significance testing may provide the basis of whether or not to remove an item that has been flagged as DIF. Generally, the weakness with relying solely on significance testing is that, it is susceptible to large sample size. Therefore, for studies with relatively larger samples like the current study, an item with very small or negligible practical significance may be flagged as DIF due to the large sample size. Contrary, in a relatively small sample study, an item with a large magnitude of DIF may not be statistically significant, suggesting the presence of no DIF. Notably, significance testing is purely a function of sample size.

In response to the call for estimation of effect size in DIF, Wright and Oshima (2015) developed a more current effect size estimate for Raju's NCDIF based on ETS' Mantel–Haenszel categories at various levels of the 3-parameter (*a*, *b*, and *c*). Put differently, Wright and Oshima's procedure takes into consideration the IRT model involved, discrimination, and difficulty parameters.

Conclusively, despite the several methods of detecting DIF magnitude and direction highlighted above they seem to be a consensus that DIF detection alone is not enough. Once DIF is identified, subsequent item content analysis by subject experts must follow for a deeper and comprehensive understanding of its sources of origins (Liao & Yao, 2021; Özkan, & Güvendir, 2021; Karami & Salmani Nodoushan, 2011; Buzick, & Stone, 2011; Ferne & Rupp, 2007). In sum, merely reporting on DIF based on parameter estimates is not enough to inform item developers or policy.

Conceptual Review

This section evaluates different key concepts on psychometric properties of test items.

Concept of Validity

According to Rust (2007), in psychometrics, validity is one of the four fundamental principles by which assessment quality is judged. Rust further defined validity as the extent to which a test or examination assesses what it is intended to assess. Denga (2003) agrees with this and defines validity as "the extent to which a test is true, accurate, or relevant in measuring the trait it is meant to measure". The implication of these two definitions is that any tool that cannot accurately measure what is expected of it is useless. Validity in IRT refers to the extent to which individual test and test items rank well in the ability that the test items measure, that is, the ability of any test to assess individuals according to their ability and rank items according to their level of difficulty (Hambleton, 1994; Bich & Talib, 2018).

But a review of the literature shows that validity has changed in several ways in recent years. Messick (1989) said that validity should be seen as the meaning of a test score, not as a property of a test or any other assessment tool. Standards for Educational and Psychological Testing (AERA, APA, and NCME, 2014) and Messick (1989) looked at validity as a single idea that considers evidence from many different sources. Evidence must match the types of meaning or interpretation that are wanted for a measure.

Validity is a single idea, but it can be broken down into different parts to draw attention to issues and differences that might otherwise be overlooked or downplayed. For example, the social effects of how well you do in school or what the meaning of score means when it's put to use (American Educational Research Association, American Psychological Association, National Council for Measurement in Education, 1999). They came up with five lines of evidence for the validity of Educational and Psychological Testing based on (a) test content, (b) response processes, (c) internal structure, (d) relationships to other variables, and (e) the results of testing.

A construct is something that can be measured or put into practice by linking it to something that can be seen (Fulcher & Davidson, 2007). In the field of testing, it is generally understood that a construct is what we want to test. This means that it is the same as the traits, abilities, performance, and characteristics of the test subjects that we want to test (Fulcher & Davidson, 2007; Bachman, 1995). So, it must show proof that the content is relevant, that the sample is representative, and that the technical quality of what is being measured is good. So, tests are operational definitions of constructs because they show how the thing being measured works (Fulcher & Davidson, 2007). Messick (1989) says that construct validity is "an integration of any evidence that has something to do with how test scores are interpreted or what they mean" (p. 17). In the end, the definitions of "construct" point to a broad approach based on test content, the way items are put together inside, and how the scores are interpreted.

So, content validity is defined in terms of subject matter sampling, how it shows up in the desired behaviour, or in how people react to items drawn from a hypothetical universe of response situations. Constructs must be present in how people react to all sample items. So, the correlation between the intended independent variable (the construct) and the proxy independent variable (the indicator or trait) that is used can be used to measure construct validity.

To this end, when an indicator of content is expressed as a multiple-item instrument which is the case in this study, factor analysis (FA) is used to validate the construct. FA is a commonly used and widely promoted procedure for developing and refining psychological assessment instruments to obtain evidence of the construct validity of measures (Tavakol & Wetzel, 2020). Furthermore, a strong association between construct validity and FA provides evidence based on test content and evidence based on internal structure, which are key components of construct validity (Wetzel, 2012). Establishing construct validity for the interpretations from a measure is key to the quality of quality assessment.

Factor analysis is divided into two main categories namely, Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Williams, Onsman, & Brown, 2010). With EFA, the researcher does not anticipate the number or nature of the variables and, as the name suggests, is exploratory in nature. That is, it allows the researcher to explore major dimensions to generate a theory or model from a relatively large set of latent constructs often represented by a set of items (Pett, Lackey et al. 2003; Swisher, Beckstead et al., 2004; Thompson 2004; Henson & Roberts, 2006). While in CFA, the researcher uses this approach to test and design a theory (CFA is a form of structural equation modeling) or model, and unlike EFA, it has a priori theory-based assumptions and expectations regarding the number of factors and which factor theories or models fit best.

Factor analysis uses statistical procedures for simplifying a set of complex variables and exploring relationships between multiple variables/items (Tavakol & Wetzel, 2020). It unearths hidden patterns to illustrate how those patterns overlap and show what features are seen in multiple patterns This approach is used to analyse relationships between test items within a subset of test takers' responses in order to analyse the dimensionality between different items (Bandalos, & Finney 2018; Koyuncu, & Kline, 2016). Factor analysis is a useful method for figuring out how much construct validity there is. So, the identification or description of a factor or dimension is an inference about the existence and nature of a construct like an ability, trait, or other psychological function. The degree of correlation between a test and a factor, known as its factor loading, is an indicator of the test's validity with respect to the construct.

For its execution, factor analysis can be conducted using various statistical software such as SAS, SPSS, and R (Tavakol & Wetzel, 2020). The exploratory procedures statistically analyze the interrelationships between the instrument items and domains to uncover the unknown underlying factorial structure (dimensions) of the construct of interest. Factors that contribute more to construct will have the highest factor loadings. Factor loadings are like correlation coefficients in that they can vary from -1 to 1. The closer the factors are to -1 or 1, the more they influence the variable (Pett, Lackey et al., 2003). Therefore, a factor loading of zero would indicate no effect.

For the EFA process to be easier to understand, a correlation matrix should be used to show the relationships between the different variables. Henson and Roberts (2006), researchers use the correlation matrix the most. Hair et al. (1995) and Darroch (2003) used a different rule of thumb to classify these loadings: 0.30 = minimal, 0.40 = important, and.50 = practically significant. If none of the correlations are higher than 0.30, the researcher should think again about whether or not factor analysis is a good statistical method. In other words, a factorability of 0.3 means that the factors explain about 30% of the relationship in the data. In practice, this means that a third of the variables share too much variance, making it hard to tell if they are correlated with each other or with the dependent variable (multicollinearity).

In sum, looking at strands of validity evidence as identified by (American Educational Research Association, American Psychological Association, National Council for Measurement in Education, 1999; 2014), factor analysis provides evidence based on content and internal structure collectively are indicators of construct validity.

Even so, construct-irrelevant variance is the most obvious threat to validity that comes from wrongly using a test or misinterpreting what the scores mean (Messick, 1989; Henning, 1987). They said that construct-irrelevant variance is made up of variables that have nothing to do with the thing being measured. It happens when test results are affected by things that have nothing to do with the thing being measured. Bachman (1990), which Zheng (2017) cites, says that a person's background knowledge, personality, characteristics, test-taking strategies, and general intellectual or cognitive abilities may all be irrelevant, and that people should try to keep these kinds of influences to a minimum. So, variables that make it hard to understand scores or ratings in a meaningful way on a regular basis rather than by chance are examples of construct irrelevant variance.

Test bias is a big problem to construct validity, so test bias analyses should be done to look at test items (Ford & Scandura, 2018). When there is test bias, it changes how a psychological construct is measured. But the fact that there is no test bias does not mean that the test is valid (Obine,2008). In other words, the absence of bias test is important, but it is not enough.

Concept of Reliability

The reliability of scores from psychological or educational tests tells us a lot about how accurate the measurements are (Chan ,2014). Reliability of scores from a psychological scale or test refers to the consistency of measurement and is an essential component to ensure the validity of test scores (American Educational Research Association, American Psychological Association, and National Council on Educational Measurement, 2014). Culligan (2008) defines reliability as a measure of the consistency of the application of an instrument to a particular population at a particular time.

According to Bean and Bowen, (2021), the main function of a test is to estimate a score that represents an examinee's position along the continuum of the measured construct. Using IRT, the score will be expressed using the theta (θ) metric and will typically range between 3 and +3. Examinees with higher test scores correspond to greater levels of ability. In estimating the scores of examinees, we are concerned with the precision of our estimates. Consequently, it is essential to determine how close an estimate is to a population value by employing precision measures such as standard errors and confidence intervals.

The conventional evaluation of the reliability and precision of score estimates rely on an internal consistency reliability coefficient (e.g., alpha, omega) and a standard error of measurement coefficient derived from a classical test theory model. These coefficients assume that test reliability and estimated precision using confidence intervals for standard error of measurement are identical for all scores (DeMars, 2010). In contrast, in IRT models, the item information function, conditional standard error function, and conditional reliability allow for the identification of optimal scores (Demars, 2010; Bean & Bowen, 2021).

Information is a statistical term for how well an item and a test can predict how well a student will do on a certain level (Baker & Kim, 2017). Information is calculated both at the item level and at the test level, with more accurate score estimates coming from higher levels of information. Item-level information shows how well each item adds to the accuracy of estimating a score. Information about a test is the sum of the information values of the items that make up the test or instrument (Baker & Kim, 2017). One important thing about both item and test information is that both depend on (DeMars, 2010). So, conditional reliability means that an item may be less (or more) reliable at different points along ability continuum.

An item information function summarizes the extent to which items provide statistical information about the latent trait. Figure 7 illustrates the relationship between item information and conditional standard errors. The solid Information line (θ) represents the information function of the item. The range $0 \leq \theta \leq +2$ on the item provides the most information. Standard error (line with dots) shows how estimate precision varies across, with smaller values corresponding to greater estimate precision.

Classical approach to reliability was replaced by the Test Information Function (TIF) and the conditional SEM, which is written as the inverse of the square root of the TIF. These can be estimated at any level on the score scale to show how precise the measurement is (Hambleton, Swaminathan, &

Rogers, 1991; AERA, APA, & NCME, 2002). Thus, in IRT, reliability varies across the range of theta and therefore there isn't one value.

This IRT property is unique to item response models and makes it easy for test developers to choose which items to use based on their own preferences and how those items will affect the overall test information. For example, if the goal is to choose test takers with high levels of the latent trait, items that give more information on the right side of the scale may be chosen to make sure scores at the top of the ability scale are as accurate as possible (Hambleton et al., 2016). They also say that standard errors are inversely proportional to test information functions, so the two must be used together for meaningful interpretation.

Standard Error of Measurement

Estimates of the standard error (ie, the standard error of the parameters, SE) in IRT are a function of the set of respondent-endorsed item responses, which is the square root of the reciprocal of the sum of the information provided by all response items expressed in the equation below. In IRT, each item provides information about the respondent's level or ability, represented by item information Ii (θ). The standard error of the estimate, SE(θ), is the square root of the reciprocal of the sum of the information across all answered items (Embretson,1996):

$$SE(\theta) = \sqrt{\frac{1}{I(\theta)}} = \sqrt{\frac{1}{\sum_i I_i(\theta)}}$$

In sum, the more information a test provides on a particular latent trait, the better the measurement and the smaller the measurement errors at that level (Hambleton et al., 2000). The amount of item information is proportional to the standard error of estimate (SEE) for each possible θ (De Ayala, 2013).

Smaller SEE means more confidence in the estimate of θ and thus more information about individuals with that value of θ. In agreement, DeMars, (2010), posits that an item provides the greatest amount of information near the difficulty value ("b") because there is the least amount of variability (error) near this value (see Figure 7).

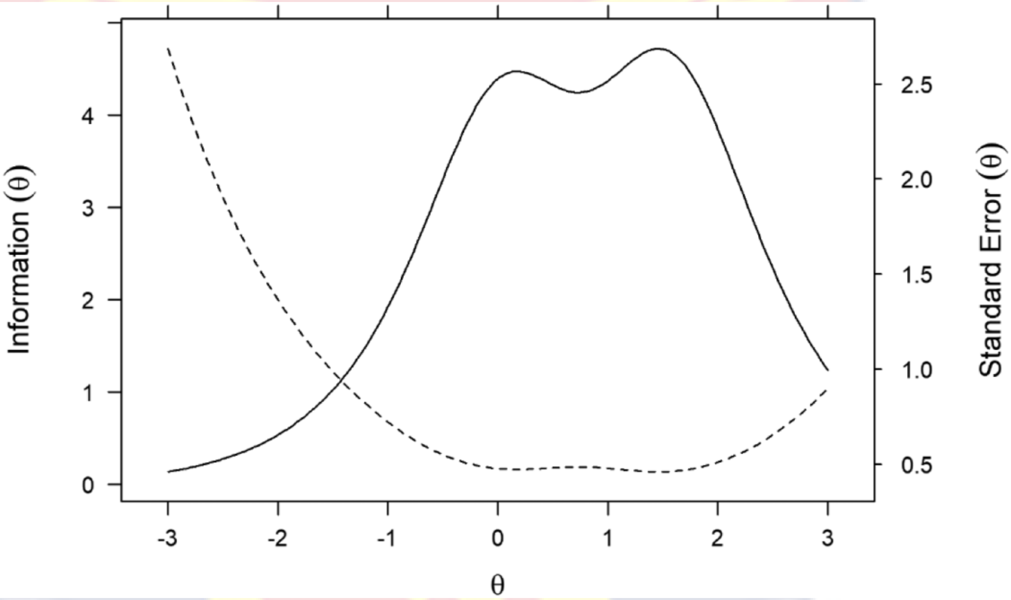


Figure 7- *Item information function*
Source: Bean and Bowen (2001)

Concept of Differential Item Functioning

Bias analysis date back to the start of the 20th century (McNamara & Roever, 2006). Researchers were hard at work at the time making tests to measure "raw intelligence. Several studies done at the time, though, showed that the socioeconomic status of the test-takers was a factor that made it hard to tell what was going on. So, they tried to get rid of some of this difference by getting rid of items that did different things for people with high and low socioeconomic status. In the 1960s, bias studies moved away from intelligence tests and toward areas where social justice and fairness were important (Angoff, 1993). This brought up how important fairness is in testing.

Test fairness and bias are also part of the field of psychology. Quantitative analyses are often used to find or stop unfair tests by figuring out if test items or test scores mean the same thing for different groups of test takers (Camilli, 2013). Differential item functioning (DIF) procedures were used to figure out when differences between groups were caused by test design and not by real differences in skill levels. This was done as part of an analysis of how well scores on individual items were comparable across groups. In educational measurement, a test item shows DIF when examinees from two subpopulations have different chances of giving a certain answer, even when the trait being measured is taken into account. For example, let's say that two groups of people with the same level of ability (females and males) have different chances of getting a question right. Since the item shouldn't show anything other than the subject's ability, the link between sex and item score should be gotten rid of.

Differential item functioning (DIF) analyses were first done in the 1960s because people were worried that cognitive ability tests were biased against people from certain groups (Angoff, 1993). The original goal of DIF analyses was and still is to find these items early on in the process of making a test so that they can be changed or taken off the final version. DIF could be caused by many things in different backgrounds. For example, people from different cultures might have a different understanding of how a test question is worded. Gender, race, native language, and socioeconomic status are the things that are most often studied.

DIF analyses were first used in educational testing to see if items on a test were unfair to, say, women or a certain ethnic group, even after the overall test ability of that group was considered. DIF analysis is also a good way to learn about personality constructs (Sheppard, Han, Colarelli, Dai, & King, 2006). Test bias shows that there are psychological differences between different subgroups. Even though the first DIF analyses were done to learn more about differences in overall test scores, it is important to appreciate that it is done at item level. This is because DIF analyses try to separate differences in performance on each item from differences in performance on an overall measure.

In fact, DIF analyses account for differences in the total performance. in ability, which is often called "impact," by matching respondents based on an estimate of the latent ability being measured using the score on a matching subtest of items that are thought not to be functioning differently. Then, the performance of these matched subsets of respondents (such as males and females) on each item is compared to see if there are any differences. Zumbo (1999) says that for measurement experts, DIF often means that the test or measure is picking up a type of systematic but construction-irrelevant variance. Also, group membership is linked to the source of construct-irrelevant variance.

There is difference between "item impact" and "DIF" (Park, 2006), so "item impact" may be present when test takers from different groups have different chances of passing an item because their abilities are measured differently. When this happens, there are "true" differences between the groups in the ability being tested by the item, which shows up in how people do on the item. Zumbo (1999) also explains the issue by saying that when examinees from different groups rate an item differently, item impact is clear because there are real differences between the groups in the ability being measured by the item, and item bias happens when some aspect of the test item or testing situation that isn't related to the test's purpose makes one group less likely to score well. As he goes on to explain, the difference between item impact and item bias is that group differences are caused by either relevant or irrelevant test characteristics.

Differential item functioning (DIF) is present for a test item when two subpopulations with the same level of a trait have different chances of getting it right. When there is a DIF item, the same true trait levels for examinees from different subpopulations could lead to different total test scores or estimates of trait levels, even though the true trait levels are the same.

DIF is needed for item bias to happen, but it's still not enough. In other words, there is no item bias if an item does not show DIF. Still, when DIF is clear, further analyses of item bias (such as content analysis or empirical evaluation) are needed to prove that an item is biased. In any DIF study, there are at least two groups: the focal group and the reference group. The group that might be at a disadvantage is the focal group, which might be a group of minorities. The group that might benefit from the test is known as the reference group.

Differential Item Functioning is the same as statistical bias, which is when one or more of the statistical model's parameters are under- or overestimated (Camilli, 2006; Wiberg, 2007). If an item has DIF, the source(s) of this difference should be looked into to make sure it's not a case of bias. Any item that is marked as having DIF is biased if and only if the source of variance has nothing to do with what the test is trying to measure. In other words, it is a source of variance that has nothing to do with the construct, and the groups perform differently on a given item because of a grouping factor (Messick, 1994). DIF is not proof that the test is unfair. It's evidence of bias if, and only if, the thing that causes DIF doesn't have anything to do with the test's main idea. If that factor is part of the construct, it is called "impact" instead of "bias." It is up to the person judging the item to decide if the real cause of DIF is part of the thing being judged. Usually, a group of experts is asked to weigh in on the interpretations to give them more weight.

DIF analysis only gives part of the answer to questions about fairness. It only looks at how different two groups did on a single item. So, DIF is not used when there are no groupings in a test. But when things are put into groups, there is a chance that they favor one group over another. If this happens, the test might not be fair for the group that did not do as well. So, in these situations, DIF analysis should be used to solve the problem. But psychometricians need to know the different kinds of DIF to help them analyze and understand it. Different types of DIF can be made based on different things. For instance, French and Miller (1996) say that there are two kinds of DIF: uniform and non-uniform. Uniform DIF is when one group always likes something more than another group across the ability continuum. That is, an item has uniform DIF if it has the same amount of DIF no matter what the value of is. On the other hand, non-uniform happens when test-takers' ability level and how well they do on an item interact in a way that causes the direction of DIF to change along the ability scale. When there is non-uniform DIF, the size of the effect changes based on. In other words, Teresi (2004) explains that "uniform DIF means that the DIF is in the same direction across the whole spectrum of disability," while "nonuniform DIF" means that an item favors one group at some levels of ability and another group at other levels (p. 2). In terms of item response theory, nonparallel item characteristic curves show that DIF is not uniform. Also, nonuniform DIF is much harder to understand because there is an interaction between the ability level of the test takers and the group they belong to (Park, 2006).

When the b-parameters (which measure how hard an item is) are different between groups (called the Reference group and the Focal group), uniform DIF is found. In this case, item characteristic curves (ICCs) do not cross (as can be seen on Figure 8). This means that test takers in one group with the same trait levels would have a lower chance of getting this question right than those in the other group. On the other hand, nonuniform DIF is found when either the 'a' parameter (item discrimination) or both the 'a' and 'b' parameters are different between groups. Since items discriminate between the two groups in different ways, their ICCs are not parallel and may cross. So, questions with nonuniform DIF are less discriminating for one group than for the other, while questions with uniform DIF are harder for one group to get right (Woods, 2008). Figure 8 shows the ICCs for DIF that is both uniform and not uniform.

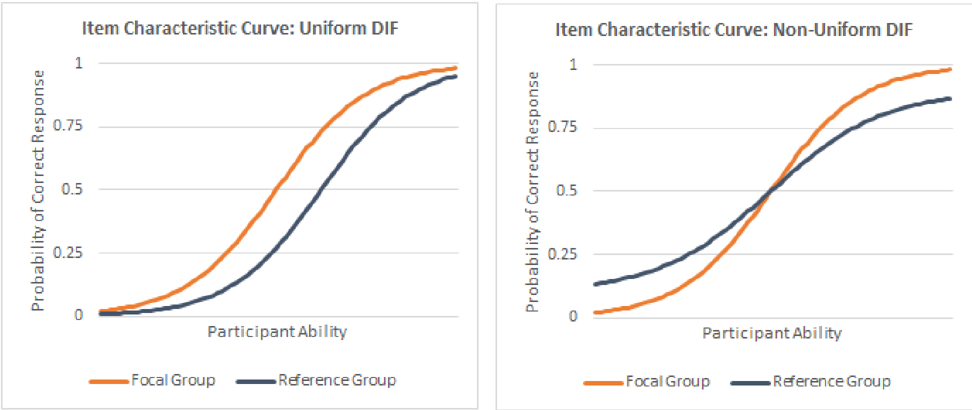


Figure 8- ICCs for uniform and non-uniform DIF

Even though the literature has talked a lot about uniform and nonuniform DIF for dichotomous items, it's likely that the same distinctions can be made for polytomous items, which are scored on more than two categories (e.g., performance assessment items or Likert-type items). For a polytomous item with k categories, however, there are k – 1 ICCs, which makes these classifications much more complicated and difficult to use (Kim, Cohen, Alagoz, & Kim, 2007). DIF can happen with polytomous items, and most of the methods developed to find DIF in dichotomous items have generalizations that can be used to find DIF in polytomous items. So, in practice, looking at polytomous items for DIF is very similar to looking at dichotomous items for DIF.

Concept of Achievement Test

An achievement test is designed to measure how much knowledge a person has gained through teaching in a particular subject. Ali (2006) viewed performance tests as a management tool for individuals or groups, as incentives to encourage specific desired responses on behalf of individual or group ability. In order for the obtained data to be considered valid, each tool (ie., the test) must have certain characteristics (Ezeh & Onah, 2005). A measurement instrument must meet the requirements of reliability, validity and fairness (Anene & Ndubuisi, 2003).

There are two kinds of achievement tests: those made by the teacher and standardised ones. Exams made by teachers are called "teacher-created" exams (Onunkwo, 2002). They are tests that teachers make in their own schools to see how their students are doing. Ifeakor (2011) says that a standardised test is one that has rules. Norms are a set of descriptive data that can be used to figure out how a candidate compares to a certain reference group. Standardized tests have a set of questions, instructions, and ways to give the tests that are all the same. Most cognitive achievement tests are either essay tests or tests with several choices. Onunkwo (2002) said that an essay test is a test that asks students to answer questions and gives them the chance to organise and write down their ideas.

This study focuses on multiple-choice which is an objective test, which can take two forms: the first can be a direct question that examinees are asked to answer, and the second can be an incomplete question that examinees are asked to finish (UI Hassan, & Miller, 2019). No matter how it's set up, every multiple-choice question has two parts: the stem and the options (i.e., the answer options). The stem is the direct question or part of the question that isn't clear. The alternatives are the choices that test-takers are told to choose the most correct answer from.

Procedures for Development of an Achievement Test

There are numerous steps involved in creating the test (Ivanova, 2014; Şahin, Yildirim, & Öztürk, 2022). These include: analysis of content, the test creator should have a detailed description of the topic or content for which the test is designed. Content analysis requires the test developer to examine the primary content on which the test is based in order to determine what the content is (Anene & Ndubisi, 2003).

Instructional objectives are analyzed as the second step of test development. According to Anene and Ndubisi (2003), instructional goals are desired changes in student behaviour in a given subject. Because these are the characteristics of the subjects he or she must assess, the test developer must establish the instructional objectives. The next step, according to Okolo (2006), is to create a test specification table that will serve as a guide. It specifies the number of items that must be developed for each topic and cognitive domain, from the simplest to the most difficult. Table 4 shows the distribution of an achievement test based on levels of cognitive thinking skills.

Using the preceding example in Table 4 as a guide, determine the number of items per cell and use this information to create your test items. This guide ensures a balance between content delivered and content assessed during item creation.

Table 4- *Sample of Table of Specification for a 40-Item Achievement Test*

	Know.	Comp.	App.	Analy.	Synth.	Eval.	Total
Content	10%	30%	25%	20%	10%	5%	100%
Topic A (25%)	1	3	2	2	1	1	10
Topic B (25%)	1	3	3	2	1	0	10
Topic C (20%)	1	2	2	2	1	0	8
Topic D (30%)	1	4	3	2	1	1	12
Total	4	12	10	8	4	2	40

Note:
Know. – Knowledge
Comp. – Comprehension
App. – Application
Analy. – Analysis
Synth. – Synthesis
Eval. – Evaluation

Face validation is what a scale looks like it measures based on its different parts. Polit and Hungler (2002) say that face validity is the process of sending scale items to experts in the field for feedback. Review of Item: Anene and Ndubisi (2003) say that item review involves a close look at each test item that has been written and choosing the best ones, so that only those that have stood up to the scrutiny are used in the trial testing. For trial testing, the validated test is given to a large, representative sample of the students it was made for (Anastasi & Urbina 2002). From the above, it's clear that the narrative trial is important because it shows whether a candidate's response pattern matches what subject experts expect.

The researcher thinks that item analysis is the last step in making a test. Anene and Ndubisi (2003) say that item analysis is the process of looking at how people answered each test question. Those that pass the statistical analysis are used for the final version of the test, while those that don't pass are out CTT and IRT, depending on the measurement theory that is being used. In both frameworks, a well-designed test had to have item parameters that were in line with the theoretical scale for selecting items using test theories (Oguguo & Lotobi, 2019).

Concept of Gender

UNESCO (1999) defines gender as the socially determined characteristics of men and women that are always differentiated from those that are generally or biologically determined (sex). Therefore, gender consists of socially constructed distinctions between males and females. There are some activities or tasks that male students do better than female students, and the same is true for the other way around. The implication is that culture and society decide what gender is based on what men and women think, feel, value, and expect from each other.

Mills (2017) says that gender is an analytical concept that looks at how men and women (masculine and feminine) behave in a society. Performance differences between men and women will always be different from one society to the next. Bland (2013) says that gender is a set of differences between men and women, especially between men and women. Depending on the situation, these differences can be about anything from sex to social roles to gender identity.

As a core-subject, Agricultural science is studied by both boys and girls without exception. Due to the practical nature of the subject, some activities may necessitate physical exertion that differs between female and male students. Male students are more likely to do well at activities or tasks that involve preparing seedbeds. On the other hand, female students are more likely to do well at activities that happen after harvest, like packaging, marketing, and keeping records.

Even though the government and non-government sectors are working hard to make sure that boys and girls get the same education, there is still a big difference between how well boys and girls do in school and what they learn. Botswana only has a small number of studies that look at how the academic performance of boys and girls differs. These limited studies focused rather on the overall performance difference instead of how individual items in a test contributed to such differences.

According to UNDP (2016), in Botswana majority (57%) of women compared to men form the labour force in crop production. Their engagement in crop production (field crop and horticulture). This engagement covers planting/broadcasting seeds, weeding, harvesting, processing and post-harvest handling, while men tend to be responsible for land clearing and ploughing with cattle. The report further posits that for every 100 men who own livestock, there are only 39 women in Botswana.

However, based on the researcher's observations, gender differences in agricultural performance in Botswana are most likely the result of socialisation. For instance, women are primarily responsible for the entire cycle of crop production, small-stock and poultry, gathering firewood, veldt products for domestic consumption, and water collection (FAO, 2018). In the crop value chain, women are responsible for planting/disseminating seeds, weeding, harvesting, processing, and post-harvest handling, whereas men are typically responsible for clearing land and ploughing with cattle. Men dominate the primary agriculture production activity, which is primarily livestock farming (beef cattle production).

The impact of gender-based agriculture activities in society on teaching and learning Agriculture science in schools cannot be underestimated. Due to cultural expectations and socialization, girls and boys tend to imitate their mothers' and fathers' activities. Williams, Walaver, and Duggal (cited in Ihechu, & Madu, 2016), assert that differences in subject areas such as Agricultural Science and Economics may be attributable to social and cultural influences that generate stereotypes. These stereotypes may reduce interest in subjects dominated by women or men, including Home Economics and Agriculture science, respectively.

The researcher has observed that in Botswana boys usually join their fathers at the cattle post (*meraka*) during school vacation while girls also join their mothers at the crop fields (*masimo*). During such times, the students are likely to assist their parents in farming activities. These may be management activities like branding, castration, watering of livestock for boys and weeding, winnowing, and harvesting for girls. The experiential learning taking place is likely to be taken back to the classroom and reinforce understanding of certain topics. Moyo (2015), opines that through early-age agricultural activities interaction, boys gain exposure more than their female counterparts.

The JCE results for the time period this study covers show that there is a big difference between how well boys and girls do in agricultural science, with girls doing better than boys. This means that students' scores on Agricultural examinations in junior secondary school may be heavily influenced by their group membership rather than their ability. This view supports Moyo's (2015) assertion that in Botswana, boys perform better in agriculture as a farming practice than as a subject. Girls are more likely to be involved in crop-related farming activities than men, so tests with more items are likely to account for the observed difference.

Concept of Location - Urban/Rural

School location refers to a school site, type of buildings, usage, capacity, teachers, students, and other parameters for rationalization of rural and urban school operations (World Bank Guidelines, cited in Umar & Samuel, 2018). The location of a school has a big effect on the facilities that are available, the number of teachers, and the number of students. Junior secondary schools are spread all over the breadth and width of the country.

Notwithstanding, most teachers favour urban schools with social amenities to the detriment of rural schools with low populations and subsistence livelihoods (Pérez, 2020). In the context of Agriculture science, where large school sizes are required to house school farms for practical sessions, urban schools are likely to be disadvantaged. This is because the researcher has observed that in some instances urban schools must sacrifice their school farms to construct new classrooms to accommodate the continuously growing urban population.

According to Statistics Botswana (2011) census report, the standard definition of a modern rural area in Botswana is a location with access to a road network, at least a primary and secondary school, a health facility, a kgotla, a mobile phone network, and sometimes the national grid. However, one of the defining characteristics of rural areas is that they are administered by a chief (*Kgosi*). Historically, and in some cases still today, rural areas were dependent on agriculture and natural resources (Tisdell & Moepeng, 2010). However, in rural villages agriculture farming activities are predominately at the subsistence level as prescribed by Town and Planning Act.

According to Towns and Planning Act, Botswana is divided into three strata, namely residential (villages and towns), ploughing fields (*masimo*), and cattle posts *(meraka*). However, unlike in towns, livestock farming (cattle, goats, sheep) is allowed in rural areas. Notwithstanding, backyard farming (keeping small non-ruminants animals such as chickens and rabbits) is allowed in towns. Smallholder farming is still the most common way to make a living in rural areas. It provides a lot of jobs, food, and money, especially for women and the poor (United Nations Development Programme, 2016). This allows students in those areas the exposure and involvement in farming activities.

Studies on location and differences in academics are inconclusive. Several studies reported significant differences in performance by location (Moyo, 2015; Ihechu, & Madu, 2016; Amao, et al., 2016; Annan-Brew, 2020). However, some studies reported no significant difference in academic performance by region (Ogunyemi, 2000; Gana, 2007; Ajayi, 2009; Bulala, Rmatlala & Nenty, 2014).

Concept of School type

Although the majority of basic schools in Botswana are government-owned (public schools), there is a noticeable increase in the number of private schools in Botswana. The government owns, runs, and pays for public schools, while private schools are owned, run, and paid for by private people, parent's groups, non-profit organizations, and/or religious institutions (Thapa, 2013).

Several researchers have pointed out that there are academic disparities between private and public schools (Anigbo 2006; Gbadamosi 2007; Dixit, 2019; Kunwar, 2021). This observed disparity is also reported in a study by UNESCO-IBE (2010) on world education data in Botswana. Concerning Agriculture science (Ihechu, & Madu, 2016), reported a significant difference in academic performance in National Examinations Council (NECO) between private and public school students in Agriculture.

Various factors have been attributed to the academic disparity between public and private schools (Mathema, 2007; Sharma, 2012; Thapa, 2012). For example, Thapa (2012) cites resource disparities between private and public schools as contributing factors. This narrative is supported by several authors

(Subedi, Shrestha, & Suvedi, 2014; Bonsu, 2016), who observed that due to a lack of resources like classrooms in public schools due to high enrolments. Notwithstanding, teaching and learning resources are varied and go beyond just classrooms, for example, qualified teachers, textbooks, laboratory, and revision papers are also critical in the quality of education.

This reported high number of students per class is most likely to limit student-teacher interaction and individual assistance. This will be an even bigger impediment in subjects like agriculture which requires small and manageable groups for practical exercises which are necessary for reinforcing learning. The limited interaction is likely to compromise exposure during school-based assessment and preparations for the high-stake examination.

However, for Agriculture science teaching resources go beyond just the classroom. School farms are limited and, in some cases, non-existent in private schools compared to public schools due to lack of sufficient space. Notwithstanding, agriculture farms in schools influence academic performance (Machisu, Opondo, Nakhumicha, & Mosi, 2022). Furthermore, most of these schools cite a shortage of appropriate tools and equipment (Baliyan, Malebalwa, Keregero, & Mabusa, 2021). Due to these challenges, most private schools particularly those in urban areas, opt for taking the written practical test in replacement of school practical assessment.

Empirical Review

Characteristics of JCE Agriculture Science Multiple Choice Items based on 3-PLM

Orangi and Dorani (2010) conducted research to develop a social studies achievement test for high school students based on item-response theory (IRT). The purpose of the study was to develop a social studies achievement test for high school students (first grade) based on item response theory. The sample consisted of 321 high school students in Tehran. Multi-stage cluster sampling was used for selecting the participants. The study adopted an instrumentation research design. The first step in conducting this exam was to prepare two parallel forms of multiple-choice tests in which on one hand concentrated on the educational objectives and on the other hand on the content of the lessons.

In the first stage, the questions were looked over for any ambiguity in their wording and the students' ability to understand expressions was tested, among other things. In the second stage, which was the practical stage, the level of difficulty of the questions, the students' ability to recognize questions, and the level of interdependence of questions with the overall score were determined. Ten days passed between giving each form to the sample group. The results show that the forms that were made were very trustworthy. Considering the Item Characteristic Curve, both forms gave students with average skills the same amount of knowledge. In this analysis, a sort of rank-percentile norm was made for each sex. This study is similar to the current one in some ways. For example, the design was the same, but the sampling method and size of the samples were different. The small sample size makes it hard to trust the calculated item parameters. Also, the study was mainly about Social Studies, not agriculture science.

In a study by Oliveri and Ercikan (2011), they looked at how similar the English and French versions of the Programme for International Student Assessment (PISA) were in terms of the way they were put together and where they were different. Several methods were used to look at test characteristics and item-levels, including a look at the structure of the test data, comparisons of reliability, and test characteristic curves (differential item functioning, item parameter correlations, and linguistic comparisons). About 28,000 15-year-olds from more than a thousand schools in Canada took part in PISA 2003. This study's total sample size was 2,156 for Booklet Set 1 and 2,232 for Booklet Set 2, and no data was missing. The test item difficulty level analyses show that the two language versions of PISA are very similar. This is shown by the similarity of the internal consistency coefficients, test data structure (same number of factors and item factor loadings), and test characteristic curves for the two language versions. But the results of the test item difficulty level show that there are a number of differences between the two language versions. This is shown by the fact that a large number of items work differently, that item parameter correlations (discrimination parameters) are different, and that a number of items have linguistic differences.

The current study adds to the long list of studies which have focused on core subjects such as English, Maths General Science etc. However, like the current study the study used a large sample size. Furthermore, like the current study it also looked at item parameter estimates and item characteristic curves in relation to contribution to student achievement.

In another study, Yoon (2011) looked at the Psychometric properties of the revised Purdue spatial visualization test. In this study, 1022 undergraduate students from all majors at Purdue University in the United States of America were used to test the psychometric properties of the new version. Within the framework of CTT, Cronbach's alpha showed that the items in the measure were consistent, and a confirmatory factor analysis (CFA) showed that the Revised PSVT:R measure of the construct had a unidimensional factor structure. This study looked at four IRT models: the Rasch, 1PL, 2PL, and 3PL models. The 3PL IRT model gave the best fit between the model and the data.

Overall, the results showed that the difficulty and difference between items were within a good range. But the items that were put in order by how hard they were to rotate were not in order of how hard they were, which suggests that there are other things, like the difficulty of the 3-D shapes, that affect how hard people think an item is. The 3PL IRT model says that all 30 of the Revised PSVT: R items have a certain amount of guessing effect. Even though this study and this one is about different subjects, they both used the 3PL model to figure out the item difficulty, guessing level, and discrimination parameters. The limitation with this study is that it used a relatively small sample, and its target group are university level respondents. These two disparities may limit the comparability of the findings.

Obinne did a study in 2008 using the Item Response Theory to look at the psychometric properties of the items on the National Examination Council (NECO) and the West African Examination Council (WAEC) biology tests (IRT). The study used a method called "instrumentation." There were made, tested, and analysed research questions and hypotheses. The sample included 1800 students in their third year of high school from 36 secondary schools in both urban and rural areas of Benue State. The stratified multistage sampling method was used. From 2000 to 2002, the questions on the NECO and WAEC Biology exams were used to gather data. IRT procedures called for the maximum likelihood estimation technique (using the computer program BILOG MG) to be used to look at the research questions. The hypotheses were tested with the t-test.

It was found that the biology test questions from the two different testing bodies were both valid and reliable. The biology questions on the NECO exam in 2001 were harder than those on the WAEC exam of the same year. WEAC questions were easier to guess than NECO questions. It was suggested that all examination bodies in Nigeria use IRT procedures so that our measurement problems can be solved. This study is like the current study used high stake items administered by examination body comparably to BEC. This is important as parameter estimates are made from standardised items. It also covered a three-year period. But the study was about biology and used a small number of students as samples compared to the current study which used agriculture items with a large sample size.

Odili (2003) undertook a study on the effect of language manipulation on DIF of Biology multiple-choice test. The instruments for data collection were four namely: WAEC/SSCE biology paper 2 1999, 2000, and 2001 made up of 60 items each. Differential functioning test items used in the original language (form A) made up of 30 items; Differential test items with simplified non-technical words used (form B) made up of 30 items; Questionnaire on student's background (SES) made up of 20 items. The sample of the study was made up of 3300 senior secondary three students (male 1762, female 1538; urban 1980, rural 1320; high SES 638, low SES 2662; experimental group 512, control group 513). The DIF detection method used was the scheuneman's modified chi-square. However, he used the dependent t-test and chi-square to test the significant difference existing between the two groups in the experimental study.

The result revealed that WAEC/SSCE biology paper 2 for 1990, 2000, and 2001 contains item with significant location, gender, and socio-economic status DIF, with location having more DIF items. In addition, the manipulation of differential functioning test questions did significantly reduce DIF for the test takers. The researcher used scheuneman's modified chi-square, he did not make use of the IRT based DIF method or the purely CTT based DIF method. This could be so because it was not the focus of his study. He however, used gender, SES, and location in his study.

Orheruata and Uyigue (2018) investigated flawed items of the West African Senior School Certificate Examination (WASSCE) Agricultural Science multiple choice items across 2012 to 2014 to determine the level of flaw in the item parameters across the stated examination years using Item Response Theory (IRT). Survey research design that adopted multistage sampling technique was used in selecting a sample of 3,744 of senior secondary three (SS3) Agricultural Science students from Edo South Senatorial District. The instruments used were 2012 to 2014 WASSCE Agricultural Science multiple choice test items. The instruments (test items) were assumed to be valid and reliable by the nature of standardised instruments administered by WAEC.

The items were calibrated using EIRT computer software programmes to determine item difficulty (b), item discrimination (a) and guessing (c) parameter estimates. From these estimates, items with parameter values not within the IRT theoretical scale were flagged as flawed items for analysis. The results established that the condition of the items with flawed difficulty parameter estimates showed a percentage of 88.2, 50 and 71.4 pointed easier across 2012, 2013 and 2014, respectively. This study focused on agriculture science multiple choice items and just as the current study it also used IRT to determine parameter estimates. However, the sampling technique and subsequent sample size are different compared to the current study. The analysis was conducted on the WASSCE Agricultural Science multiple choice test items which give little information on the JCE agriculture science as curricula and context may be different.

Nkpone did a study in 2001 on how latent trait models were used to create and standardize physics achievement tests for senior high schools in Nigeria. The one-parameter logistic model and 359 students from high schools for seniors were used in the study to get estimates of the item parameter. The results showed that the items ranged from -1.49 to 0.49 in terms of how hard they were. The estimated value matches how the Physics Achievement Test (PAT) questions were written to get harder in each content area. About 22 of the 60 questions were easy, with a difficulty level of less than 0. About 37 questions were hard and had a level of more than 0. The mean estimate of difficulty was zero, and the standard deviation was 0.31, which suggests that there wasn't much difference in how people scored. The 2pl model was also used to estimate item parameters.

The results showed that the items were not too hard and had consistent discrimination indices between 1.76 and 0.39. The item's difficulty index was between 1.66 and 0.69. For the latent trait 2PL model, items with a discriminating index of more than 0.8 and an item difficulty index with a rectangular distribution from -2.0 to 2.0 have a good discriminating power. The study also tried to figure out how much each of the PAT items was off by on average. It was found that the standard errors range from 0.0578 to 0.0518, with a mean of 0.17. This is equivalent to 17 percent of the total unreliability of the variance, while 83 percent is due to the reliability of the true variance. The range of standard error values was less than 10% of the largest standard error. This means that the estimates of the difficulty indices are very accurate. This study has to do with the current one, especially in the use of IRT model. Notwithstanding, in its use of 2PL it treated guessing parameter to be constant and as such doesn't give insight into guessing which is common in multiple choice. Compared to this study, the sample size of that study was also very small.

Akanwa, Ihechu, and Nkwocha (2020) studied the effect of language manipulation on different item functioning of the Nigerian Agricultural Science multiple choice test items used by the National Examination Council in the Senior School Certificate Examination (SSCE) from 2013 to 2015. A quasi-experimental design was used with a control group that only took the test. The population was made up of 3,238 students in Imo State who were in their third year of high school and took Agricultural Science as a certificate subject. For the study, three-stage proportionate stratified random sampling was used to pick 100 students at random.

The study showed that the SSCE 2013-2015 Agricultural Science multiple-choice questions used by NECO have test items that work very differently for students based on their socioeconomic status, where they live, and their gender. It was also found that when the language of Agricultural

Science test items was made simpler, there was a big drop in how differently different items worked. This study gives us important information about DIF Agricultural Science.

There are key similarities to this study and current study; first both use agriculture multiple choice items, secondly, they all looked at secondary school examinees and thirdly they covered a three period. This makes the two studies closely comparably. However, the study looked at manipulation of language and its influence on students response while the current study purely focused agriculture achievement test. Therefore the two studies covered to different constructs. Furthermore the sample used is relatively small compared to current study.

Butakor (2022) investigated how well a math test item made by a teacher and used in one of Ghana's Senior High Schools measured up. This study used a quantitative descriptive design in which the answers of 400 randomly chosen students to a math test made by their teacher were collected and analyzed using different psychometric techniques.

The results showed that the Math test had a low but acceptable reliability coefficient of 0.61. Also, out of the 40 multiple-choice questions, only one was too hard and three were too easy. This means that 26 of the questions were just right. Based on the results of the discrimination indices, 25 of the test questions had bad or weak discrimination indices, and 4 of the questions had negative discrimination indices. The study looked at math tests for teachers and had a small sample size. The current study, on the other hand, looked at high-stakes, standardised agriculture science items and had a very large sample size. But some testing organizations, like BEC, use teachers as part of their item-writing teams. This makes it possible to compare the two studies.

Adedoyin and Mokobi (2013) did a psychometric analysis of the 2010 Botswana mathematics JC to figure out how good the multiple-choice test items were for the junior certificate mathematics exam. The first math paper had forty (40) multiple-choice questions that were based on the three-year JCE math curriculum. All 36,940 students who took the JC mathematics exam in 2010 were included in the study. SPSS software was used to pick a random sample of 10,000 students from this group. IRT (3PL) model was used to look at the psychometric parameter estimates of the forty test items, which were the item difficulty, item discrimination, and guessing value.

For each test item that fit the IRT (3PL) model, the item characteristics curves were also made. Out of the forty items, only 23 fit the 3PLM. These 23 items were used to test the JC mathematics test paper 1's psychometric properties. Based on the results of this study, twelve (12) of the twenty-three (23) items that fit the IRT model were rated as poor test items, ten (10) were rated as fairly good test items that could be revised or improved, and one (1) was rated as a good test item. But this study was done with math test questions, which shows that agriculture science hasn't been given as much attention as it should. Agriculture tests need to be taken into account as well. The study also looked at how hard the items were, how different they were from each other, and how much guessing was worth. Also, the study was done with test scores from only one year (2010). This study looks at tests in 2018, 2019, and 2020, which are three consecutive years.

Motshabi and Nenty (2012) did an ethnicity-related differential item functioning study of English language test questions in primary schools in Botswana. The study looked at how primary school students in Botswana answered questions on the 2008 Primary School Leaving Examination in English language. It did this by using three different ways to find items. Purposive sampling was used to choose 2,587 of the 41,471 students who took the exam in 2008 so that the study subjects would be a good mix of different races.

The Mantel-Haenszel, standardization, and maximum likelihood item analysis methods were used to look at both the primary and secondary data collected from and about these subjects. This was done to see if the way the items worked were similar across the three ethnic groups. About 48 percent of the 60 items on the test worked differently for learners from different ethnic groups, even though they all had the same level of English language skill. It was thought that this would have big effects on how fair cross-cultural assessment, teaching, and learning were in Botswana. Even though this study had a small sample size, it used a CTT-based DIF detection method and showed how common DIF is in BEC exams.

Difficulty parameter, discrimination parameter and level of guessing are some major characteristics of a well-structured exams. Empirical studies have shown that almost all test items have some level of discrimination index, difficulty index as well as some level of guessing. Indication that, an acceptable level of index is usually recorded among the various text characteristics. Some studies also indicates that, item difficulty and item discrimination were within an acceptable range. Confirming the fact that, any well-structured exams will have it characteristics parameters well within range. Though several studies have been conducted on difficulty parameter, discrimination parameter and level of guessing, several of these studies did not focus on Agriculture science. Also, most of these studies used a small sample size. In addition, very few studies have been found in Botswana, focusing on Mathematics and English. The current study differs from these studies due to its focus on Agriculture Science and its use of a large sample size in Botswana.

Contribution of JCE Agriculture Science Multiple Choice Items to the Measure of Students' Achievement in the Subject

Rivera, (2011) worked on test item construction and validation and his impetus for this study was addressing the need of one secondary career and technical education program, agricultural science education, which does not have a state-wide exam in USA. His study was mostly about the multiple-choice written part of the state-wide exam. More specifically, it was about two parts of making an exam: making items and making sure they are correct. Animal systems and plant systems, two of the nine parts of the test, were made using criteria-based test construction methods.

The results of this study describe a process for developing and validating items. They talk about some of the pros and cons of making test items for a diverse audience without the help of a test validation. More thought is given to the procedures used to validate test items, especially expert judgment and analytical data.

The study like the current study looked multiple choice agriculture science items. The study used item information function and item characteristic curves to validate item contribution to observed achievement which the current study also did. However, due to the geographical differences from where the studies were conducted, they maybe contextual differences. The study clearly revealed that, validated test avoids high standard errors in test construction and ensures high reliability. The results from this study provided guidance on the identification on measuring item contribution which the current study required.

Shogbesan (2017) did a study to find out how well the tests used by teachers in secondary schools measure what they are supposed to measure. He also looked at how the format of test items affected the construct validity of achievement tests in Nigerian high schools in the state of Osun. The study used a descriptive survey research design to do this. Multistage sampling was used to choose 300 students from Senior Secondary School II and 36 teachers from secondary schools at random. The Achievement Test (AT) and a score record sheet were used as research tools in the study. The AT was used to collect data that was used to figure out the construct validity of the achievement test. The test was validated by using the test blueprint to check for content validity and the Cronbach's alpha reliability, which gave a coefficient value of 0.68. Principal component analysis, scree plot, and one-way analysis of variance were used to look at the collected data.

The results showed that achievement tests made by subject-specialist teachers and used to evaluate students in the classroom have construct validity (r=0.414, df=298, p0.05) and that the format of test items has a big effect on the construct validity of the Achievement Test in secondary schools in Osun state. Lastly, there is a significant difference (F=290.25, p0.05) in how people do on the Achievement Test when it is given in different ways. The study came to the conclusion that the format of test items affects the construct validity of the Achievement Test in secondary schools in Osun state. The study also suggested that people who make tests and teachers in the classroom should understand the characteristics of each format and choose the format that best fits the purpose of a test in each situation.

The weakness with the approach of this study was it used test information instead of item level information. It is obvious that different items contribute differently to how valid and reliable the instrument is. This flaw challenges the conclusion the author test reliability and validity. Perhaps, maybe a different approach may have given individual items that contributed very little to the observed overall reliability coefficient. However, this study like the current study used descriptive design and also used large sample size.

Osadebe (2015) did a study to make a valid and reliable Economics test for Nigerian high school students. Two research questions were made to help find out if the Economics Achievement Test is valid and reliable (EAT). It is an objective test with five choices and 100 questions. A sample of 1,000 students was chosen at random to figure out how valid and reliable the test was. After analyzing the items, the results showed that students did well on the Economics achievement test and that the test had both high face validity and high content validity. The item difficulty and discrimination indices were used to figure out the test item's validity. After using the formula to correct for guessing, a "difficult index" or "p-value" of 0.5 was used for each item. Point biserial statistics were used for each item with a correction coefficient of at least 0.3 to figure out the discrimination index. Using the Kuder-Richardson formula 20, we found that the test has a reliability coefficient of 0.95.

This study looked at validity and reliability through the CTT and used composite test scores instead of individual test items, which the current study looked at through the items' contribution to achievement construct. Compared to the current study, the study had a small number of people to study. But both looked at multiple-choice questions, even though they were about different subjects.

In 2014, Osadebe did research on the Standardised subject Achievement Test for students in Nigeria's senior secondary schools. The study was based on three research questions. An expert first made sure that the standardised test in each subject was a valid and reliable tool. In this study, the test was then used to make sure everything was the same. That is, making sure that the tests for each subject are all the same. It was given to 3,000 students following the same rules, and there was no cheating. For standardization, the sex, location, and type of school of most students were considered. Using the normal curve as a theoretical foundation, the raw scores of the students were normalized by using the students' percentile rank, Z-score, T-score, and Stanine statistics as derived scores. This was the last step in turning the test into a Standardized Achievement Test. With the help of Percentile rank, Z-score, T-score, and Stanine, the results show that the test scores were spread out in a normal way. There were suggestions that making sure a standardised test is valid and reliable could be used to measure and compare students from year to year.

In constructing a test how the test may contribute to the measure of students' achievement in the subject is very important. However how valid the test items have a higher probability of predicting student achievement in the test. In the works just looked at, it was clear that an achievement test made by subject-specialist teachers and used to evaluate students in the classroom has construct validity and that the format of the test items has a big effect on the construct validity of the achievement test. For this reason, several considerations has been given to procedures used to validate test items, specifically expert judgment and analytical data. This is because, it was clearly revealed that, validated test avoids errors in test construction and ensures high achievement. However, the various forms of test validity were not given the necessary attention as a contributing factor in students' achievement in test in Agriculture Science in Botswana.

DIF of JCE Agriculture Science Multiple Choice Items in terms of Sex

Scholars, policymakers, and practitioners have all noticed that there are socially constructed differences between men and women that have big effects on their lives, and they all seem to agree on this. There is a big difference between how well boys and girls do in school, according to studies of students at different levels all over the world. Several studies have shown that girls do better in school than boys (Orabi, 2007; Dayioglu & Turut, 2007; Khwaileh & Zaza, 2010). Ghazvini and Khajehpour (2011) also said that there are differences between men and women in how well they think in an academic setting. Girls are more likely to be able to learn in a different setting than boys. Mwingi's (2014) study of students in secondary schools in Kenya, though, showed that more boys than girls passed. On the other hand, Goni et al (2015) in a study of college students didn't find a big difference between how well men and women did in school.

Huang, Wu, and Mok did a study in 2022 to look at how different items work for men and women using Poly-BW Indices. Huang, Wu, and Mok said that the current methods for detecting differential item functioning (DIF) based on item difficulty or item discrimination aren't good enough for figuring out how DIF items are related to each other. Because of this, DIF items were usually either deleted or ignored. Given how important it is to make tests with as few DIF items as possible, teachers or testing professionals need more information about what DIF items might be linked to. The goal of this study was to look at how the Poly-BW indices (discrimination, difficulty, and guessing) affected the properties of gender-related DIF items using a teacher-made math test as an example. Data from a 34-item mathematics achievement test with 1,439 seventh-grade students from Taiwan (51.01 percent boys and 48.99 percent girls) showed that the differences in the difficulty and discrimination indices between men and women were good predictors of the DIF measures estimated by the Poly Simultaneous Item Bias Test (Poly-SIBTEST) procedure with satisfactory hit rates. This study used the SIBTEST method to find DIF, while this study used the Raju are method. Agriculture is an applied science, and there are parts of it that require math skills. This makes this study relevant to the current study.

Abedalaziz, Leng, and Alahmadi (2018) used transformed item difficulty to find a gender-related difference in how items worked. The goal of the study was to see if there were any differences in how men and women did on a multiple-choice math test that was part of a high school graduation test made to match the curriculum of the eleventh grade. A gender-related DIF was found using the transformed item difficulty (TID). 1400 11th graders in the Malaysian city of Kuala Lumpur were chosen at random. In DIF indices, females had a statistically significant and consistent advantage over males on algebra questions, while males had a less consistent advantage on geometry and measurement, number and computation, data analysis, and proportional reasoning questions.

This study like the current study focuses on multiple choice items at high school level to detect DIF. The study used transformed item difficulty while the current study used Z-statistic and associated P-values to detect DIF. The use of item difficulty alone may miss items whose DIF is tied to guessing or item discrimination indices. Furthermore, the study looked at mathematics instead of agriculture which measures a different construct to agriculture ability. In any case agriculture is an applied may have items requiring mathematical computations.

Robin, Zenisky, and Hambleton (2003) did a study to find gender DIF in a large-scale science assessment and look for trends in the DIF and non-DIF items due to content, cognitive demands, item type, item text, and visual-spatial/reference factors. DIF analyses were done on the item-level answers of more than 60,000 students who took part in a large-scale state assessment program in language arts (LA), math (MA), and science (SCI). In the United States, standardized tests were given in each subject at the lower school (LS), middle school (MS), and high school (HS) levels. The DIF study was done at three grade levels and with two randomly chosen versions of the science test at each grade level. This made it easier to do the analyses (carried out in different years). A version of the standardisation procedure was used on six sets of data that each had data on 60,000 students. This version of the procedure is easy for practitioners to understand and explain.

From the study, a number of useful things were learned that could be shared with committees that make tests. As an example, when there is a DIF in science questions, multiple-choice questions tend to favour men, while open-ended questions tend to favour women. By putting together DIF data from multiple grades and years, it is more likely that important patterns will be found in the data and that writing about DIF will be based on more than just anecdotal reports. Like the current study, it used a large sample size, covered more than sets of data. This study also looked at multiple-choice questions in agriculture, which is another applied science.

Amaechi and Onah (2020) administered an Economics multiple-choice standardized test in Nigeria as part of a DIF study. Using simple and purposeful random sampling, the study's population consisted of 4,434,979 secondary school students. For data collection, they utilized Socio-Demographic Inventory (SDI) instruments and a 50-item WAEC General Economics Paper I Multiple Choice Exam. The IRT-Binary Logistic Regression method was utilized, and hypotheses were tested using the Wald test in conjunction with binary logistic regression statistics at a significance level of 0.05. They reported that, out of the thirteen items with DIF issues, eight have uniform DIF and five have non-uniform DIF. This study, like the present one, focused on standardised multiple-choice assessments at the secondary school level. Despite the relatively small sample size, the study met the minimum sample size requirements.

A study was done in the Nigerian state of Enugu by Ikeh, Ene, Ojobo, Ani, Metu, Ugwu, Owolawi, Omosowon, Oguguo, Ezugwe, and Agugoesi in the year 2021. Differential Item Functioning (DIF) was used in this study to find questions on the high school certificate exam about economics that were biased against women. A causal comparative or Ex-post factor research design was used for the study. The people who were part of the study were 2,985 SS3 students who were majoring in economics. The study used 339 SS3 Economics students as its sample size. The 50-item 2018 SSCE multiple-choice Economics test made by the West African Examination Council was used as the research tool (WAEC). A coefficient of reliability of 0.87 was found by using the Kudder-Richardson formula. Logistic regression was used to look at the study's collected data.

The study found that out of the 50 questions on the 2018 WAEC Economics exam, 14 questions, or 28%, showed significant gender DIF at the 0.05 level of significance. Only one of the 14 items that had significant gender DIF in favour of male students (2%) was found, while 13 items (26%) worked differently in favour of female students. Even though the study was mostly about economics, it also talked about high-stakes tests in high school. It also put a lot of emphasis on questions with more than one answer. Even so, the number of people in the sample was much smaller than in this study.

Ekong, Ubi, and Eni (2020) looked into the differential item functioning (DIF) of the 2018 Basic Education Certificate Examination (BECE) in Mathematics tests given by the National Examination Council (NECO) and the Akwa Ibom State government in Nigeria. Item Response Theory (IRT) was used to figure out if the tests were the same for both men and women. There were 58,281 students in the study population, and a sample of 3,810 students was chosen using a multistage sampling method. The multidimensional IRT (MIRT) package in the R programming language was used to look at the data. In terms of sex, BECE of NECO showed 23 DIF items, or 38.3%, while BECE of Akwa Ibom State showed 37 DIF items, or 61.7%. The results also showed that, in terms of performance, more questions on the two tests were easier for men than for women. It was suggested that test makers use the IRT model to figure out the parameters for selecting items to make sure of their quality before giving the test.

Though the study, like most of those reviewed focused on mathematics, it has two main features related to the current study. For example, the instruments used are high stake national examinations and relatively large samples. Furthermore, like the current the study used R programing language for analyses.

Omorogiuwa and Iro-Aghedo (2016) used a survey research design to investigate how different items worked for men and women on the 2015 Mathematics Multiple Choice Test Items (Dichotomous) Examination given by the National Business and Technical Examinations Board (NABTEB) in Nigeria. This was done by figuring out which parts of the test gave different results for men and women. Out of the 63,584 responses from examinees in Nigeria's six geo-political zones, a sample of 17,815 responses from two states in each geo-political zone was chosen. This included 11,873 men and 5,944 women who ran for office. A 50-item math multiple-choice test was used to gather information. The Area Index (Raju) method, which is one of the DIF detection methods based on item response theory, was used to find questions that were answered differently by men and women. On 17 questions (34%), male and female test-takers did not do the same, but on the other 33 questions, there was no difference (66 percent). Six of the 17 things were good for male students, and eleven were good for female students.

The study used a large sample size that increases the possibility of accurately detecting the presence of DIF. Furthermore, this study used the Area Index (Raju) method. However, this study focused on mathematics subject while the current one focused on agricultural science.

In a regional study, Woitschach, Zumbo, and Fernández-Alonso (2019) conducted a comparative and interpretive study to compare mathematics, science, and reading educational outcomes in 15 Latin American countries. This research focuses on the ecological perspective, which encompasses the examined individual, process, context, and time. These descriptions talk about certain situations, show how and what has changed at the individual, school, and national levels, and give information about how to respond. The goal is to find out why different items work differently in different environments. The study looked at the sixth-grade science tests that 12,657 students from 2,609 schools in 15 countries took in 2013. In the multivariate Bernoulli logistic regression model, the variance of the distribution was worked out step by step.

The approach by Woitschach, Zumbo, and Fernández-Alonso (2019) has a weakness in that it reports the presence of DIF across countries. The study does not specify how the reported 32 percent is distributed across countries. Nonetheless, this study demonstrates that DIF is more influenced by context. As agriculture science is linked to cultural farming practices, this makes the study important to the current study. Yao and Chen (2020) investigated whether there are gender differences in the test items of the General English Proficiency Test for Children (GEPT-Kids). A descriptive research design based on a two-stage mixed method was used. In the first phase, the performance data of 492 participants from five Chinese-speaking cities were analysed with the Mantel-Haenszel (MH) method to determine gender DIF. In the second stage, items showing DIF were subjected to content analysis by three experienced raters to identify sources of DIF. The results showed that 3 items with small gender DIF were detected by statistical methods and 3 items were identified as biased.

The flaw in this study is that it used CTT based DIF detection method which is amenable to a small sample size. The study does not describe in detail the nature or types of GEPT items used. Thus, making it difficult to know if they were dichotomous or polytomous. Furthermore, the study is based on English subject at the elementary school level. The current study, however, focuses on agricultural science which is science based. Notwithstanding, the study has one important feature in that it goes beyond mere identification of DIF items to expert content judgment.

Siamisang and Nenty (2012) looked at how items worked differently for girls and boys on the 2007 TIMSS test taken by students from Botswana, Singapore, and the United States. They used a quantitative method that included the Scheunemann Modified Chi-square and the Mantel Haenszel (M-H) differential item functioning analysis. Findings from Chi-square and (M-H analysis showed that there were only four differences between male and female students in mathematics and four differences between male and female students in integrated science across the three countries. But the DIF found did not work in a way that was significantly different between men and women from the countries. The M-H analysis showed that all of the items that were tested for gender DIF had either a small or no DIF.

The conclusion of the study by Siamisang and Nenty (2012) was based on how well the study's goals and research questions were met. The study's results showed that there were some differences between men and women, but the size of these differences was small across the three countries. But the DIF analysis between countries showed that the differences between Singapore and USA students were statistically significant but small, while the differences between Botswana and Singapore students and between Botswana and USA students were both statistically significant and quite large. This study used the CTT method Mantel Haenszel to find DIF, while this study used the IRT 3PL method. Both studies, though, used large samples and focused on multiple-choice questions, which makes them similar.

Kalaycioglu and Berberoglu did a study in 2011 to find items that worked differently for men and women. They looked at the content of the items to find possible sources of DIF. Finally, they looked at how DIF items affected the criterion-related validity of test scores in the quantitative section of the university entrance exam (UEE) in Turkey. The study looked at the DIF of items based on how they related to the subject matter, how they tested cognitive skills, and how they were formatted. It seemed that DIF in favor of male students was caused by higher-order cognitive skills and figural or graphical representations in the item content. DIF against males could be caused by routine algorithmic calculations. Out of all the things that were looked at, cognitive skills as measured by items seem to be the most likely to cause gender DIF. But the criterion-related validity of the quantitative part of the UEE was not at risk because of DIF items. Notably, Kalaycioglu and Berberoglu's (2011) study of the DIF items and gender differences showed that there are differences in how men and women choose items on measures that are used for the same purpose. This study gave information about DIF, but it didn't explain how DIF was analyzed or how many samples were used.

Annan-Brew (2020) did a study using the cross-sectional design on the West African Senior Secondary Certificate Examination (WASSCE) in core subjects to look into gender and location differential item functioning (DIF) in Ghana. Six research hypotheses and one research question were written for the study. From the 273,289 people who took the test between 2012 and 2016, a sample of 36,035 candidates was chosen. This sample was made up of 8,994 English Language, 8,935 Math, 9,089 Integrated Science, and 9,017 Social Studies candidates. There were two kinds of test items for each subject. We used MH, LR, and IRT DIF detection methods to find things that had DIF.

The results showed that there was a big difference in how the items worked for men and women. There was also a big difference in how things worked depending on where they were, since all three methods found things that worked differently in each of the five regions under study. When it came to finding items with DIF, Logistic regression, Mantel Haenszel, and 3PL Item Response Theory all agreed on a lot of things. Like the current study, this one looked at DIF over time over a number of years. This study is important to the current study because it compared indices from CTT and IRT-based methods, including 3PL, which was used in the current study. The study was relevant to the present because it used multiple-choice questions and a large sample size.

In a quantitative study, Adedoyin (2010) looked at math test questions from the Botswana Junior Certificate Examination to see if they were biased against men or women. From the 36,000 students who took the Botswana Junior Certificate exam, a random sample of 4000 students' answers to mathematics paper 1 were chosen. This was done to find test items that were biased against men or women. Two thousand were men and two thousand were women. There were 38 test questions on the test paper. The study used 3PL (Multilog software) item response theory (IRT) statistical analysis to find the test items that were biased toward one gender over the other. This made the ICC for both the male and female groups. The study compared the ICC curves for the male and female groups and found that 5 of the 16 test questions that fit the 3PLM statistical analysis were biased toward one gender or the other. Both the male and female item characteristics curves were different, which shows that these items were skewed toward a certain gender. The following eleven (11) things were not important out of the sixteen. But there were five things that were important. Using IRT methodology (ICC), it was found that the biased items on the 2004 Botswana mathematics paper 1 exam would have caused a difference in test scores between the male and female groups. The only goal of the study was to find biased math test questions based on gender. The study didn't focus on Agriculture Science, and the number of people who took part in it was smaller than usual. Against this idea, the current study tried to figure out how the different parts of the Botswana Agriculture Science work.

The goals of a study conducted by Moyo (2015) study were to analyse student performance on the 2013 Botswana General Certificate of Education

(BGCSE) Agriculture Examination, determine the dimensionality and fairness of the examination for all students, and determine implications for differential item functioning. All 12784 students who took the 2013 BGCSE agricultural test made up the study's population. The study examined the psychometric parameter estimates of 40 test items using the 1PL, 2PL, and 3PL models. For each test item that fit into one of the three IRT models, dimensionality analysis and the chi-square test were performed. The logits test for t-test significance was used to undertake DIF analysis for each item depending on gender and location type (p .05). Eight items fit the 3PL, however just one item fit the 2PL. Twenty-nine (29) of the forty (40) items were gender-based DIF, with twelve (12) favouring females and seventeen (17) favouring men, The results of this study about sex-based DIF. The results showed that 18 had location based DIF, of which 10 favoured rural students and 8 favoured urban students.

The national examination tool was looked at as part of this study, and the results showed that the 2013 BGCSE Agriculture Examination was not the same for all students nor was it fair. So, it was suggested that test designers and organizations in charge of exams improve the quality of their test items by, among other things, using IRT psychometric analysis to validate DIF. In response to this recommendation, the current study wants to look at not just DIF, but also the difference between items, how hard items are, and how much guessing is involved. Also, the current study has a wider scope as it takes tests of three consecutive years, 2018, 2019 and 2020. The current study seeks to provide a more comprehensive picture of the agriculture science examinations. Furthermore, Moyo’s study focused on Agriculture Science at the BGCSE, whiles the current study focused on JCE which vary item samples. Further, the study did not provide possible source of DIF on the identified items.

Differential item functioning (DIF) studies have shown that the way test items work differently for men and women. This means that when students are given multiple-choice questions, there are different DIF questions for female and male students. Some subjects sometimes like women more than men, while other subjects also like men more than women. For example, some studies showed that females consistently did better than males on algebra-related questions, but males did better on geometry and measurement-related questions, but not as often. Again, women did better than men on the grammar, language use, and cloze tests. On the other hand, men do better on tests of vocabulary and word order. Both men and women do better on the part of the test that tests reading comprehension. Also, it was decided that DIF couldn't be explained by the item's format alone. These works were able to show that the amount and direction of DIF depends not only on gender but also on the subject area or on how the subject area and item format work together. But there isn't much proof in the case of Botswana.

DIF of JCE Agriculture Science Multiple Choice Items Based on School Location

Obiebi-Uyoyou asssessed how different math questions worked on the Senior Secondary School Certificate Examination in the Delta Central senatorial district in the year 2023. Five research questions and five hypotheses were used to plan the study. The goal of this study was to find out how different math questions did on the Senior Secondary School Certificate Examination in the Delta Central Senatorial District of Delta State. For this study, a research method called "ex-post facto" was used. All class three students from senior secondary schools in the Delta central senatorial district made up the sample for this study. It was done with a method called "proportional stratified random sampling." The validity and reliability of the WAEC/SSCE 2021 math multiple-choice questions and the socioeconomic status were used to collect data.

The Chi-square test was used to test the hypotheses, while the L-R method was used to answer the research questions. The results showed that the WAEC/SSCE 2021 math multiple-choice test items worked differently depending on gender, location, socioeconomic status, school type, and who owned the school. Twenty-four items out of a total of fifty were said to have gender DIF. In terms of where the examinees lived (rural or urban), 27 out of 50 items were marked as DIF. Lastly, DIF was present in twenty-one of the fifty items on a multiple-choice math test that had to do with the school the test takers went to (public and private school).

This study is important in that it covered all three demographic variables (gender, location, and school type) that are investigated in the present study. However, the study did not go further to point out whether items favoured females or males, urban or rural, and private or public-owned schools. This is a serious shortfall that may not assist item developers how to address the reported DIF. This is in view that gender is a social construct that is viewed in relation to how the socialisation of males and females influences their learning. The study also used a relatively small sample of 375 which in IRT-based studies is very small and may not accurately identify all DIF-related items.

Abedalaziz (2012) carried out a study titled “Exploring DIF: comparison of CTT and IRT method. The tool for collecting data was a math test with 60 items that could be scored only one way. The sample of the study was made up of 1280 students (656 males and 624 female). The DIF detection method used are Area index (IRT based), transformed item difficulty (CTT based), b-parameter difference (IRT based) and Scheuneman’s chi-square (CTT based).

TID shows that 35% of the items revealed DIF, b-parameter difference shows that 75% of the items revealed DIF, Area index shows that 77% of the items revealed DIF and Scheuneman’s chi-square shows that 50% of the items revealed DIF. Both the area index and chi-square methods agreed that 23 items showed DIF and seven did not. So, 56 percent of what they say agrees with each other. Both the b-difference and the chi-square (Scheuneman) agreed that 25 of the items showed DIF and 21 did not. So, 85 percent of what they say agrees with each other. Both the TID and the chi-square agreed that 16 items showed DIF and 23 did not. So, 72 percent of the time, they agree with each other. Both the b-difference and Area index methods agreed that 27 items showed DIF and 5 items did not. This means that there is 59 percent agreement between the two methods. Both the area index and the TID methods agreed that 16 of the items showed DIF and 6 did not.

As a result, they agree with each other 41% of the time. Both the TID and b-difference methods agreed that 17 items showed DIF and 20 items did not. This means that 69 percent of the time, they agreed with each other. The study found that chi-square and b-parameter difference had the highest level of agreement (85 %), while Area index and TID had the lowest level of agreement (41 %). Females had a statistically significant and consistent advantage over males on questions about relations and functions, polynomial functions, and trigonometric functions. Men, on the other hand, had a less consistent advantage on triangle-related questions, so it was decided that differences between men and women in math may be related to the content. This study is very important to the current study as it demonstrated that robustness of Raju area method in detecting DIF. This is very relevant since the current too adopted the Raju method.

Calhoun, Goeman, and Tsethlikai (2014) did a study on the 1986 Common Entrance Examination in Mathematics test takers from 522 rural areas and 512 cities Akwa Ibom and the Cross-River States in Nigeria to see if there was a difference in how they worked based on where they lived. For this study, 522 people lived in rural areas and 512 people lived in cities. She used the modified Scheuerman chi-square (SS2) process, the transformed item difficulties (TID-450), and the item discrimination methods as her three detection methods.

The SS2 approach found that out of the total of 33 multiple-choice exam items, 13 of them contained biased information. The transformed item difficulty uncovered a potential for bias in five of the test's items, while the discrimination method found differential item functioning (DIF) in nine of the test's items overall (out of a total of 33). Her finding provided evidence that was consistent with the idea that there is a geographical bias in mathematics success exams. However, the limitation of this study is that it used relatively small sample size and transformed item difficulty which have been reported to be less robust in detecting DIF.

Chime (2012) deveoped an Economics Achievement Test and make sure it was accurate. The study took place in the Enugu Education Zone of the state of Enugu. Six of the twenty-six secondary schools in the Enugu Education Zone that offer economics were chosen at random. One thousand twenty-five SS II students from the schools that were chosen were chosen at random. The researcher came up with two hypotheses and four research questions to guide the study. A table of specifications was made and used. It was based on the research questions, hypotheses, design, and methods. Based on the level of cognitive ability measured by the test, a 50-item test was made, and experts checked the draft AT to make sure it was correct. Mean, standard deviation, and the Kudder-Richardson (KR-20) reliability method were used to look at the data.

Based on the results of the analysis, it was found that: the developed achievement test instrument for Senior Secondary schools has high psychometric properties in terms of difficulty and discrimination index; the instrument has a high reliability index. The study showed that there was a big difference in average achievement between students in schools in cities and those in schools in rural areas. The difference was in favour of students in cities. Then, it was suggested that when using the Standardized Achievement Test, the location of a school should be taken into account.. The study used overall achievement mean to detect sex difference. This cannot be said to address how individual items function and contributed to observed sex achievement difference. However, this study is worthy looking at as DIF investigation starts because of observed disparities between subgroups.

Yohanna and Muhammad, in their study in 2022, looked at how location and gender affected how well students did in WASSCE Agricultural Science in the Zaria Educational Zone from 2014 to 2018. They chose the design after the fact. We got information about how well students did in schools that were chosen on purpose. With these kinds of data, simple descriptive statistics and an independent T-test were used. The results showed that 51.57 percent and 48.43 percent of the 2612 students who took the test were from rural and urban areas, respectively, and that 62.75 percent and 37.25 percent of the students were male and female, respectively. There were also statistically significant differences in how well students did based on whether they were male or female ($t = 3.99$) or whether they lived in a city or a rural area ($t = 2.76$). Also, the mean scores of students in rural areas (43.69 points) and women (43.21 points) were higher than those of students in cities (39.19 points) and men (40.50).

The study found that location and gender are factors that affect how well students do in Agricultural Science. To increase female students' interest in and performance in Agricultural Science, the study suggested making sure all students have the same educational rights, no matter what gender they are, and giving them better facilities, trained and qualified teachers, supervision, and monitoring.

The study looked at Agricultural Science which is also the focus of the current study. However, the study relied on mean scores and T-tests to make the comparison of the different demographic variables. This does not give much detail on which items functioned differentially. Unlike the current study that focuses on items, this study focused on the whole test.

Using IRT, Mokobi and Adedoyin (2014) did a quantitative study to find items that were different for rural and urban schools on the 2010 Botswana Junior Certificate Examination Mathematics Paper 1. The study also reported disparities between girls and boys in rural and urban settings. Math paper 1 of the 2010 Botswana Junior Certificate Examination had forty (40) multiple choice questions. For this study, 4000 students who took the 2010 Botswana Junior Certificate Examination Mathematics paper 1 were chosen at random from a group of 36940 students. The group of students chosen at random was made up of 2000 male students, 1000 of whom went to schools in rural areas and 1000 to schools in cities. The other 2,000 students were girls. One thousand came from urban schools and the other thousand came from rural schools. Though the study focused on mathematics, like the current study it used 3PLM to analyse its data. However, the sampling and sample size were different. Furthermore, it focused on senior high students and used smaller sample whole the current study looked at junior high level using large sample size.

Parameter estimates statistics were used to make Item Characteristics Curves (ICCs) for the rural/urban, rural/urban, and rural/urban groups that were similar in terms of gender. The ICCs for the groups were compared to find items that were biased toward rural or urban locations. The study's results showed that six (6) of the 24 items that fit the IRT (3PLM) model were biased toward rural or urban locations. The study also found that three (3) items were biased toward rural or urban location for men and six (6) items were biased toward rural or urban location for females.

Almost every research work, there has been evidence confirming the difference in Differential item functioning (DIF) of test items with reference to the location of school. There has been evident of a significant difference in mean achievement between students in schools located in urban areas and those in rural areas and in favour of urban students. Even with teachers DIF results found a significant difference in DIF associated with teachers' content coverage ratings and the theoretical item classification in rural and urban schools. Some additional studies also made it clear that, with DIF, there is a geographical bias in mathematics success exams. However, possible source on this gab of differences in DIF concerning urban and rural schools has not been explained in Agriculture science in Botswana.

DIF of JCE Agriculture Science Multiple Choice Items based on Type of School (Public or Private)

Ogbebor and Onuka, (2012) investigated items that are biased using differential item functioning approach in relation to school type (private and public schools), and 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 research type of design. The study sample comprised students in Delta State, Nigeria. A sample of 447 candidates was used. The test contained administered 60 items. Logistic regression was used to analysis the data. The study revealed that 17% of items were biased in relation to school type and 13% items in relation to school location.

This study like the current study focused on national examination. The national examinations are expected to follow similar or closely related procedures in standardisation and validation process. Notwithstanding, the study used a relatively small sample size as compared to the current one.

Osadebe and Agbure, in 2020, looked at the DIF in Social Studies multiple-choice questions on the Basic Education Certificate Examination (BECE). It used a research design called "ex-post facto." All Junior Secondary school students in Delta central senatorial district were part of this study's group. They used a method called "proportional stratified random sampling." The number of students in this study's sample was 500. The 2014 BECE Social Studies multiple choice questions were used to get the information. SPSS and WINSTEPS were used to look at the data. To answer the research questions, descriptive statistics were used, and independent chi-square tests were used to test the hypotheses.

The results showed that the way an item works depends on gender, location, socioeconomic status, school type, and who owns the school. This study, like the current one, used descriptive statistics to answer research questions and focused on the Junior secondary level. But the study was mostly about things from social studies, which may be different from agriculture science in terms of content and context. The sample size is small compared to the current study, which could mean that the size of the effect is different.

Scarpati, Wells, Lewis, and Jirka (2011) used differential item functioning (DIF) and latent mixture model analyses to find out why students in private schools and public schools in the USA performed differently on a large-scale math test. The results of a state accountability test in math for 8th graders were looked at. Using information from a recent (2007) state accountability assessment of arithmetic proficiency for pupils in Grade 8, the

DIF and mixed DIF analyses were done. There were 39 mathematics assessment items, 34 of which were given a dichotomous score.

The study used the 34 items with dichotomous scores. Moreover 73,000 pupils took the test, and 12,268 of them had special needs and needed various test accommodations. More than 73,000 students participated, of which 12,268 were students from public schools. DIF analyses showed that people who went to public schools and those who went to private schools performed differently on tests. Latent performance class analyses showed that differences in performance were linked to the difficulty and ability of the item as well as the fact that the school was private and in a rural area. Results back up validity studies that use mixture models that can take school type, academic skills, and accommodations as context variables. The study focused on mathematics while current study looked at agriculture science, but both looked at dichotomous items. the study used item difficulty to determine DIF while current study used item difficulty, discrimination and guessing parameter converted into Z-statistics. However, the two studies used large sample which are comparably.

Ayva Yorü and Atar (2019) looked at questions from the mathematics section of the Centralised High School Entrance Placement Test (HSEPT), which was given in 2012 by the Ministry of National Education in Turkey. The goal of this study was to find out if the mathematics subtest questions on the Centralised High School Entrance Placement Test (HSEPT), which was given by the Ministry of National Education in Turkey in 2012, show DIF based on gender and school type. For this purpose, the DIF of the 20 items on the mathematics subtest of the HSEPT in 2012 was calculated using the

SIBTEST, Breslow-Day, Lord's [chi-squared], and Raju's area measurement methods, and the items were found to have DIF based on these methods. The HSEPT, which eighth-graders took in 2012, was used to gather data for the study. After the missing data were taken out of the data set, DIF analyses were done on the math subtest of 1,063,570 (females: 523,939, men: 539,631, public school: 1,025,979, private school: 37,591). This study used a descriptive research method because its goal was to find out performance difference. According to the methods that were used, the number and level of DIF in the items with DIF varied by gender and type of school. This study's results show that there are at least two ways to figure out the DIF.

The strength with the approach by Ayva Yörü, and Atar, (2019) is that three different IRT methods of detecting DIF were used. This allowed for comparison of the robustness of these methods in detecting DIF. However, the study focused on mathematics while current study focused on agriculture science. Furthermore, the study does not give a detailed description of difference between public and private schools. This makes it difficult to compare the context of school type in Turkey and Botswana where the current study was carried out.

Dogan and Ogretmen (2008) wanted to see how similar and different the techniques were in practice. They did this by comparing Chi-square, the Mantel-Haenszel and logistic regression techniques, and the differential item functioning (DIF) determination techniques. The study was done with a group of 3,345 students chosen from the 600,000 students in Turkey who took the Selection and Placement Examination for both public and private secondary schools in 2003. The study's data came from the answers students gave on the science part of the Selection and Placement Examination for both private and public secondary education. Only groups of schools of the same type were used for the DIF analysis. The study's results showed that these techniques made a big difference in DIF analysis depending on the type of school.

The strength in the study was the use of CTT and IRT methods in detecting DIF. The like the current one used high stake examination and large sample was used. This study amplified the importance of choosing the right DIF detection method in DIF analyses.

Evidently more literature exists that focuses on the number of differential item functioning (DIF) items which exist when comparing students in different schools**.** Differential item functioning (DIF) was found to exist in large amount of items. In some studies about 40%–50% of the items exhibited some form of DIF and also latent performance class analyses revealed that performance differences were associated with item difficulty and ability in addition to private and rural status of the school. In the same manner, other results from other studies also indicated that these techniques provided a significant difference in DIF analysis based on the school type. However, private schools and government school comparison in Botswana in terms DIF has not been ascertained in Agriculture Science as per the literature reviewed.

Conceptual Framework

The conceptual framework is presented in Figure 9. The student achievement in Agriculture Science is the latent variable that is made observable through a characteristics of a set of 40 Agriculture Science test items taken in 2018, 2019 and 2020.

As shown in Figure 9 (i.e., the conceptual framework), the solid lines proposed that observed response manifested through interaction of item characteristics and student latent ability in agriculture multiple choice item. In this regard, a relatively easy item is expected to function at the left-hand side (negative) of the ability continuum, whereas a difficult item is expected to function at the right-hand side (positive) of the ability continuum. The study further conceptualised that each test item across the three examination years clearly discriminates low achievers from the high achievers. Put differently, each item should provide precise information about the examinee's ability on the item. Lastly, this study anticipates that there will be little or no guessing for each item across the three examination years.

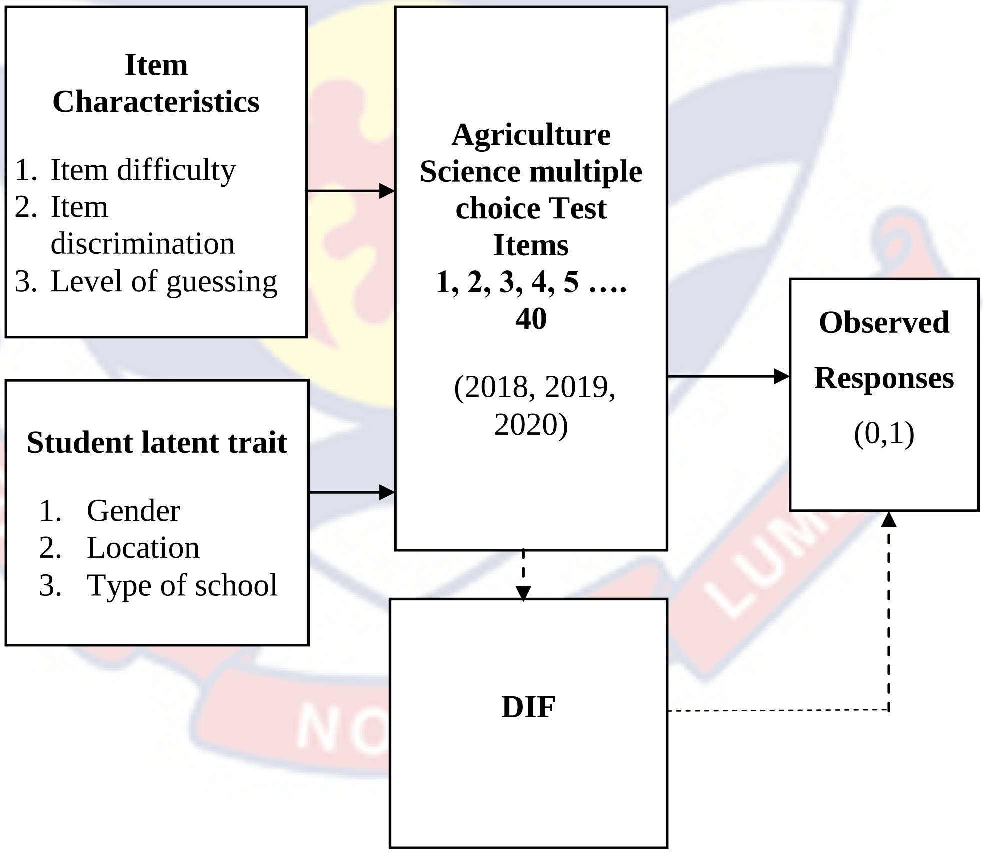


Figure 9- Conceptual Framework

Therefore, examinee's response to an item should solely be accounted for by their knowledge on the item, and not necessarily guessing. In practice, examinees taking the items are of different sex, from different school locations and from private as well as public schools. From the framework, the dashed arrows from group are interaction effects that changed the meaning of the construct being measured (achievement in agriculture). This implies that interaction effect from either gender, location, or school type influenced the observed responses in some of the items. Interaction effects between groups, conditions, and items are considered a construct-irrelevant source of variance – in other words, the item functioned differently and were not invariant over groups or conditions.

To put it another way, the item difficulties did not remain constant throughout all groups. When referring to a test's item difficulty, the phrase 'non-invariant over groups' indicates that the level of difficulty changes when it is given to different groups of people whose latent abilities are comparable to one another. This study used the IRT DIF detection approach in identifying the interaction effects seen at the item level between different items and different groups.

The study had projected the presence of no DIF in all of the test items across the three-year examination periods on the basis of gender. This is premised on the fact that for quality test items, regardless of the roles society assign to a particular sexual orientation, there is an equal probability that examinees of same ability would the get the item correct. In this regard, the test items are said to be favouring none of the gender categories. This is an ideal situation, since examinees taking the JCE in Botswana are exposed to the same content, and therefore, all item should not exhibit any differentiation.

In addition, study had projects that regardless of the location within which the schools of examinees are situated, they were expected to have same probability of getting an item right, provided they have the same level of ability. This is an indicative of a quality test item. For that matter, any difference in examinees' performance on an item is attributable to their knowledge that specific item but not the location of their school. In a similar manner, whether an examinee enrolled in a private school or public school, with the same level of ability, they are expected to exhibit same likelihood of getting an item right. This position of the current study emanates from the fact that both private and public schools in Botswana run the same curriculum. In the event of a test item favouring one group over the other, then the quality of such item(s) is questionable. Summarily, the framework for this study provides the lens through which the psychometric properties of the BEC test items must be examined.

Chapter Summary

In most of the studies on psychometric properties showed that there were items with poor parameter indices and also showed DIF by gender, location, and type of school. Furthermore, there were differences in how well boys and girls do on agricultural science items, even when the tests are closely tied to the curriculum. Different studies used different IRT detection methods, but most of them used 3PL IRT statistical analysis. This method is reported as most robust with its greatest limitation being the large sample size requirement. The studies point to the model fit and sample size as the basis for determining ideal IRT DIF detection method. Most studies focused on mathematics and science with very few on agriculture. From the literature search there seem to be very few studies on the quality of test used in Botswana.

Even though SIBTEST, Likelihood-Ratio test and Lord's Chi-square approaches have limitations of quantifying the magnitude of DIF most studies used them. Raju which is the strongest of all the DIF detection approach was found to be under-utilised. The contribution of items to construct being measured was also missing in all the reviewed studies. Based on this, there was the need for a study which uses the Raju approach and this was exactly what this study sought to do.

CHAPTER THREE

RESEARCH METHODS

Introduction

The research methods of this present study are discussed in this chapter. The first part of the chapter focuses on the main idea behind the study. The goals of this study were met by using quantitative methods and a positivist worldview. The next section is about how the research was set up, who was in the study area, how samples were chosen, what instruments were used, and how data were collected, processed, and analysed.

Research Paradigm

When choosing a research paradigm, it is expected that the research follows the paradigm's assumptions, beliefs, norms, and values and be guided by them. So, it's important to show that the researcher understands each of these elements.

This study was based on the positivists' view of the world. Based on this, the researcher worked under the assumption that reality is stable and objective and can be seen and measured (Ansari, Pahnwar, & Shah, 2017; Cohen, Manion, & Morrison, 2013). In this study, the item quality is seen as a stable variable, so students with high ability in Agriculture Science should always be able to demonstrate that. Because of this, "student ability" is a variable that is observable, is objective, and can be measured. Also, the researcher was guided by the ideas of proof, causal links, and verification between the pieces of information used and recognizing factors that affect outcomes (Creswell, 2009). According to this study, students respond to items by assigning numbers to show how much they know about agriculture science. Since this was the case, the main method of research for this study was quantitative, which is in line with the ideas of positivists.

Furthermore, in consideration of data analysis, for instance, the positivist paradigm indicates that the data collected is quantitative and most likely to be assessed statistically. The use of the interpretive paradigm, on the other hand, relates to qualitative information-gathering and analysis techniques and methodologies. The researcher subscribes to the Positivist's worldview. The researcher believes that an objective world exists independently of the person and consists of causally interacting things that may be observed. This is because the researcher will be able to statistically analyse the agricultural test via the use of a quantitative technique (Leavy, 2017). Quantitative research, according to Apuke (2017), is the exploitation and analysis of numerical data using statistical approaches to answer questions such as who is doing it and how much is being done.

The IRT which underpins this study is probabilistic and estimates the ability of examinees. In this study, the quality of multiple-choice items as a variable, is characterised as stable, such that an examinee who has high achievement ability in Agriculture Science should consistently meet the demands of items in the test. This makes the variable "achievement ability" observable, objective, and quantifiable. Given the different demographic dispositions of examinees, the researcher was guided by the ideologies of interactions and unwanted effects of DIF to observed outcomes.

Research Design

A descriptive research design was used for the current research. This type of research seeks to explain events, symptoms, or specific groups of people based on numerical data from tests. In this study, the characteristics of the test items were explained according to the nature of the agricultural questions used in 2018, 2019, and 2020 Botswana Junior Certificate Examination. This design is relevant to this study as it allows the analysis of available data to describe the JCE agriculture science multiple choice items.

As noted in the literature review, studies across the globe amplify the significance of high-stakes examination in decision-making. In this sense, it would be very useful to use the descriptive research design to describe the psychometric properties of the JCE agriculture multiple choice items. This would help verify the quality of JCE agriculture multiple choice items that meet the IRT 3PL criteria and are suitable to accurately capture students' abilities.

In addition, IRT detects DIF and unidimensionality for all items, therefore, has great strength to determine the accuracy and core dimensions of the 2018-2020 JCE agriculture multiple choice items. This is believed to provide the researcher with guidance on how to develop and monitor a test instrument from administration to the next (Boone & Canterbury, 2005). As a result, the results of this study will provide insight into Botswana's assessment particularly JCE agriculture multiple choice items.

Study Area

The Botswana Examination Council is parastatal under the Ministry of Basic Education (MoBE). The council was established in 2001 to manage and conduct national examinations (Republic of Botswana, 2001). The council currently runs Primary School Leaving Examination (PSLE), Junior Certificate Examination (JCE), and Botswana General Certificate of Secondary Education (BGCSE). The Directorates of Products Development and Standards leads the item developed with assistance of the test Committees, which includes stakeholders from various departments of MOBE. As the custodian of Junior Certificate Examination (JCE) the BEC keeps all records of student responses examination items. Therefore, relied solely on their archive to data retrieval.

The Botswana Examination Council is the only testing body tasked with examinations at basic education level (primary school and secondary school levels). Few studies have been conducted to understand the quality of examination offered by BEC (Moyo, 2015; Adedoyin, 2010). These previous studies limited scope to fairness which is quality and item quality and raised concerns about item bias and (2) the researcher has experienced the development of agriculture items before which serve as a motivation to conduct this study.

Population

One of the problems faced by a researcher is to determine the population of the people whom to makes assertions. At this point, the researcher must be able to ascertain the population's size. As such, the term ‘study population’ refers to the broad population that the researcher wishes to investigate. The population of a study group is a target group that has aspects which the researcher is interested in studying. These members are a group of people or subjects who share one or more characteristics of interest to the researcher (Leavy, 2017).

For this study, the scores of individual items of all 123,218 examinees who wrote the junior certificate examination from 2018 to 2020 constituted the population of the study. Those students had completed the three-year agriculture curriculum at the time of data collection. The study targeted all responses to individual multiple-choice items in each of the three years.

Demographic information of the examinees

The data obtained spanned three years from 2018 to 2020, including data on gender, school type and school location. The distribution of cases across the three years is shown in Table 5.

Table 5- *Demographic Information of Students who sat for the BEC Agricultural Examination from 2018 to 2020 (N=123,218)*

Variable	2018		2019		2020	
	n	%	*n*	%	*n*	%
Gender						
Male	20096	49.8	20049	49.3	20852	49.4
Female	20247	50.2	20617	50.7	21357	50.6
School Type						
Public	38321	95.0	38508	94.7	40314	95.5
Private	2022	5.0	2158	5.3	1895	4.5
School Location						
Urban	17023	42.2	17696	43.5	17994	42.6
Rural	10208	25.3	9840	24.2	10300	24.4

Source: BEC 2018-2020

For the three years, the sex distribution showed that more female students sat for the BEC examination as compared to male students (see Table 5). For example, females constituted 50.2% (vs. 49.8% of males), 50.7% (vs. 49.3% of males) and 50.6% (vs. 49.4%) for 2018, 2019 and 2020 year respectively. Records displayed in Table 5 show that public schools dominate in the public-private school proportions. Over 90% of the students who participated in the BEC examination were affiliated with public schools with about 5% coming from private schools. The information in Table 5 also shows that most of the students went to schools in urban schools (over 40 %), while only 24 % to 25 %percent went to schools in rural areas during the three years of study.

In terms of responding to the examination items, the study targeted 123,218 examinees who wrote 40-item test, and this translates to 4,928,720 cases of data point obtained from the Botswana Junior Certificate Examination from 2018, 2019, and 2020 years, respectively. Each student was expected give one response for each item.

Sampling Procedures

Amedahe and Asamoah-Gyimah (2015) show how important sampling is for getting a full picture of the whole population. Sample surveys are needed because it's hard to cover everyone, they save time and money, and they cover the whole population in a short amount of time and get similar results.

Several investigations into the effects of sample size and test length in IRT have been conducted. These studies have suggested at least 200 examinees to 5000 with a minimum of 20 items (Hambleton & Cook, 1983; DeMars, 2010; Guyer & Thompson, 2011; Akour & Al-Omari, 2013). Despite their varied findings, all the studies agreed that there is a positive association between sample size and stability in item parameter estimates. Thus, IRT is a large sample size-based theory.

In line with these findings, coupled with the conditions necessary for sample survey, the researcher decided to use census approach. The study used cases and data points. Thus, the sample size was all 123 218 examinees resulted in 4,928, 720 data points from the 120 items responded to over three years. This was the case because the data on the scores of the examinees was found readily available in binary format on an Excel sheet making complete coverage of the population easy and economical. Furthermore, this would allow BEC the opportunity to compare their CTT-generated item analysis indices (thus, their usual analytic procedure) with the IRT without doubt of sampling error contribution.

Data Collection Instrument

The JCE is a test of how well students have done in school. It is given to students in their third year of junior secondary school. The test shows how well the students have learned what they are supposed to learn, which in this case is about agriculture science. As a public testing body, BEC reviews how its tests are given on a regular basis so that there aren't too many mistakes, and the scores are very accurate.

This study used secondary data on students' responses to agriculture science multiple choice items for the years 2018-2020 (Appendix A). The Agriculture Science multiple-choice items are developed from crop science, soil science, animal husbandry, agriculture economics and agricultural engineering. The paper has four options from, which one correct answer is chosen while the remain three options serve as distractors. For the three years under study agricultural economics and engineering sections had the least items.

Student records contained information about each student's responses, the school location and whether the school is public owned or privately owned. Therefore, no specific instrument was developed for this study. Data on students' responses to JCE agriculture multiple choice items that were collected between 2018 and 2020 was used. According to Bulala and Malema (2019), intensive panel-based content analysis and face validation is carried out by Botswana Examination Council by a subject specialist. Therefore, it is assumed that the instrument was reliable.

Data Collection Procedures

Following approval of the proposal, ethical clearance was obtained from the Institutional Review Board (IRB) of the University of Cape Coast (UCC). The Department of Education and Psychology at UCC also issued a letter of introduction, which the researcher used to write another letter asking for permission to access and use the data (see Appendix B). Copies of these letters with researchers' curriculum vitae were sent to BEC formally requesting permission to access and use the data. This was done after initial contact with the acting director of the Directorate of Research and Policy Development to inform the council in advance and discuss with them what the study wanted to achieve and the need for the study. Student-by-item responses for the years 2018, 2019, and 2020 were collected as text data files on the Microsoft Excel platform and were shared through email.

Ethical Consideration and Data Management

The study considered ethical issues such as IRB approval, confidentiality, data de-identification, and anonymity of data. The researcher applied for ethical clearance from the UCC IRB (see Appendix C) and Botswana Examination Council (see Appendix D). Moreover, privacy was not compromised as data details were not disclosed to any third party. This is supported by Hasan (2021) who posited that handling secondary data analysis should not result in the disclosure of sensitive information about individuals.

It is the BEC's policy that students' responses to examination items should be kept confidential from anyone outside the BEC. To ensure confidentiality, BEC de-identified the data before sharing it with the researcher. There were no names of students or schools on the data. For informed consent, it was provided by BEC. Indeed, during the examination phase, students authorize the council to use the data for management and administration purposes as well as for future research purposes (Jol & Stommel, 2016). Based on this, informed consent was provided during the examination review, which also allowed the council to provide the data to any party for the purposes mentioned above.

To keep data protected from third parties' passwords and codes were created. A hard copy of the data was kept under lock. Also, a softcopy was uploaded to the email and dropbox account and the researcher's external drive for backup purposes. The data was kept for a year after collection before it was destroyed. This time frame allowed the researcher to address any suggestions regarding the analysis before the final research work is submitted and approved. Such a level of privacy is considered sufficient to ensure the protection of human rights.

Data Processing and Analysis

The data were checked for mistakes and cleaned up to ensure uniformity between data. The data retained official codes for the school and test items, but each examinee was given a unique five-digit code, like 00001, 00002, etc. Table 6 shows the codes for the answers.

An analysis of the data using SPSS version 23 was done for demographic information. Prior to the DIF analysis, first, the test items were checked to determine the dominance of the first factor (unidimensionality), secondly assessed for model fit, before IRT psychometric analysis are done to estimate the test item's parameters using data from students’ responses.

Table 6- *Codes for Different Criterion and Demographic Variables*

Variables	Code
Gender	Male = 1; Female = 2
Location	Urban = 1; Rural = 2
School type	Public = 1; Private = 2
Correct response	1
Wrong response	0

Item information curves and test information curves were also used to depict poor and good items as well as the whole tests.

The test for unidimensionality in this study was through exploratory factor analysis. It is used to find out if there exists a dominant trait in agriculture multiple-choice. This helps to determine the construct validity of the test. Test for model fit was also assessed by comparing the values of -2Log likelihood from different models to indicate which model represents a better fit for the data.

The DIF analysis was conducted using the DifR package within the R-studio environment. According to Oshima and Morris, (2008D) DifR uses the differential functioning of items and tests (DFIT) technique for assessing DIF. The procedure has the capability of analysing both dichotomous and polytomous test items and being able to detect uniform and non-uniform DIF (Oshima & Morris, 2008). Primarily, the IRT Raju estimation procedure was utilised (Oshima, Raju, & Nanda, 2006), providing the Raju Z statistics and their associated p-values. It must be noted that DIF approaches like Mantel-Haenszel statistics (Holland & Thayer, 1988) and Simultaneous Item Bias Test (SIBTEST) have their statistics expressed on a scale and thus, the value of the statistics automatically gives an idea of which group the flagged item favours (Zieky, 2003).

Unlike Mantel-Haenszel and SIBTEST procedures, the group with which the DIF flagged items favours is unknown within the Raju area DIF method, and thus, the Z statistics only tell whether DIF is present or not. This limitation is addressed by comparing the difficulty thresholds and discrimination indices of the groups being compared to identify the group favoured by DIF item (Oshima & Morris, 2008). The effect size estimation for the DIF flagged items was performed using the approach proposed by Wright and Oshima (2015) using the ETS's effect size categories at given levels of *a*, *b*, and *c* parameters as the benchmark. Following the interpretation and recommendations of the Educational Testing Services (ETS), only DIF items with large effect size (C-level) were considered as DIF items. This approach has been widely accepted and has been successfully adopted in previous studies (Yörü & Atar, 2019). Table 7 summarises analysis procedures done.

Table 7- *Summary of the Analysis Procedure*

No.	Research Questions	Analysis Plan
1.	What are the characteristics of the JCE agriculture multiple choice items organised by BEC based on the (a) difficulty parameter, (b) discrimination parameter, and (c) level of guessing?	1st **Phase**: Assumptions were tested for 3PL. They include dimensionality, local independence, equal discrimination indices, minimal guessing, and power test administration. Internal structure 2nd **Phase**: Conducted a statistical test to determine overall model fit, item fit and person fit. (likelihood-ratio chi-square statistic) 3rd **Phase:** Conducted 3PL analysis and report on difficulty, discrimination, and guessing parameters.
2.	What is the contribution of the organized by BEC to the measure of the student's achievement in the subject?	The following indicators were generated: Item (and test) characteristics curves, Item (and test) information function with their curves, and marginal reliability (precision level against standard error)
3.	What is the level of DIF of the JCE agriculture multiple choice items administrated by BEC; in terms of sex?	**Software:** DifR package within the R-studio environment was used.
4.	What is the level of DIF of the JCE agriculture multiple choice items administered by BEC regarding school location?	1st **Phase**: Measurement invariance assumption was conducted to test for invariance across school locations. 2nd **Phase**: DIF analysis using IRT was conducted.
5.	What is the level of DIF of the JCE agriculture multiple choice items administered by BEC regarding school type (government and private)?	1st **Phase**: Measurement invariance assumption was conducted to test for invariance across school types. 2nd **Phase**: DIF analysis using IRT was conducted.

Chapter Summary

The study was grounded in the positivists' paradigm using the descriptive quantitative approach. The study covered Agriculture science examination candidates only. The population was targeted to public and private school candidates which were 123,218 for the 2018-2020 exam review period resulting in 4,928,720 cases of data point. The study used secondary data on students obtained from the Directorate of Research and Policy Development (DRPD), BEC. All candidates were involved in this study. Thus, the sample size was 4,928,720 obtained using the census approach.

Ethical approval was received from UCC's Ethics Review Board after the proposal was approved. This was followed by an introductory letter received from the Department of Education and Psychology. A letter was also obtained from one of the supervisors. The researcher then wrote another letter and attached it to the first two together with a copy of the proposal. This package was then sent to BEC to officially seek permission to access and use the data on student responses. This was done after an initial meeting with the Director of DRPD-BEC to initially brief the Board and discuss with them the scope of my research work and its significance to BEC (Creswell, 2012). Key ethical issues of anonymity, data de-identification, and confidentiality were maintained. The data were analysed by conducting psychometric analyses using 3PL-IRT procedure.

CHAPTER FOUR

RESULTS AND DISCUSSION

The study assessed the psychometric properties of JCE agriculture multiple choice items organised by BEC from 2018 to 2020. This chapter presents the results as well as the discussion based on the objectives of the study. The chapter addresses the specific questions which guided the study. The chapter concludes with a discussion of the results from the analysis.

Preliminary Analyses

The preliminary analyses included two key sections. The first sub-section highlights the exploration of the response distribution for each item for the three years. This is followed by another section which outlines the main assumptions underlying the use of IRT.

Descriptive analyses of the Data

An initial descriptive analysis, including a binomial test, was carried out to explore the data regarding the correct or wrong response provided by the students for the three sets of results. Tables 8 to 10 present the descriptive information.

As presented in Table 8, the responses to the 2018 BEC examination items varied from item to item. Whereas for some items an overwhelming number of examinees had the item right, this was the opposite for other items. Taking item 13 for instance, over 80% of the students correctly answered the question with close to 19% choosing the incorrect response.

Similarly, for item 27, about 10% selected an incorrect response with nearly 90% getting the answer right. Other items showed a different trend of results, with a majority of the students selecting the incorrect response. An example is items 8 and 9 where over 70% of the examinees had the item wrong. Other items include item 36 which had a distribution of 73.1% of examinees getting the question incorrect.

Table 8- *Binomial Test for 2018 Batch of Result (N=40343)*

No.	R	n	pr	*p*	*No.*	*R*	n	pr	p
Q1	0	18677	0.463	< .001	Q21	0	26389	0.654	< .001
	1	21666	0.537	< .001		1	13954	0.346	< .001
Q2	0	23999	0.595	< .001	Q22	0	34323	0.851	< .001
	1	16344	0.405	< .001		1	6019	0.149	< .001
Q3	0	16400	0.407	< .001	Q23	0	8895	0.220	< .001
	1	23943	0.593	< .001		1	31447	0.780	< .001
Q4	0	12841	0.318	< .001	Q24	0	9406	0.233	< .001
	1	27502	0.682	< .001		1	30936	0.767	< .001
Q5	0	19462	0.482	< .001	Q25	0	25409	0.630	< .001
	1	20881	0.518	< .001		1	14933	0.370	< .001
Q6	0	19327	0.479	< .001	Q26	0	22176	0.550	< .001
	1	21016	0.521	< .001		1	18166	0.450	< .001
Q7	0	27143	0.673	< .001	Q27	0	4353	0.108	< .001
	1	13200	0.327	< .001		1	35989	0.892	< .001
Q8	0	31272	0.775	< .001	Q28	0	31147	0.772	< .001
	1	9071	0.225	< .001		1	9195	0.228	< .001
Q9	0	29231	0.725	< .001	Q29	0	16892	0.419	< .001
	1	11112	0.275	< .001		1	23450	0.581	< .001
Q10	0	16892	0.419	< .001	Q30	0	27170	0.673	< .001
	1	23451	0.581	< .001		1	13172	0.327	< .001
Q11	0	14101	0.350	< .001	Q31	0	14642	0.363	< .001
	1	26242	0.650	< .001		1	25699	0.637	< .001
Q12	0	24773	0.614	< .001	Q32	0	10748	0.266	< .001
	1	15570	0.386	< .001		1	29593	0.734	< .001
Q13	0	7548	0.187	< .001	Q33	0	23219	0.576	< .001
	1	32795	0.813	< .001		1	17121	0.424	< .001
Q14	0	22550	0.559	< .001	Q34	0	11844	0.294	< .001
	1	17793	0.441	< .001		1	28495	0.706	< .001
Q15	0	15550	0.385	< .001	Q35	0	23757	0.589	< .001
	1	24793	0.615	< .001		1	16582	0.411	< .001
Q16	0	27716	0.687	< .001	Q36	0	29472	0.731	< .001
	1	12627	0.313	< .001		1	10865	0.269	< .001
Q17	0	26827	0.665	< .001	Q37	0	27643	0.685	< .001
	1	13516	0.335	< .001		1	12692	0.315	< .001
Q18	0	18214	0.451	< .001	Q38	0	21427	0.531	< .001
	1	22129	0.549	< .001		1	18899	0.469	< .001
Q19	0	13424	0.333	< .001	Q39	0	23004	0.571	< .001
	1	26919	0.667	< .001		1	17264	0.429	< .001
Q20	0	19976	0.495	0.052	Q40	0	16033	0.405	< .001
	1	20367	0.505	0.052		1	23572	0.595	< .001

R-response; 0- wrong response; 1- correct response
p- significant value; pr- proportion of responses

The distribution for the year 2019 also showed a mixed result; whereas some items showed an even distribution between correct and incorrect responses, others revealed either overwhelmingly correct endorsement or incorrect responses (see Table 9).

Table 9- *Binomial Test of 2019 Batch of Result*

No.	R	n	pr	*p*	No.	R	n	pr	p
Q1	0	20302	0.499	0.762	Q21	0	24121	0.593	< .001
	1	20364	0.501	0.762		1	16544	0.407	< .001
Q2	0	21954	0.540	< .001	Q22	0	11887	0.292	< .001
	1	18712	0.460	< .001		1	28778	0.708	< .001
Q3	0	22427	0.551	< .001	Q23	0	2389	0.059	< .001
	1	18239	0.449	< .001		1	38276	0.941	< .001
Q4	0	23578	0.580	< .001	Q24	0	4231	0.104	< .001
	1	17088	0.420	< .001		1	36434	0.896	< .001
Q5	0	2983	0.073	< .001	Q25	0	12298	0.302	< .001
	1	37683	0.927	< .001		1	28367	0.698	< .001
Q6	0	29408	0.723	< .001	Q26	0	11925	0.293	< .001
	1	11258	0.277	< .001		1	28740	0.707	< .001
Q7	0	30568	0.752	< .001	Q27	0	13137	0.323	< .001
	1	10098	0.248	< .001		1	27528	0.677	< .001
Q8	0	11302	0.278	< .001	Q28	0	29369	0.722	< .001
	1	29364	0.722	< .001		1	11296	0.278	< .001
Q9	0	10447	0.257	< .001	Q29	0	29478	0.725	< .001
	1	30219	0.743	< .001		1	11186	0.275	< .001
Q10	0	29489	0.725	< .001	Q30	0	18261	0.449	< .001
	1	11177	0.275	< .001		1	22403	0.551	< .001
Q11	0	31972	0.786	< .001	Q31	0	9673	0.238	< .001
	1	8694	0.214	< .001		1	30990	0.762	< .001
Q12	0	10627	0.261	< .001	Q32	0	23579	0.580	< .001
	1	30039	0.739	< .001		1	17084	0.420	< .001
Q13	0	12759	0.314	< .001	Q33	0	35107	0.863	< .001
	1	27907	0.686	< .001		1	5556	0.137	< .001
Q14	0	8662	0.213	< .001	Q34	0	9559	0.235	< .001
	1	32004	0.787	< .001		1	31103	0.765	< .001
Q15	0	34313	0.844	< .001	Q35	0	12937	0.318	< .001
	1	6352	0.156	< .001		1	27724	0.682	< .001
Q16	0	34583	0.850	< .001	Q36	0	19804	0.487	< .001
	1	6082	0.150	< .001		1	20854	0.513	< .001
Q17	0	20556	0.505	0.027	Q37	0	15000	0.369	< .001
	1	20109	0.495	0.027		1	25654	0.631	< .001
Q18	0	6404	0.157	< .001	Q38	0	22388	0.551	< .001
	1	34261	0.843	< .001		1	18253	0.449	< .001
Q19	0	18748	0.461	< .001	Q39	0	22501	0.554	< .001
	1	21917	0.539	< .001		1	18097	0.446	< .001
Q20	0	16218	0.399	< .001	Q40	0	21334	0.532	< .001
	1	24447	0.601	< .001		1	18780	0.468	< .001

R-response; 0- wrong response; 1- correct response
p- significant value; pr- proportion of responses

For items like 5, 9, 22, 24 and 31, a larger proportion of the examinees selected the right option, usually ranging between 70% to over 90% examinees. The data further discovered that there were some items (like 11, 15, 16, 28 and 33) which saw about 70% to 90% of the examinees choosing the incorrect answer (see Table 9).

Table 10- *Binomial Test of 2020 Batch of Result*

No.	R	n	pr	*p*	No.	R	n	pr	p
Q1	0	19851	0.470	< .001	Q21	0	30267	0.717	< .001
	1	22358	0.530	< .001		1	11942	0.283	< .001
Q2	0	21063	0.499	0.690	Q22	0	26144	0.619	< .001
	1	21146	0.501	0.690		1	16065	0.381	< .001
Q3	0	16507	0.391	< .001	Q23	0	4781	0.113	< .001
	1	25702	0.609	< .001		1	37428	0.887	< .001
Q4	0	6867	0.163	< .001	Q24	0	19852	0.470	< .001
	1	35342	0.837	< .001		1	22357	0.530	< .001
Q5	0	14581	0.345	< .001	Q25	0	20171	0.478	< .001
	1	27628	0.655	< .001		1	22038	0.522	< .001
Q6	0	5970	0.141	< .001	Q26	0	22470	0.532	< .001
	1	36239	0.859	< .001		1	19739	0.468	< .001
Q7	0	26447	0.627	< .001	Q27	0	9331	0.221	< .001
	1	15762	0.373	< .001		1	32878	0.779	< .001
Q8	0	13055	0.309	< .001	Q28	0	17261	0.409	< .001
	1	29154	0.691	< .001		1	24948	0.591	< .001
Q9	0	17232	0.408	< .001	Q29	0	18674	0.442	< .001
	1	24977	0.592	< .001		1	23535	0.558	< .001
Q10	0	11758	0.279	< .001	Q30	0	18019	0.427	< .001
	1	30451	0.721	< .001		1	24190	0.573	< .001
Q11	0	18318	0.434	< .001	Q31	0	6256	0.148	< .001
	1	23891	0.566	< .001		1	35952	0.852	< .001
Q12	0	16337	0.387	< .001	Q32	0	40462	0.959	< .001
	1	25872	0.613	< .001		1	1746	0.041	< .001
Q13	0	12860	0.305	< .001	Q33	0	17399	0.412	< .001
	1	29349	0.695	< .001		1	24809	0.588	< .001
Q14	0	18508	0.438	< .001	Q34	0	5417	0.128	< .001
	1	23701	0.562	< .001		1	36791	0.872	< .001
Q15	0	10617	0.252	< .001	Q35	0	15014	0.356	< .001
	1	31592	0.748	< .001		1	27194	0.644	< .001
Q16	0	10444	0.247	< .001	Q36	0	16512	0.391	< .001
	1	31765	0.753	< .001		1	25695	0.609	< .001
Q17	0	9993	0.237	< .001	Q37	0	27423	0.650	< .001
	1	32216	0.763	< .001		1	14782	0.350	< .001
Q18	0	31353	0.743	< .001	Q38	0	13522	0.320	< .001
	1	10856	0.257	< .001		1	28672	0.680	< .001
Q19	0	35561	0.842	< .001	Q39	0	13243	0.314	< .001
	1	6648	0.158	< .001		1	28902	0.686	< .001
Q20	0	9865	0.234	< .001	Q40	0	10210	0.246	< .001
	1	32344	0.766	< .001		1	31336	0.754	< .001

R-response; 0- wrong response; 1- correct response
p- significant value; pr- proportion of responses

The results presented in Table 10 showed the distribution of responses who sat for the BEC examination for the year 2020. As expected, performances on the items varied from one item to another. For question 2, for instance, an almost equal number of examinees selected either the right option (49.9%) or the incorrect response (50.1%). This result distribution for question 2 shows that examinees who chose the correct option were not significantly different from those who selected the incorrect option. In other instances, the response distribution was significantly dominated by examinees who correctly answered the question, as in the case of items 1, 6, 10, 23, 27, 31, and 34, among others. What is peculiar about these items is that over 70% of the examinees correctly answered the questions. The analysis also revealed some items where the majority of the examinees selected a wrong answer (e.g., items 18 and 19). A particular point in case is item 32 which recorded over 95% of the examinees getting the answer wrong.

Assumptions underlying the use of IRT and model selection

Unidimensionality assumption

To determine if the items measured a single construct, dimensionality of the data was checked before the analyses of the objectives. First, an exploratory factor analysis (EFA) was carried followed by a confirmatory factor analysis (CFA) with IRT. The dimensionality results are shown in Figure 10 and Table 11.

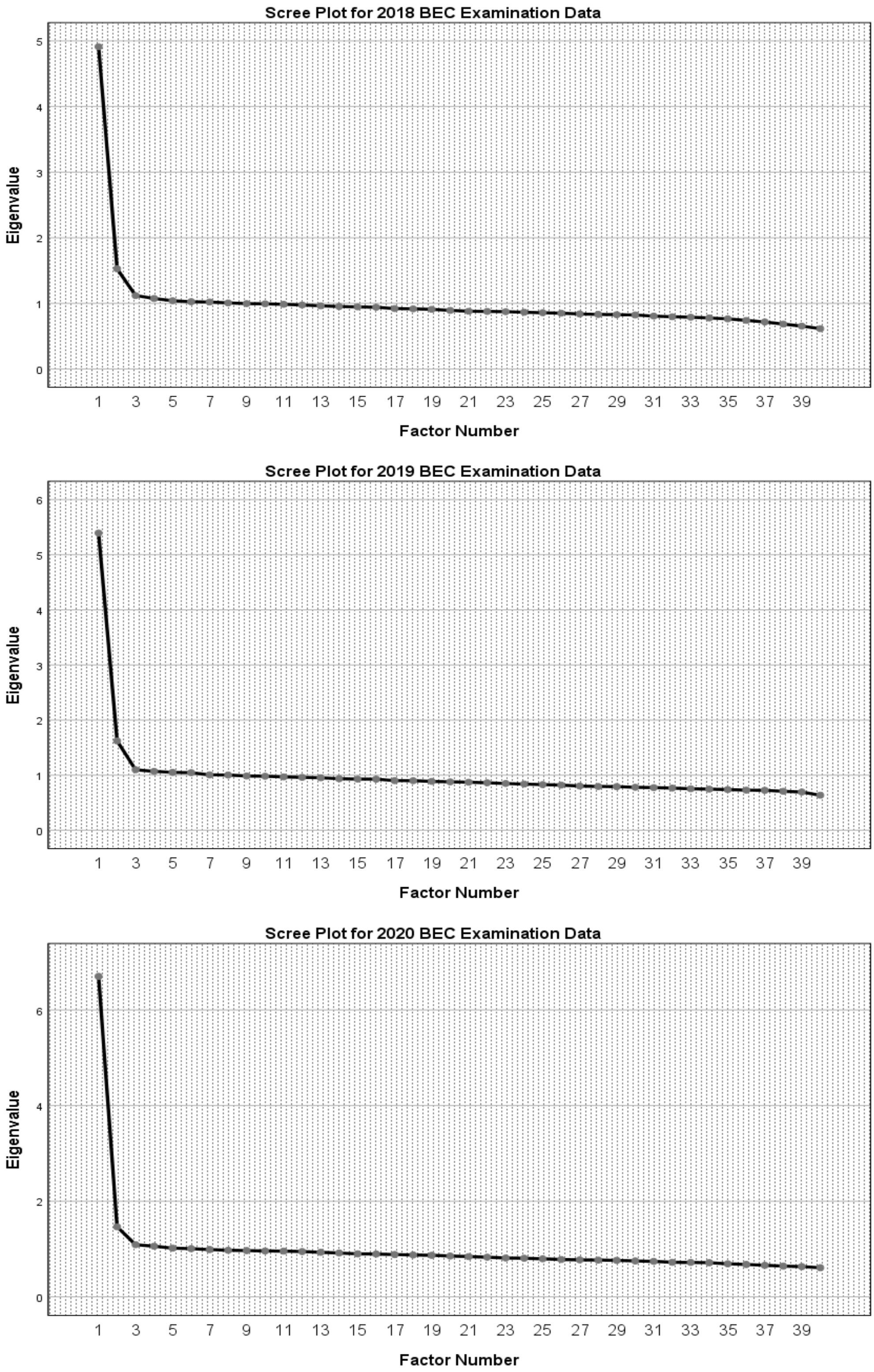


Figure 10- Scree plots for BEC Examination Results for 2018 to 2020

As an assumption to the use of EFA, the appropriateness of the analysis was tested for the three-year data using the Kaiser-Meyer-Olkin Measure (KMO) of sampling adequacy and Bartlett's test of sphericity (χ^2). The EFA analyses showed that the data satisfies the criteria for conducting EFA analysis (2018: KMO=.931, $\chi^2(780)$= 120650, p<.001; 2019: KMO= .942, $\chi^2(780)$= 142264, p<.001; 2020: KMO= .960, $\chi^2(780)$= 202410, p<.001).

The outcome of the EFA (unweighted least square estimation), as shown in the scree plots revealed a competing factor exploration between 1-factor and 2-factor models. Following the EFA analysis, the 1-factor and 2-factor CFA models were fitted to derive the most suitable model for the data. The outcome of the CFA results is shown in Table 11.

Table 11- *Model Fit Indicators for the Unidimensional and 2-factor Models*

Year	Models	AIC	BIC	Log-Likelihood	p-value
2018	Unidimensional	1959862	1960206	-979891.1	
	2-factor	1987561	1996310	-970123.4	<.001
2019	Unidimensional	1806355	1806700	-903137.5	
	2-factor	1900343	1908926	-879455.2	<.001
2020	Unidimensional	1870618	1870964	-935269.0	
	2-factor	1899533	1896812	-986218.0	<.001

AIC- Akaike information criterion; BIC- Bayesian information criterion

The model fit indicators for the CFA revealed that the 1-factor models across the three-year data showed significant adequate fit over the 2-factor model. The AIC and BIC together with the log-likelihood test showed lower values for the 1-factor models (2018, LL: -979891.1 vs. -970123.4, p<.001; 2019, LL: -903137.5 vs. -879455.2, p<.001; 2020: -935269.0 vs. -986218.0, p<.001).

Local independence

The local dependency hypothesis was tested to find out whether the responses to the items are attributed to the attribute being measured and other variables. By inspecting the local dependency results (Appendix E), it was revealed that the majority of the item pairs had a low local dependency, with values less than 1. Very few item pairs showed moderate dependency concerns with values between 1 and 5.

Model selection

Before the specific model (i.e., 1PL, 2PL, or 3PL) was selected, three indicators were inspected. First, the difficulty parameters for all three models were examined and varying levels of difficulty parameters across the three-year data were observed. Secondly, the guessing parameter for the 3PL model was scrutinized; it was found that guessing was prevalent for some items that could not be ignored. Finally, the model fit indicators for the three models were also compared (see Table 12). These three methods were combined to decide on the specific model to select for the analysis.

For the three years, the 3PL model was found to be more appropriate for the data set. This conclusion was drawn after comparing the model fit indices of the three logistic models. For the 2018 BEC examination results, for instance, it was discovered that the 3PL model was superior to the 1PL (-979891.0 vs. -952274.8, LRT=55232.7, $p<.001$) and the 2PL models (-954938.8 vs. -952274.8, LRT=5328.01, $p<.001$). Similarly, the 3PL model appeared appropriate over the 1PL and 2PL models for the 2018 and 2019 data sets. Comparing the model fit indicators with the varying levels of difficulty and guessing, it was more appropriate to use the 3PL model.

Table 12- *Model Comparison for 1PL, 2PL and 3PL Models*

Year	Contrast	Model	AIC	BIC	Log. Likelihood	LRT	df	*p*-value
2018	1PL vs. 2PL	1PL	1959862	1960206	-979891.1			
		2PL	1910038	1910726	-954938.8	49904.69	40	<.001
	2PL vs. 3PL	2PL	1910038	1910726	-954938.8			
		3PL	1904790	1905822	-952274.8	5328.01	40	<.001
	1PL vs. 3PL	1PL	1959862	1960206	-979891.0			
		3PL	1904790	1905822	-952274.8	55232.7	80	<.001
2019	1PL vs. 2PL	1PL	1806355	1806700	-903137.5			
		2PL	1762175	1762864	-881007.5	44260.05	40	<.001
	2PL vs. 3PL	2PL	1762175	1762864	-881007.5			
		3PL	1755698	1756731	-877728.9	6557.04	40	<.001
	1PL vs. 3PL	1PL	1806355	1806700	-903137.5			
		3PL	1755698	1756731	-877728.9	50817.09	80	<.001
2020	1PL vs. 2PL	1PL	1870618	1870964	-935269.0	58641.48		
		2PL	1812056	1812748	-905948.2		40	<.001
	2PL vs. 3PL	2PL	1812056	1812748	-905948.2			
		3PL	1807668	1808706	-903713.9	4468.7	40	<.001
	1PL vs. 3PL	1PL	1870618	1870964	-935269.0			
		3PL	1807668	1808706	-903713.9	63110.18	80	<.001

Research Question One

What are the characteristics of the JCE agriculture multiple choice organised by BEC based on the (a) difficulty parameter, (b) discrimination parameter, and (c) level of guessing?

This research question sought to assess the properties of the test items in the Agricultural examination organised by BEC focusing on three indicators, namely, difficulty, discrimination, and guessing. The three-year data were analysed separately with each year having its different item properties. Table 13 presents the details of the analysis for the 2018 batch of results.

As presented in Table 13, the guessing parameter ranged from 0 to .423, indicating that some of the items were susceptible to guessing for the 2018 JCE agriculture multiple choice. Comparing these guessing parameters with the acceptable random guessing factor of .25 for four option multiple choice items, five items (Q14, Q16, Q24, Q31, Q38) were found to be susceptible to guessing with guessing parameter values ranging from .272 to .423.

Most of the items were found to function in the higher-ability group. However, items Q2 (b=-4.449), Q7 (b=18.339), Q25 (b=14.076), and Q32 (b=-3.888) had difficult threshold values beyond the acceptable -3 to 3 range (Table 13). This anomaly suggests that these four items were problematic and should be checked, deleted or revised. Not surprisingly, these four items (i.e., Q2, Q7, Q25, and Q32) together with four other items (i.e., Q11, Q20, Q21, and Q33) showed poor discrimination indices. This was based on the cut-off point of a<.50 (Barker, 2001).

Table 13- *Item Parameters of 3PL for the Year 2018*

No.	Guessing (*c*)	Difficulty (*b*)	Discrimination (*a*)	P(x=1|z=0)
Q1	0.002	-0.213	0.751	0.541
Q2	0.018	-4.449	-0.097	0.405
Q3	0.010	-0.638	0.612	0.600
Q4	0.000	-0.797	1.224	0.726
Q5	0.000	-0.115	0.648	0.519
Q6	0.009	-0.077	0.959	0.523
Q7	0.078	18.339	0.054	0.327
Q8	0.125	2.419	0.990	0.199
Q9	0.207	1.709	2.292	0.222
Q10	0.147	-0.041	0.932	0.582
Q11	0.000	-1.338	0.489	0.658
Q12	0.211	1.058	1.841	0.310
Q13	0.000	-1.193	1.950	0.911
Q14	0.312	1.129	2.233	0.363
Q15	0.000	-0.560	0.994	0.636
Q16	0.282	2.488	1.683	0.293
Q17	0.113	1.018	1.517	0.269
Q18	0.235	0.311	2.063	0.499
Q19	0.188	-0.314	1.684	0.699
Q20	0.015	0.107	0.096	0.505
Q21	0.010	1.528	0.458	0.338
Q22	0.099	1.948	2.472	0.107
Q23	0.059	-1.286	1.151	0.826
Q24	0.423	-0.332	1.709	0.791
Q25	0.239	14.076	0.112	0.369
Q26	0.200	0.671	1.891	0.376
Q27	0.000	-1.649	2.041	0.967
Q28	0.214	2.769	2.073	0.216
Q29	0.137	-0.038	1.800	0.583
Q30	0.100	2.039	0.570	0.314
Q31	0.414	0.540	1.160	0.618
Q32	0.000	-3.888	0.265	0.737
Q33	0.002	-2.317	-0.134	0.424
Q34	0.081	-0.651	1.733	0.775
Q35	0.150	0.843	1.267	0.367
Q36	0.177	1.995	1.331	0.231
Q37	0.187	1.804	1.151	0.278
Q38	0.270	1.056	1.192	0.432
Q39	0.157	0.692	1.559	0.371
Q40	0.004	-0.544	0.762	0.604
Cut-off	(*c*) ≤ 0.25	(*b*) -3 – +3	(*a*) ≥ 0.5	

Two out of these eight non-functional items had a negative discrimination index. Table 14 presents the outcome of the analysis of the achievement of students in Agriculture for the year 2019.

Table 14- *Item Parameters of 3PL for the Year 2019*

No.	Guessing (c)	Difficulty (*b*)	Discrimination (*a*)	P(x=1|z=0)
Q1	0.249	0.598	1.805	0.439
Q2	0.111	0.581	0.885	0.444
Q3	0.140	0.499	1.919	0.379
Q4	0.270	0.964	2.933	0.311
Q5	0.000	-2.301	1.456	0.966
Q6	0.000	3.104	0.316	0.272
Q7	0.201	2.015	2.111	0.212
Q8	0.276	-0.516	1.137	0.741
Q9	0.000	-1.486	0.811	0.769
Q10	0.223	2.543	1.282	0.252
Q11	0.138	2.480	1.130	0.187
Q12	0.000	-1.340	0.901	0.770
Q13	0.176	-0.402	1.822	0.732
Q14	0.000	-1.753	0.853	0.817
Q15	0.077	1.669	2.451	0.092
Q16	0.120	2.308	2.249	0.125
Q17	0.115	0.260	1.854	0.453
Q18	0.081	-1.298	1.864	0.925
Q19	0.000	-0.995	0.158	0.539
Q20	0.232	0.087	1.412	0.592
Q21	0.000	1.114	0.349	0.404
Q22	0.170	-0.513	1.775	0.762
Q23	0.000	-1.982	2.426	0.992
Q24	0.000	-2.155	1.261	0.938
Q25	0.001	-0.810	1.377	0.753
Q26	0.116	-0.693	1.332	0.749
Q27	0.262	-0.197	1.835	0.697
Q28	0.271	-3.761	-1.538	0.273
Q29	0.008	-38.428	-0.026	0.275
Q30	0.104	0.017	1.415	0.547
Q31	0.000	-3.208	0.374	0.769
Q32	0.191	1.020	1.146	0.383
Q33	0.123	-3.195	-1.768	0.126
Q34	0.351	-0.441	2.097	0.816
Q35	0.119	-0.642	1.087	0.707
Q36	0.376	1.506	1.006	0.488
Q37	0.045	-0.494	1.167	0.657
Q38	0.241	0.914	1.501	0.394
Q39	0.258	1.316	0.985	0.418
Q40	0.191	0.750	1.111	0.436
Cut-off	$(c) \leq 0.25$	(b) -3 – +3	$(a) \geq 0.5$	

The outcome of the analysis, as shown in Table 14, depicts that 6 out of the 40 items appeared to be susceptible to guessing for the 2019 set of questions. Specifically, items Q8 (*c*=.276), Q27 (*c*=.262), Q28 (*c*=.271), Q34 (*c*=.351), Q36 (*c*=.376), and Q39 (*c*=.258) showed guessing parameter values greater than the recommended cut-off of .25. It was found that the difficulty of the items was fairly distributed between less difficult items and more difficult ones. However, 5 of the items (Q6, Q28, Q29, Q31, and Q33) exhibited threshold parameters outside the range of the acceptable range of -3 to 3. The data also showed that items Q6 (*a*=0.316), Q19 (*a*=0.158), Q21 (*a*=0.349), Q28 (*a*=-1.538), Q29 (*a*=-0.026), Q31 (*a*=0.374), and Q33 (*a*=-1.768) had poor item discrimination indices.

For the 2020-year BEC examination data, the details for the item parameters are shown in Table 15. The 2020 BEC examination data showed that 6 items were susceptible to guessing, they include Q2 (*c*=.259), Q13 (*c*=.478), Q14 (*c*=.303), Q26 (*c*=.253), Q30 (*c*=.296), and Q40 (*c*=.257). More than half of the items were considered less difficult because those items functioned at the lower ability of the examinees. Two of the items, however, were problematic as their difficulty indices were outside the threshold range of -3 to 3. Items Q11 (*a*=0.477), Q19 (*a*=-1.038), Q28 (*a*=0.217), Q32 (*a*=-1.283), Q35 (*a*=0.437), and Q37 (*a*=-0.380) had poor discrimination indices.

Table 15- *Item Parameters of 3PL for the Year 2020*

No.	Guessing (*c*)	Difficulty (*b*)	Discrimination (*a*)	P(x=1|z=0)
Q1	0.166	0.211	2.786	0.464
Q2	0.259	0.754	1.266	0.465
Q3	0.170	-0.073	2.185	0.618
Q4	0.001	-2.921	0.600	0.852
Q5	0.185	-0.266	1.592	0.678
Q6	0.000	-1.785	1.301	0.911
Q7	0.199	1.089	1.777	0.300
Q8	0.145	-0.564	1.316	0.724
Q9	0.213	0.083	1.602	0.580
Q10	0.133	-0.596	2.019	0.800
Q11	0.005	-0.564	0.477	0.569
Q12	0.171	-0.113	1.458	0.620
Q13	0.478	0.313	1.702	0.671
Q14	0.303	0.969	0.591	0.554
Q15	0.000	-1.061	1.367	0.810
Q16	0.000	-2.321	0.505	0.764
Q17	0.150	-0.850	1.601	0.826
Q18	0.122	1.542	1.517	0.199
Q19	0.132	-3.847	-1.038	0.148
Q20	0.235	-0.826	1.293	0.804
Q21	0.180	1.944	1.260	0.245
Q22	0.180	1.119	1.342	0.329
Q23	0.040	-1.523	2.222	0.969
Q24	0.040	-0.039	0.996	0.529
Q25	0.157	0.244	1.882	0.483
Q26	0.253	1.064	1.049	0.438
Q27	0.043	-1.182	1.343	0.837
Q28	0.011	-1.626	0.217	0.592
Q29	0.199	0.205	1.626	0.533
Q30	0.296	0.531	0.999	0.557
Q31	0.000	-1.749	1.281	0.904
Q32	0.000	-3.011	-1.283	0.021
Q33	0.006	-0.602	0.616	0.594
Q34	0.192	-1.201	2.673	0.969
Q35	0.005	-1.396	0.437	0.650
Q36	0.136	-0.143	1.898	0.626
Q37	0.001	-1.688	-0.380	0.346
Q38	0.124	-0.422	2.161	0.749
Q39	0.079	-0.912	0.818	0.704
Q40	0.257	-0.606	1.604	0.796
Cut-off	(*c*) ≤ 0.25	(*b*) -3 – +3	(*a*) ≥ 0.5	

Research Question Two

What is the contribution of the JCE agriculture multiple choice items organised by BEC to the measure of the student's achievement in the subject?

This research question examined how each item of the BEC Agricultural examination tests contributed to the measurement of students' achievement in the subject. To achieve this purpose, the item characteristic curves and item information curves were inspected (see Figures 11 – 13).

The ICCs for the items (in Figures 11 – 13) revealed that the items were fairly distributed across the ability continuum. While the majority of the items for each year were found to have good discrimination power, they only functioned best either among low-ability examines or the high-ability examinees. However, some items were found problematic and as such failed to accurately measure the construct along the ability continuum. For example, in 2018, items Q2, Q7, Q11, Q20, Q21, Q25, Q32, and Q33 failed to establish any meaningful relationship between the probability of correctly answering a question and the ability level of the examinees.

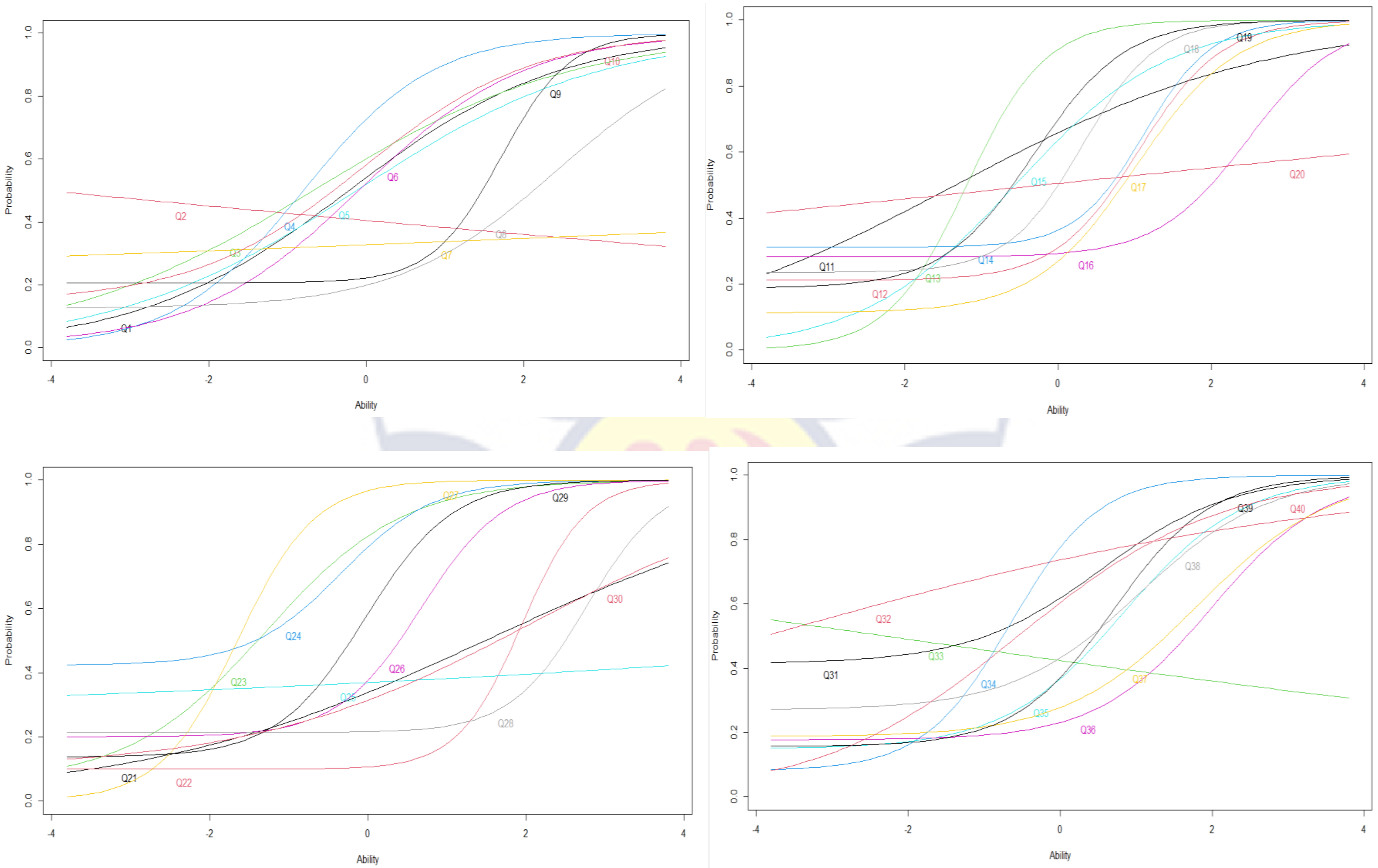


Figure 11- ICC's (3PL) for the 2018 Batch of Results for the 40-MC Items

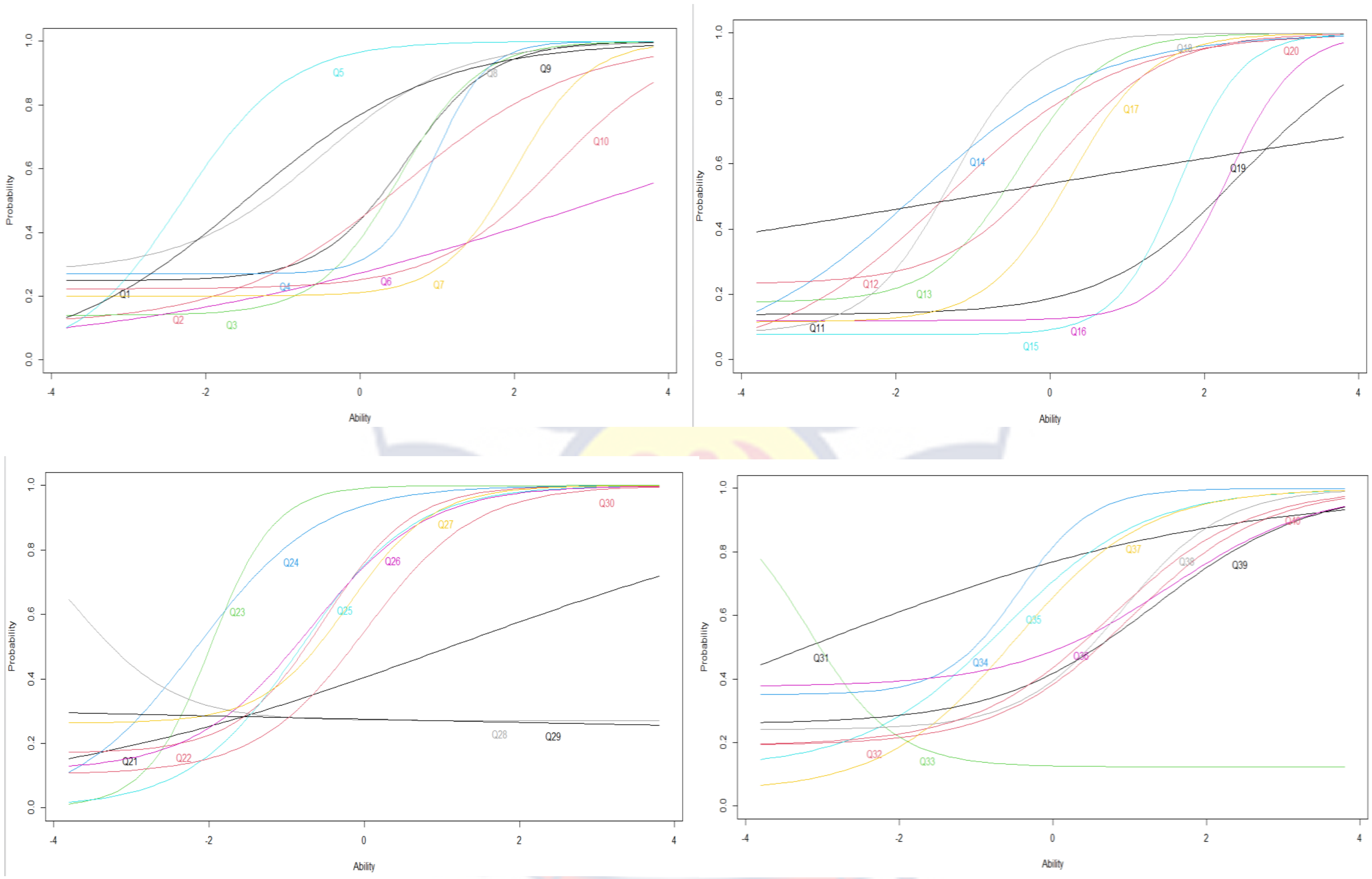


Figure 12- ICC's (3PL) for the 2019 Batch of Results for the 40-MC Items

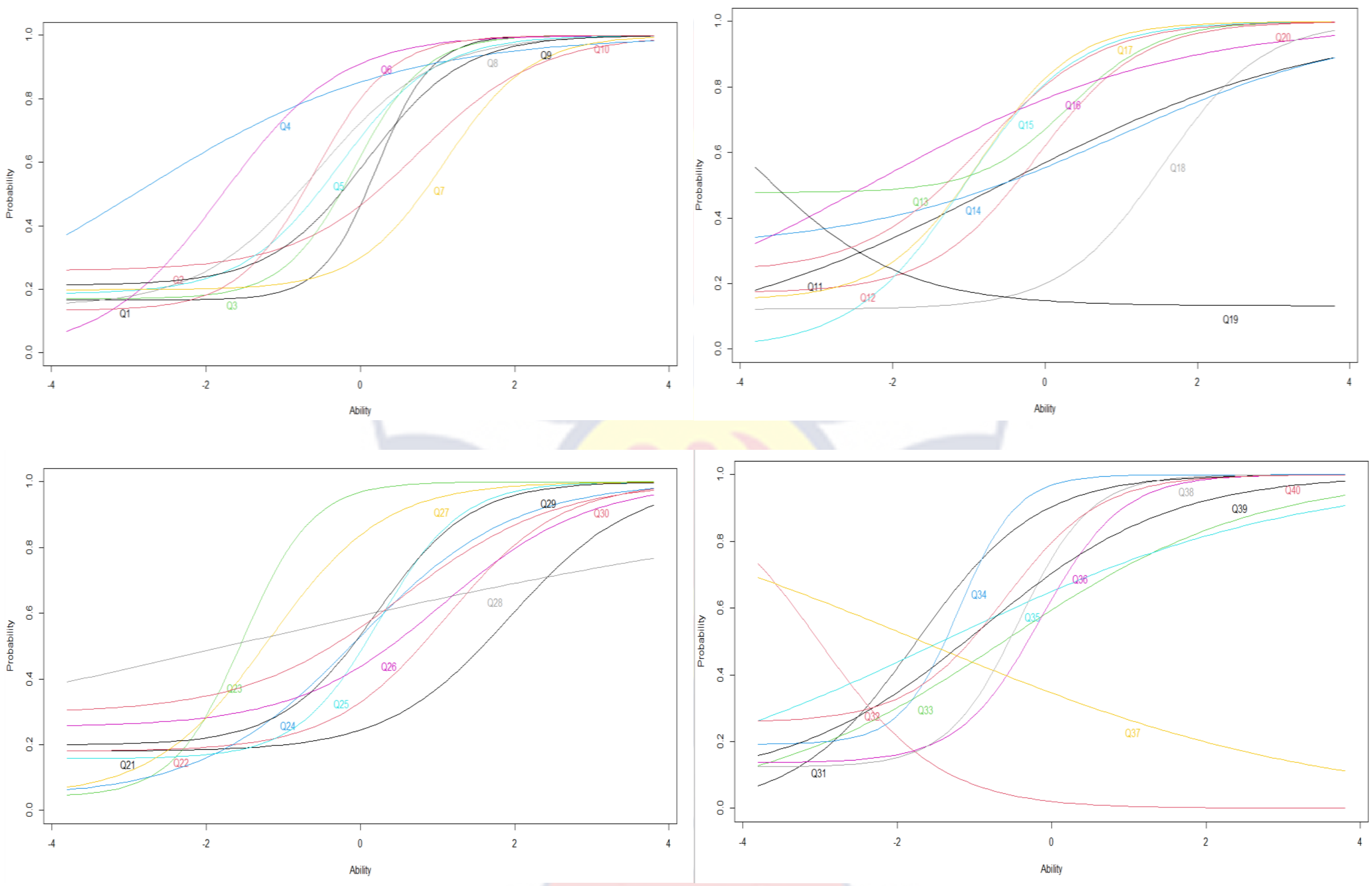


Figure 13- ICC's (3PL) for 2020 Batch of Results for the 40-MC Items

A similar pattern of results was discovered for the test items in the year 2019 (i.e., Q6, Q19, Q21, Q28, Q29, Q31, and Q33) and 2020 (i.e., Q19, Q28, Q37, Q32, and Q35). Some of these items had very poor ICCs to the extent that a negative relationship was revealed between the probability of correctly answering a question and the ability level of the examinees. Examples of items are items Q2 and Q33 in 2018, items Q28, Q29, and Q33 in 2019, and items Q19, Q37 and Q32 in 2020. Given the ICCs, it appears these problematic items are contributing very little to the measurement of examinees' ability in the Agriculture subject.

The inspection of the IIF for the items across 2018 to 2020 showed that whereas some items contributed significantly to the measure of students' ability, other items failed to contribute meaningfully to the measurement. For the three years, the items had varying levels of contribution of empirical information to the measurement of the construct under investigation. For the year 2019, item Q22 contributed the largest empirical information to the measurement of students' ability in Agriculture (see Figure 14). Following this, 8 of the items (Q2, Q7, Q11, Q20, Q21, Q25, Q32, and Q33) provided very little information in measuring the achievement construct.

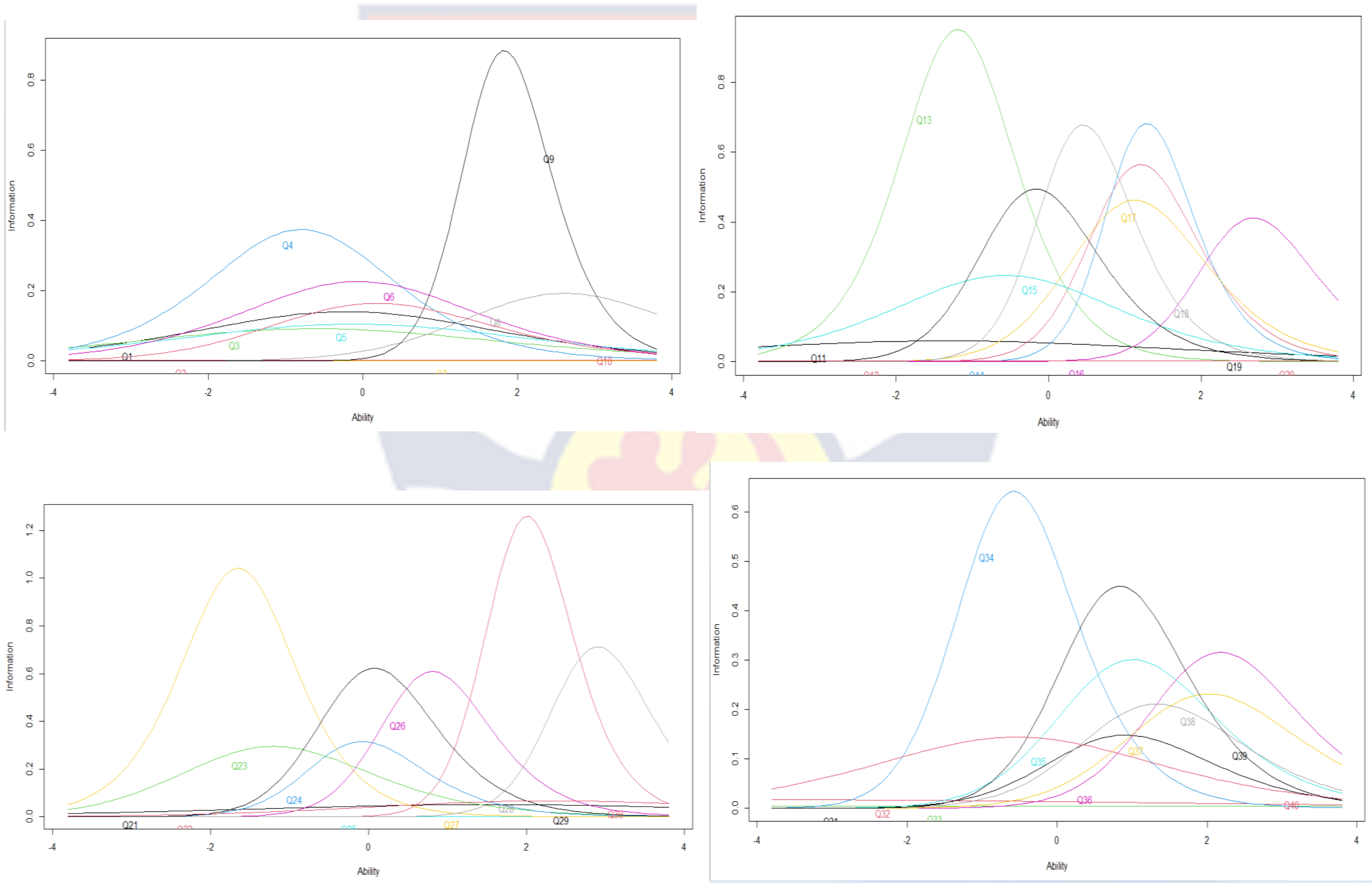


Figure 14- Item Information Function for 40-MC Items in 2018

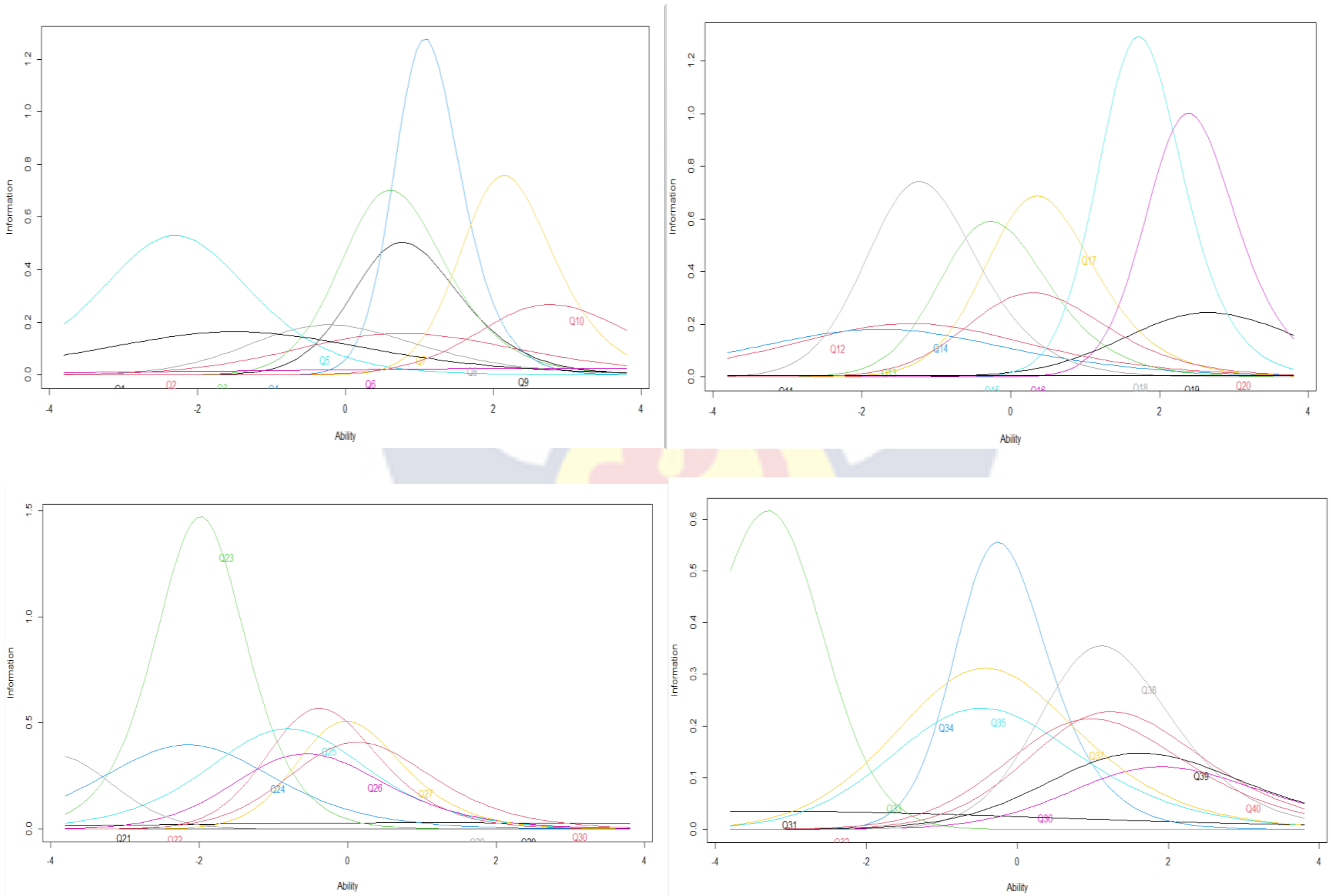


Figure 15- Item Information Function for 40-MC Items in 2019

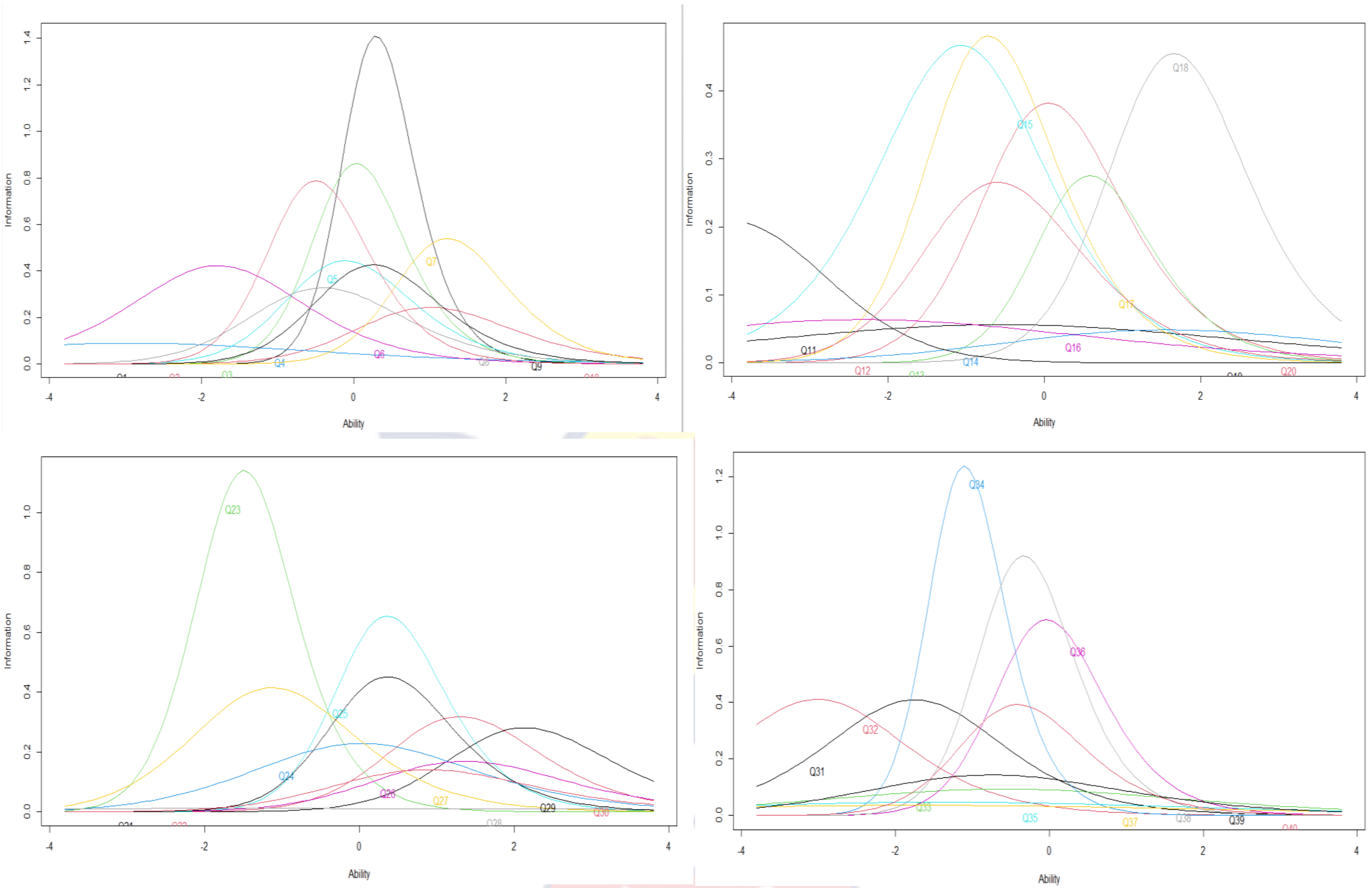


Figure 16- Item Information Function for 40-MC Items in 2020

In 2019, whereas item 15 had the highest empirical information, items Q6, Q19, Q21, Q28, Q29, Q31, and Q33 had little contribution to the measure of students' ability (Figure 15). For the year 2020, item Q1 had the smallest measurement error in quantifying students' ability while items Q11, Q19, Q28, Q32, Q32, and Q37 had the largest measurement error in the measurement of students' traits in Agriculture (Figure 16).

Research Question Three

What is the level of differential item functioning (DIF) of JCE agriculture multiple choice items administrated by BEC in terms of sex?

The study further examined the extent of DIF of the test items administered by BEC in terms of gender. Before the DIF analysis, measurement invariance for gender was tested across three years (i.e., 2018-2020). The measurement invariance hypothesis was satisfied by conducting the multiple indicator confirmatory factor analysis (see Appendix F). The DIF analyses were performed using the Raju area method. It is important to point out that the study only focused on items which were flagged as having large DIF (Appendix G). The details of the analysis are shown in Tables 16 to 18. Items which were flagged for DIF were matched with assessment syllabus to identify DIF source.

Results as presented in Table 16 showed that for the 2018 batch of data, 7 items (Q3, Q11, Q21, Q23, Q24, Q32, and Q40) were considered as gender DIF items with large effect sizes. Out of the 7 items, 6 of them showed uniform DIF whereas one showed non-uniform DIF (i.e. Q3). For the items which showed uniform DIF, 4 of them favoured male students whereas 2 items favoured female students.

Table 16- *Gender DIF of MC Items of BEC from 2018*

No.	Z-Statistics	p-value	Effect size	Remark	Group favoured
Q1	-84.9156	0.0000 ***	B	No DIF	---
Q2	-4.5230	0.0000 ***	A	No DIF	---
Q3	-84.6239	0.0000 ***	C	DIF	Non-uniform
Q4	-139.0196	0.0000 ***	A	No DIF	---
Q5	-76.6692	0.0000 ***	B	No DIF	---
Q6	-97.6571	0.0000 ***	B	No DIF	---
Q7	0.3205	0.7486	---	No DIF	---
Q8	7.8630	0.0000 ***	B	No DIF	---
Q9	5.8542	0.0000 ***	A	No DIF	---
Q10	-95.0691	0.0000 ***	B	No DIF	---
Q11	-73.3932	0.0000 ***	C	DIF	Male
Q12	-24.5724	0.0000 ***	A	No DIF	---
Q13	-170.3358	0.0000 ***	A	No DIF	---
Q14	-29.3799	0.0000 ***	A	No DIF	---
Q15	-123.7884	0.0000 ***	B	No DIF	---
Q16	5.2185	0.0000 ***	A	No DIF	---
Q17	-34.6626	0.0000 ***	B	No DIF	---
Q18	-117.1942	0.0000 ***	B	No DIF	---
Q19	-148.9238	0.0000 ***	A	No DIF	---
Q20	0.2827	0.7774	----	No DIF	---
Q21	-4.7567	0.0000 ***	C	DIF	Female
Q22	7.1502	0.0000 ***	A	No DIF	---
Q23	-112.8999	0.0000 ***	C	DIF	Male
Q24	-109.8000	0.0000 ***	C	DIF	Male
Q25	-0.1566	0.8756	----	No DIF	---
Q26	-67.8566	0.0000 ***	B	No DIF	---
Q27	-160.0909	0.0000 ***	A	No DIF	---
Q28	3.7010	0.0002 ***	A	No DIF	---
Q29	-143.9365	0.0000 ***	B	No DIF	---
Q30	0.1009	0.9197	---	No DIF	---
Q31	-80.0888	0.0000 ***	B	No DIF	---
Q32	-20.5072	0.0000 ***	C	DIF	Male
Q33	-3.2691	0.0011	B	No DIF	---
Q34	-161.5018	0.0000 ***	A	No DIF	---
Q35	-42.7420	0.0000 ***	B	No DIF	---
Q36	2.5643	0.0103*	A	No DIF	---
Q37	-1.3707	0.1705	-----	No DIF	---
Q38	-42.7157	0.0000 ***	B	No DIF	---
Q39	-56.4434	0.0000 ***	B	No DIF	---
Q40	-96.7713	0.0000 ***	C	DIF	Female

Effect Size: A- negligible, B- moderate, C- large
Focal group= Male; Reference group=Female

The DIF results for the 2019 year BEC examination items are also presented in Table 17.

Table 17- *Sex DIF of MC Items of BEC from 2019*

No.	Z-Statistics	p-value	Effect size	Status	Group favoured
Q1	2496.9306	0.0000 ***	B	No DIF	---
Q2	1905.9155	0.0000 ***	C	DIF	Female
Q3	2984.6323	0.0000 ***	B	No DIF	---
Q4	1960.9703	0.0000 ***	B	No DIF	---
Q5	1028.7907	0.0000 ***	A	No DIF	---
Q6	250.4898	0.0000 ***	C	DIF	Male
Q7	266.8119	0.0000 ***	A	No DIF	---
Q8	1512.1762	0.0000 ***	C	DIF	Female
Q9	1132.7906	0.0000 ***	A	No DIF	---
Q10	86.1706	0.0000 ***	A	No DIF	---
Q11	301.8550	0.0000 ***	A	No DIF	---
Q12	1566.8783	0.0000 ***	B	No DIF	---
Q13	2771.0920	0.0000 ***	A	No DIF	---
Q14	1086.8557	0.0000 ***	B	No DIF	---
Q15	1402.1217	0.0000 ***	A	No DIF	---
Q16	170.2242	0.0000 ***	A	No DIF	---
Q17	3193.4729	0.0000 ***	B	No DIF	---
Q18	2098.1737	0.0000 ***	A	No DIF	---
Q19	430.7812	0.0000 ***	C	DIF	Male
Q20	2377.6610	0.0000 ***	B	No DIF	---
Q21	579.8569	0.0000 ***	C	DIF	Male
Q22	2658.7426	0.0000 ***	B	No DIF	---
Q23	1411.9698	0.0000 ***	A	No DIF	---
Q24	855.6864	0.0000 ***	A	No DIF	---
Q25	2556.8913	0.0000 ***	B	No DIF	---
Q26	2359.1034	0.0000 ***	B	No DIF	---
Q27	2571.8662	0.0000 ***	B	No DIF	---
Q28	0.0000	.9999	---	No DIF	---
Q29	0.0000	.9999	---	No DIF	---
Q30	2876.1211	0.0000 ***	B	No DIF	---
Q31	234.9568	0.0000 ***	A	No DIF	---
Q32	1483.6352	0.0000 ***	A	No DIF	---
Q33	6.9642	0.0000 ***	A	No DIF	---
Q34	2182.6032	0.0000 ***	B	No DIF	---
Q35	1952.0469	0.0000 ***	B	No DIF	---
Q36	1113.4004	0.0000 ***	B	No DIF	---
Q37	2454.2698	0.0000 ***	B	No DIF	---
Q38	1901.0912	0.0000 ***	B	No DIF	---
Q39	1238.1385	0.0000 ***	B	No DIF	---
Q40	1954.2785	0.0000 ***	A	No DIF	---

Effect Size: A- negligence, B- moderate, C- large
Focal group= Male; Reference group=Female

The results from the gender DIF for the items in the 2019 BEC examination (see Table 17). It was found that 5 out of the 40 items were flagged as uniform DIF with large effect sizes (i.e., Q2, Q6, Q8, Q19, and

Q21). Whereas 3 of the DIF items favoured males (i.e., Q6, Q19, Q21), 2 of the DIF items favoured female students (i.e., Q2, Q8).

The outcome of the DIF analysis for the year 2020 regarding BEC examination items are shown in Table 18. For the 2020 year BEC examination results, 8 of the items (i.e., Q14, Q26, Q28, Q30, Q31, Q34, Q35, and Q36) were identified as gender DIF items with moderate effect size, with 4 favouring female students (i.e., Q4, Q30, Q31, Q38), 3 favouring male students (i.e., Q26, Q28, Q37) and 1 non-uniform DIF (i.e., Q34).

Table 18- *Gender DIF of MC Items of BEC from 2020*

No.	Z-Statistics	p-value	Effect size	Status	Group favoured
Q1	-113.0960	0.0000 ***	B	No DIF	---
Q2	-59.1830	0.0000 ***	B	No DIF	---
Q3	-74.3641	0.0000 ***	B	No DIF	---
Q4	12.3334	0.0000 ***	A	No DIF	---
Q5	-43.1151	0.0000 ***	B	No DIF	---
Q6	13.3470	0.0000 ***	A	No DIF	---
Q7	-86.5088	0.0000 ***	A	No DIF	---
Q8	-12.4828	0.0000 ***	B	No DIF	---
Q9	-64.5915	0.0000 ***	B	No DIF	---
Q10	-27.3806	0.0000 ***	B	No DIF	---
Q11	-36.8935	0.0000 ***	A	No DIF	---
Q12	-58.5373	0.0000 ***	B	No DIF	---
Q13	-5.1945	0.0000 ***	B	No DIF	---
Q14	-9.3622	0.0000 ***	C	DIF	Female
Q15	0.8157	0.4146	---	No DIF	---
Q16	9.2290	0.0000 ***	A	No DIF	---
Q17	-6.4850	0.0000 ***	A	No DIF	---
Q18	-75.4907	0.0000 ***	B	No DIF	---
Q19	3.5092	0.0004 ***	A	No DIF	---
Q20	-7.8345	0.0000 ***	B	No DIF	---
Q21	-45.7500	0.0000 ***	A	No DIF	---
Q22	-81.9318	0.0000 ***	B	No DIF	---
Q23	7.7135	0.0000 ***	A	No DIF	---
Q24	-69.6991	0.0000 ***	B	No DIF	---
Q25	-84.0933	0.0000 ***	B	No DIF	---
Q26	-53.9401	0.0000 ***	C	DIF	Male
Q27	-3.1085	0.0019	---	No DIF	---
Q28	-5.9518	0.0000 ***	C	DIF	Male
Q29	-71.6503	0.0000 ***	B	No DIF	---
Q30	-29.2159	0.0000 ***	C	DIF	Female
Q31	15.5675	0.0000 ***	C	DIF	Female
Q32	13.2447	0.0000 ***	A	No DIF	---
Q33	-25.8029	0.0000 ***	B	No DIF	---
Q34	7.2255	0.0000 ***	C	DIF	Non-uniform
Q35	0.1038	0.9173	---	No DIF	---
Q36	-58.4986	0.0000 ***	B	No DIF	---
Q37	0.0960	0.9235	C	DIF	Male
Q38	-35.3714	0.0000 ***	C	DIF	Female
Q39	0.0642	0.9488	---	No DIF	---
Q40	-1.3430	0.1793	---	No DIF	---

Effect Size: A- negligible, B- moderate, C- large
Focal group= Male; Reference group=Female

Research Question Four

What is the level of differential item functioning of the JCE agriculture multiple choice items administered by BEC with reference to school location?

This research question examined the level of DIF of the items administered by BEC with regard to school type. Just like the previous research question, the DIF analysis was performed using the Raju area method (Appendix H) after the measurement invariance assumption was satisfied (see Appendix F). Tables 19 to 21 highlight the details of the analyses.

As presented in Table 19, 3 items were identified as DIF with regard to school location in the 2018 BEC examination. Whereas 2 items favoured students from rural schools (i.e., Q21, Q24) and 1 item favoured students from urban schools (Q32). The outcome of the DIF analyses with reference to school location for the 2019 batch of results is shown in Table 20.

Table 19- *School Location DIF Analysis of the 2018 BEC Examination*

No.	Z-Statistics	*p*-value	Effect size	Status	Group favoured
Q1	3.0418	0.0024 **	B	No DIF	---
Q2	0.5807	0.5614	---	No DIF	---
Q3	1.3582	0.1744	---	No DIF	---
Q4	-0.3274	0.7434	---	No DIF	---
Q5	10.2519	0.0000 ***	B	No DIF	---
Q6	5.7298	0.0000 ***	B	No DIF	---
Q7	0.0032	0.9974	---	No DIF	---
Q8	4.4257	0.0000 ***	B	No DIF	---
Q9	5.9884	0.0000 ***	A	No DIF	---
Q10	3.9875	0.0001 ***	B	No DIF	---
Q11	0.5667	0.5709	---	No DIF	---
Q12	12.6312	0.0000 ***	B	No DIF	---
Q13	-0.9027	0.3667	---	No DIF	---
Q14	11.1439	0.0000 ***	A	No DIF	---
Q15	8.9976	0.0000 ***	B	No DIF	---
Q16	3.9335	0.0001 ***	A	No DIF	---
Q17	12.3380	0.0000 ***	B	No DIF	---
Q18	9.6038	0.0000 ***	B	No DIF	---
Q19	5.1660	0.0000 ***	B	No DIF	---
Q20	1.4100	0.1585	----	No DIF	---
Q21	6.0786	0.0000 ***	C	DIF	Rural
Q22	6.2705	0.0000 ***	A	No DIF	---
Q23	-1.2949	0.1954	---	No DIF	---
Q24	4.1153	0.0000 ***	C	DIF	Rural
Q25	- 0.4198	0.6746	----	No DIF	---
Q26	8.5048	0.0000 ***	B	No DIF	---
Q27	-3.6379	0.0003 ***	B	No DIF	---
Q28	0.5011	0.6163	---	No DIF	---
Q29	5.6071	0.0000 ***	B	No DIF	---
Q30	3.5349	0.0004 ***	B	No DIF	---
Q31	0.5047	0.6138	---	No DIF	---
Q32	-2.6767	0.0074 **	C	DIF	Urban
Q33	-0.5166	0.6055	---	No DIF	---
Q34	-1.3986	0.1619	---	No DIF	---
Q35	4.9362	0.0000 ***	B	No DIF	---
Q36	3.2345	0.0012 **	B	No DIF	---
Q37	4.1465	0.0000 ***	B	No DIF	---
Q38	7.2789	0.0000 ***	B	No DIF	---
Q39	8.2348	0.0000 ***	B	No DIF	---
Q40	-1.1102	0.2669	---	No DIF	---

Effect Size: A- negligence, B- moderate, C- large

Focal group= Urban; Reference group=Rural

Table 20- *School Location DIF Analysis of the 2019 BEC Examination*

No.	Z-Statistics	p-value	Effect size	Status	Group favoured
Q1	30.8186	0.0000 ***	B	No DIF	---
Q2	12.2290	0.0005 ***	C	DIF	Rural
Q3	8.2768	0.0040 **	B	No DIF	---
Q4	5.2636	0.0218 *	B	No DIF	---
Q5	0.0447	0.8325	---	No DIF	---
Q6	7.0796	0.0078 **	C	DIF	Non-uniform
Q7	0.6780	0.4103	---	No DIF	---
Q8	10.9972	0.0009 ***	B	No DIF	---
Q9	2.7975	0.0944	---	No DIF	---
Q10	7.2668	0.0070 **	A	No DIF	---
Q11	1.5232	0.2171	---	No DIF	---
Q12	210.2101	0.0000 ***	B	No DIF	---
Q13	3.1219	0.0772	---	No DIF	---
Q14	4.3381	0.0373 *	B	No DIF	---
Q15	5.0845	0.0241 *	A	No DIF	---
Q16	1.4320	0.2314	---	No DIF	---
Q17	11.9890	0.0005 ***	B	No DIF	---
Q18	9.9930	0.0016 **	A	No DIF	---
Q19	5.0390	0.0248 *	C	DIF	Urban
Q20	2.6315	0.1048	---	No DIF	---
Q21	5.5841	0.0181 *	C	DIF	Urban
Q22	2.1453	0.1430	---	No DIF	---
Q23	10.7528	0.0010 ***	A	No DIF	---
Q24	1.2041	0.2725	---	No DIF	---
Q25	6.7990	0.0091 **	B	No DIF	---
Q26	9.4514	0.0021 **	B	No DIF	---
Q27	2.1678	0.1409	---	No DIF	---
Q28	10.8151	0.0010 ***	A	No DIF	---
Q29	10.8976	0.0010 ***	A	No DIF	---
Q30	5.8436	0.0156 *	B	No DIF	---
Q31	26.9191	0.0000 ***	A	No DIF	---
Q32	0.1682	0.6817	---	No DIF	---
Q33	0.0733	0.7867	---	No DIF	---
Q34	6.3698	0.0116 *	B	No DIF	---
Q35	10.4784	0.0012 **	B	No DIF	---
Q36	0.2221	0.6374	---	No DIF	---
Q37	1.3919	0.2381	---	No DIF	---
Q38	0.0637	0.8007	---	No DIF	---
Q39	12.3676	0.0004 ***	B	No DIF	---
Q40	19.4735	0.0000 ***	A	No DIF	---

A- negligible, B- moderate, C- large
Focal group= Urban; Reference group=Rural

The outcome of the DIF analysis with regards to school location revealed that 4 items exhibited DIF with a large effect size; 1 item favoured students in rural schools, 2 items favoured urban schools, and 1 item was found as non-uniform DIF. The results for 2020 DIF analysis with reference to school location is presented in Table 21.

Table 21- *School Location DIF Analysis of the 2020 BEC Examination*

No.	Z-Statistics	p-value	Effect size	Status	Group favoured
Q1	-17.1452	0.0000 ***	B	No DIF	---
Q2	-16.1151	0.0000 ***	B	No DIF	---
Q3	-6.0834	0.0000 ***	B	No DIF	---
Q4	10.5908	0.0000 ***	A	No DIF	---
Q5	1.0039	0.3154	--	No DIF	---
Q6	15.3151	0.0000 ***	A	No DIF	---
Q7	-22.4022	0.0000 ***	A	No DIF	---
Q8	5.1225	0.0000 ***	B	No DIF	---
Q9	-3.6748	0.0002 ***	B	No DIF	---
Q10	5.6082	0.0000 ***	B	No DIF	---
Q11	-1.6944	0.0902	--	No DIF	---
Q12	-4.7022	0.0000 ***	B	No DIF	---
Q13	8.4155	0.0000 ***	B	No DIF	---
Q14	-0.6344	0.5258	--	No DIF	---
Q15	9.9484	0.0000 ***	A	No DIF	---
Q16	6.3988	0.0000 ***	A	No DIF	---
Q17	10.4671	0.0000 ***	A	No DIF	---
Q18	-26.1384	0.0000 ***	B	No DIF	---
Q19	-0.0040	0.9968	---	No DIF	---
Q20	11.9006	0.0000 ***	B	No DIF	---
Q21	-16.9429	0.0000 ***	A	No DIF	---
Q22	-23.4792	0.0000 ***	B	No DIF	---
Q23	21.0633	0.0000 ***	A	No DIF	---
Q24	-11.0152	0.0000 ***	B	No DIF	---
Q25	-22.5354	0.0000 ***	B	No DIF	---
Q26	-15.4023	0.0000 ***	C	DIF	Urban
Q27	12.5973	0.0000 ***	A	No DIF	---
Q28	0.9795	0.3273	---	No DIF	---
Q29	-8.6258	0.0000 ***	B	No DIF	---
Q30	-0.6151	0.5385	---	No DIF	---
Q31	15.0831	0.0000 ***	C	DIF	Urban
Q32	13.5272	0.0000 ***	A	No DIF	---
Q33	3.3378	0.0008 ***	B	No DIF	---
Q34	21.2654	0.0000 ***	C	DIF	Urban
Q35	4.5132	0.0000 ***	B	No DIF	---
Q36	-8.1100	0.0000 ***	B	No DIF	---
Q37	8.5468	0.0000 ***	C	DIF	Urban
Q38	2.4565	0.0140 *	A	No DIF	---
Q39	6.3892	0.0000 ***	B	No DIF	---
Q40	10.4977	0.0000 ***	C	DIF	Urban

Effect Size: A- negligible, B- moderate, C- large
Focal group= Urban; Reference group=Rural

The outcome of the results for DIF regarding school location showed that 5 items were flagged as DIF (see Table 21). All the items which showed large effect sizes were in favour of the students in urban schools.

Research Question Five

What is the level of differential item functioning of the JCE agriculture multiple choice items administered by BEC with reference to school type (public and private)?

This research question determined the extent to which the test items administered by BEC are susceptible to DIF with regard to school type. Satisfying the measurement invariance hypothesis (see Appendix F), DIF analyses were performed using the Raju area method (Appendix I). Subject matter content area for items with large DIF effect sizes were probed in relation to practices done in both public and private schools. The details of the DIF analyses are shown in Tables 22 – 24.

The results, as shown in Table 22, revealed that some of the items for the 2018 data set were flagged as DIF for school type. Among the items, 9 of them (i.e., Q3, Q11, Q18, Q19, Q21, Q24, Q27, Q32, and Q40) were identified as DIF with a large effect size, with 4 items favouring students from public schools, 1 item favouring students in private schools, and 4 items had non-uniform DIF.

Table 22- *School Type DIF of MC Items of BEC 2018*

No.	Z-Statistics	p-value	Effect size	Status	Group favoured
Q1	-15.3574	0.0000 ***	B	No DIF	---
Q2	-0.7788	0.4361	---	No DIF	---
Q3	-8.7462	0.0000 ***	C	DIF	Public
Q4	-33.1116	0.0000 ***	A	No DIF	---
Q5	-12.9703	0.0000 ***	B	No DIF	---
Q6	-13.8358	0.0000 ***	B	No DIF	---
Q7	0.8757	0.3812	---	No DIF	---
Q8	9.4580	0.0000 ***	B	No DIF	---
Q9	-0.4306	0.6668	A	No DIF	---
Q10	-27.4304	0.0000 ***	B	No DIF	---
Q11	-20.6688	0.0000 ***	C	DIF	Non-uniform
Q12	-1.9155	0.0554	---	No DIF	---
Q13	-21.8659	0.0000 ***	A	No DIF	---
Q14	-1.8674	0.0618	---	No DIF	---
Q15	-23.5940	0.0000 ***	B	No DIF	---
Q16	-0.1396	0.8890	---	No DIF	---
Q17	-0.2029	0.8392	---	No DIF	---
Q18	-19.2013	0.0000 ***	C	DIF	Public
Q19	-47.1887	0.0000 ***	C	DIF	Public
Q20	1.4863	0.1372	----	No DIF	---
Q21	9.4129	0.0000 ***	C	DIF	Public
Q22	5.6195	0.0000 ***	A	No DIF	---
Q23	-37.6014	0.0000 ***	A	No DIF	---
Q24	-56.5108	0.0000 ***	C	DIF	Non-uniform
Q25	1.6239	0.1044	----	No DIF	---
Q26	-7.9645	0.0000 ***	B	No DIF	---
Q27	-32.6170	0.0000 ***	C	DIF	Non-uniform
Q28	8.7950	0.0000 ***	A	No DIF	---
Q29	-34.9544	0.0000 ***	B	No DIF	---
Q30	7.2378	0.0000 ***	B	No DIF	---
Q31	-28.8598	0.0000 ***	B	No DIF	---
Q32	-3.6872	0.0002 ***	C	DIF	Private
Q33	-1.2774	0.2015	---	No DIF	---
Q34	-55.7885	0.0000 ***	A	No DIF	---
Q35	-2.9898	0.0028 **	B	No DIF	---
Q36	3.5939	0.0003 ***	B	No DIF	---
Q37	8.4278	0.0000 ***	B	No DIF	---
Q38	-2.8166	0.0049 **	B	No DIF	---
Q39	-4.5561	0.0000 ***	B	No DIF	---
Q40	-2.6341	0.0084 **	C	DIF	Non-uniform

Effect Size: A- negligible, B- moderate, C- large

Focal group= Public; Reference group=Private

Table 23, presents the results output of DIF concerning school type for JCE agriculture multiple choice items for the year 2019.

Table 23- *School Type DIF of MC Items of BEC 2019*

No.	Z-Statistics	p-value	ES	Status	Group favoured
Q1	164.7765	0.0000 ***	B	No DIF	---
Q2	267.8438	0.0000 ***	C	DIF	Non-uniform
Q3	859.4688	0.0000 ***	B	No DIF	---
Q4	268.0798	0.0000 ***	B	No DIF	---
Q5	117.2414	0.0000 ***	A	No DIF	---
Q6	110.0004	0.0000 ***	C	DIF	Public
Q7	20.5505	0.0000 ***	A	No DIF	---
Q8	529.4389	0.0000 ***	B	No DIF	---
Q9	274.6414	0.0000 ***	A	No DIF	---
Q10	63.8295	0.0000 ***	A	No DIF	---
Q11	62.0251	0.0000 ***	A	No DIF	---
Q12	788.4921	0.0000 ***	B	No DIF	---
Q13	1146.1015	0.0000 ***	A	No DIF	---
Q14	140.4486	0.0000 ***	B	No DIF	---
Q15	237.0792	0.0000 ***	A	No DIF	---
Q16	27.5541	0.0000 ***	A	No DIF	---
Q17	788.3225	0.0000 ***	B	No DIF	---
Q18	503.8972	0.0000 ***	C	DIF	Public
Q19	52.1971	0.0000 ***	C	DIF	Public
Q20	1082.1182	0.0000 ***	B	No DIF	
Q21	225.1918	0.0000 ***	C	DIF	Public
Q22	1277.3066	0.0000 ***	B	No DIF	---
Q23	458.9868	0.0000 ***	C	DIF	Non-uniform
Q24	187.0494	0.0000 ***	A	No DIF	---
Q25	720.8329	0.0000 ***	B	No DIF	---
Q26	1329.4653	0.0000 ***	B	No DIF	---
Q27	1474.1963	0.0000 ***	B	No DIF	---
Q28	0.7785	0.4363	A	No DIF	---
Q29	-0.1877	0.8511	A	No DIF	---
Q30	1042.0361	0.0000 ***	B	No DIF	---
Q31	84.4973	0.0000 ***	A	No DIF	---
Q32	281.3691	0.0000 ***	A	No DIF	---
Q33	-0.0213	0.9830	A	No DIF	---
Q34	918.0110	0.0000 ***	B	No DIF	---
Q35	731.5656	0.0000 ***	B	No DIF	---
Q36	90.1157	0.0000 ***	B	No DIF	---
Q37	1327.7992	0.0000 ***	B	No DIF	---
Q38	250.7233	0.0000 ***	B	No DIF	---
Q39	170.7803	0.0000 ***	B	No DIF	---
Q40	327.4885	0.0000 ***	A	No DIF	---

Effect Size: A- negligible, B- moderate, C- large
Focal group= Public; Reference group=Private

Out of the 40 items for the 2019 BEC examination, 6 items (Q2, Q6, Q18, Q19, Q21, and Q23) were identified as DIF for school type with large effect sizes. Of these 6 items, 4 of them favoured students from public schools, 2 items had non-uniform DIF and none of these 6 items favoured private school students. Table 24 further outlines the output of the DIF analysis with respect to school type for the 2020 BEC examination.

Table 24- *School Type DIF of MC Items of BEC 2020*

No.	Z-Statistics	p-value	ES	Status	Group Favoured
Q1	59.9724	0.0000 ***	B	No DIF	---
Q2	82.9950	0.0000 ***	B	No DIF	---
Q3	78.9253	0.0000 ***	B	No DIF	---
Q4	4.7555	0.0000 ***	A	No DIF	---
Q5	51.0395	0.0000 ***	B	No DIF	---
Q6	13.3890	0.0000 ***	A	No DIF	---
Q7	19.7060	0.0000 ***	A	No DIF	---
Q8	41.8400	0.0000 ***	B	No DIF	---
Q9	65.6893	0.0000 ***	B	No DIF	---
Q10	58.3915	0.0000 ***	B	No DIF	---
Q11	33.8984	0.0000 ***	A	No DIF	---
Q12	62.8805	0.0000 ***	B	No DIF	---
Q13	17.6240	0.0000 ***	B	No DIF	---
Q14	21.9007	0.0000 ***	C	DIF	Public
Q15	32.8032	0.0000 ***	A	No DIF	
Q16	8.5349	0.0000 ***	A	No DIF	
Q17	34.2732	0.0000 ***	A	No DIF	
Q18	6.8778	0.0000 ***	B	DIF	
Q19	0.0000	1.0000	---	No DIF	
Q20	33.3937	0.0000 ***	B	No DIF	
Q21	0.0000	1.0000	---	No DIF	
Q22	18.9891	0.0000 ***	B	DIF	
Q23	24.2187	0.0000 ***	A	No DIF	
Q24	57.1388	0.0000 ***	B	No DIF	
Q25	44.3276	0.0000 ***	B	No DIF	
Q26	6.5555	0.0000 ***	C	DIF	Public
Q27	40.1680	0.0000 ***	A	No DIF	
Q28	4.8347	0.0000 ***	C	DIF	Private
Q29	59.2710	0.0000 ***	B	No DIF	
Q30	26.8392	0.0000 ***	C	DIF	Public
Q31	16.6792	0.0000 ***	A	No DIF	
Q32	8.3747	0.0000 ***	C	DIF	Non-uniform
Q33	42.4219	0.0000 ***	B	No DIF	
Q34	44.0459	0.0000 ***	C	DIF	Non-uniform
Q35	12.2590	0.0000 ***	B	No DIF	
Q36	62.1479	0.0000 ***	B	No DIF	
Q37	4.2380	0.0000 ***	C	DIF	Non-uniform
Q38	75.9294	0.0000 ***	C	DIF	Non-uniform
Q39	22.3766	0.0000 ***	B	No DIF	
Q40	27.5038	0.0000 ***	B	No DIF	

Effect Size: A- negligible, B- moderate, C- large

Focal group= Public; Reference group=Private

The 2020 BEC examination items saw 8 items flagged as DIF with the large effect size for school type. Out of these 8 items, 3 items favoured students from public schools, an item favoured those from private schools and 4 items showed as non-uniform DIF (Table 24).

Discussion

Qualities of the JCE Agriculture Multiple Choice Items for 2018-2020.

The findings from the study showed that although there are items that were functional and accurately measured the construct of student ability in the agriculture subject, there were challenges identified with some of the items. First, between 12.5% and 15% of the items were prone to guessing for 2018, 2019, and 2020 BEC examinations. Expectedly, the tendency of students to guess on a multiple test item is very high since it offers the students a chance to obtain a correct mark for an item they had little knowledge about. This phenomenon is more likely to occur in high-stakes testing environments where failure to pass the examination has dire consequences for the individual. Thus, it is not surprising that this research reported some level of guessing, although relatively minimal. Previous studies have also demonstrated the presence of guessing factor on multiple-choice tests in some high-stakes examinations such as Senior School Certificate Examinations (SSCE), National Examination Council (NECO), West African Examination Council (WAEC) (Asuquo et al., 2022; Jimoh et al., 2020; Ogunbamowo et al., 2019). Whereas some of the earlier studies reported high guessing levels, others reported minimal factors, however. Jimoh et al. (2020), for example, reported 36% guessing on the 2016 WASSCE Economics multiple-choice test and 35% on the NECO Economics test. The discrepancies in the findings of this study and other previous studies may stem from variations in the subjects/course (e.g., Economics, Mathematics) studied.

Another important finding from this study is that the items had varied levels of difficulty across the 3-year data; whereas some items functioned at the lower ability level of the examinees, others functioned best at the higher ability level. However, between 2 to 5 items from each of the years showed a dysfunctional difficulty index indicating that these items were faulty and should be removed/modified. It sounds good that the items had varying difficulty levels and thus, those items that were dysfunctional were few. This understanding is premised on the fact that item difficulty may contribute significantly to the variations in students' performance (Cobbinah, Annan-Brew, & Quansah, 2022). The findings of the difficulty levels of items appear to contradict observations of previous studies. Existing studies that examined the difficulty indicators of national-wide exams revealed a moderate level of difficulty (Deborah, Temitope, & Peter, 2020; Ibrahim, 2023; Onah, & Jiwueze, 2015). Unfortunately, these studies focused on the overall difficulty, ignoring the specific items and thus, there is little clarity regarding the difficulty parameter details of the items.

The results further showed that, for the 3 years, 15% to 20% of the items exhibited poor discrimination. While the number of items with poor discrimination index was quite low, these items are still of concern since all items sum up to the conceptualization/measurement of the student's ability construct. Such items have limited functionality and utility in terms of scaling examinees into those with sufficient mastery over the content area and those without (Nitko, 2001; Quansah, & Cobbinah, 2021). A notable concern from this study's results is that about 5% to 7.5% showed negative discrimination indices. A key factor that could explain this occurrence is the possibility that the agriculture tasks covered complex materials such that it would be probable for examinees to choose the correct answer without having mastery of the task (Matlock-Hetzel, 2011). From a broader perspective, poor distractor functioning can contribute to a poor discrimination index or even negative discrimination. Even though this study did not analyze the effectiveness of the distractors for the items with poor discrimination, it is suspected that the presence of weak distractors could be a possible cause of the dysfunction of the items flagged as having poor discrimination.

Given that some of the items had problems with guessing, difficulty, and discrimination indices, it was expected that these items would provide minimal information to the measurement of students' ability in agriculture. These items had very poor ICCs to the extent that a negative relationship was revealed between the probability of correctly answering a question and the ability level of the examinees. Based on the nature of ICCs, these problematic items are contributing very little to the measurement of examinees' ability in the Agriculture subject. The implication is that these items (not many, though) cannot help in scaling examinees into their performance profiles and consequently, affecting the variabilities in students' achievement in agriculture subjects. This issue may lead to high measurement error and low reliability of the results obtained from such tests (Cobbinah et al., 2022).

Sex DIF in the JCE Agriculture Multiple Choice Items

Gender DIF was found to be prevalent in each of the three-year data. In 2018, 7 items (17.5%) showed DIF (large effect size) with 6 being uniform (4 favours male and 2 favours female) and 1 being non-uniform. For the 2019-year data, 5 items (12.5%) showed uniform DIF (3 favours male and 2 favours female). Eight (8) items (20%) were flagged as DIF in 2020, with 4 items favouring female students, 3 favouring male students, and 1 being non-uniform DIF. Similar levels of DIF have been reported in earlier studies that focused on nationwide examinations (Amaechi & Onah, 2020; Ekong et al., 2020; Ikeh et al., 2021; Omorogiuwa & Iro-Aghedo, 2016; Woitschach et al., 2019; Yao & Chen, 2020). Ikeh and colleagues, for instance, found that 28% items of the 2018 SSCE Economics items were flagged as DIF items. More closely related to the findings of this study is the outcome of Adedoyin's (2010) research which found that over 13% of the items were tagged as DIF in the Botswana Junior Certificate Examination in Mathematics paper. Taken together, it appears that gender DIF is common in most nationwide and large-scale assessments due to several extraneous variables that are irrelevant to the trait being measured and this subsequently affects test performance as well as fairness (Penfield & Lee, 2010). To further understand the nature of DIF present in the dataset, the items that were flagged as gender DIF were evaluated; this activity has been rarely undertaken in previous studies.

In summary, 20 items out of 120 items for the years 2018, 2019, and 2020 examination review period had large gender DIF. The breakdown shows that 10 items favoured the male examinees, 8 items favoured the female examinees, with two items showing non-uniform DIF.

Gender DIF items favouring male examinees

An examination of item content for the male-favoured DIF items points to various agriculture content areas contributing to the phenomena over the three years under study. More specifically, items that showed DIF favouring male examinees were largely associated with livestock management and related activities in animal production and other activities that require manpower. It is plausible that DIF occurred due to a minor secondary ability difference between female and male groups, which was not measured by the test.

A closer look at the 2018 *question samples 1a* and *1b*, which showed gender DIF in favour of male examinees, it is observed that these items are associated with picture reading about management activities in animal production.

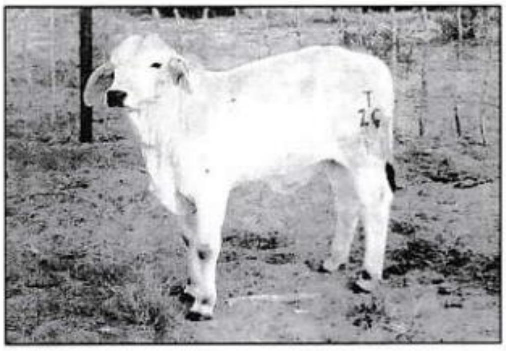
The diagram below shows an animal handling structure. Use it to answer question 23.

23. What is the name of the structure shown?

A Dip tank
B Spray race
C Cattle crush
D Holding parlour

Item Sample 1a: Question 23 of the 2018 BEC Agriculture Paper

The diagram below shows a calf that can be easily identified by the owner. Use it to answer question 24.



24. Which method of identification makes the calf easily identifiable?

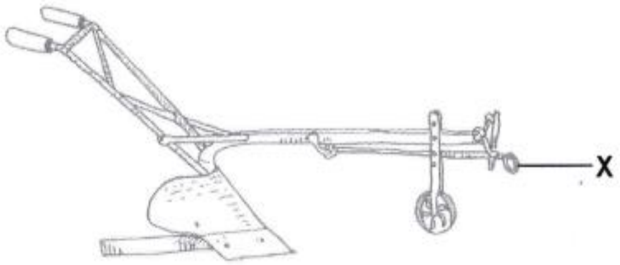
A Branding
B Notching
C Tagging
D Tattooing

Item Sample 1b: Question 24 of the 2018 BEC Agriculture Paper

The male dominance and advantage observed in both items (*Item Samples 1a & 1b*) have their attributed roots from the farming and cultural practices in Botswana. More explicitly, the DIF observed in favour of males points to the exposure males have due to cultural farming practices inherent in the country. Traditionally, looking after livestock, particularly, cattle is a men/boy's domain (UNDP, 2016). School-going boys usually tend to cattle over the weekends or on long school vacations at the designated cattle posts (Zewdu et al., 2016). It is during this period that they are exposed to livestock management activities such as castration, external parasites control, branding, and many more. This exposure is likely to enrich their classroom experience over females on related topics due to the lack of livestock (cattle) in schools which makes teaching such topics abstract. Consequently, items related to livestock management are likely to function favourably to male students as observed over the three years examination period.

Item Sample 1c deepens the understanding of gender DIF that is accounted for by the experiences of males due to their involvement in farming activities (Zewdu et al., 2016). Male examinees are more likely to be favoured in this item (on the functions of a mouldboard plough) because of the use of cattle and donkeys as sources of draught power by some households in Botswana. During the ploughing season, some livestock is brought from cattle posts (*Meraka*) to ploughing fields to serve as draught power. This arrangement thus allows males to have a better understanding of implements such as ploughs used to till the soil.

11. What is the function of the part labelled **X** on the implement shown below?



A To cut the soil deep
B To make pulling easier
C To turn the furrow slice
D To balance the implement

Item Sample 1c: *Question 11 of the 2018 BEC Agriculture Paper*

An interesting twist to the findings on the evaluation of male-favoured items is projected in *Item Sample 1d*; a question on chicken production and management. Although female examinees are expected to be favoured in this content area due to their intense involvement in this area (FAO, 2018), this happened not to be the case.

26. A layer feeds on 60 g of layers mash a day.

How much layers mash will be fed to 100 layers in 5 days?

A 5 kg
B 30 kg
C 60 kg
D 100 kg

Item Sample 1d: *Question 26 of the 2020 BEC Agriculture Paper*

Notably, *Item Sample 1d* has an inherent secondary ability (mathematical ability) that is required for the examinee to be successful on the item. It must be noted that this mathematical ability is not the ability of interest in this test. It is possible that males are favoured in this item because of their generally superior ability in mathematics they have over their female counterparts. This understanding is not surprising as it has been consistently reported in earlier studies that there exists a male advantage on questions about geometry and measurements, numbers and computations, data analysis, and proportional reasoning which is the focus of this item (Abedalaziz, Leng, & Alahmadi, 2018; Adedoyin, 2011).

Except questions 11 (2018), 23 (2018), 24 (in 2018), and 26 (in 2020), all other items (i.e., Q32 in 2018; Q6, Q19, & Q21 in 2019; Q28 & Q37 in 2020) DIF was in favour of male examinees were from contents such as economics in livestock and crop diseases, chicken production and management, and horticulture farming and related activities. Interestingly, these items showed poor psychometric properties (e.g., poor or negative discrimination, high guessing, and difficulty index beyond the recommended range). Item samples are shown in *Item Sample 1e to 1h*.

6. Which disease is likely to have attacked the crop?

A Rust
B Smut
C Streak virus
D Mosaic virus

Item Sample 1e: *Question 6 of the 2019 BEC Agriculture Paper*

Use the information below to answer question 19.

Mr Tin's chickens are observed to have greenish diarrhoea, thick liquid running out of their nostrils as well as coughing and sneezing.

19. Which disease is likely to have attacked the chickens?

A Mareks
B Newcastle
C Coccidiosis
D Fowl typhoid

Item Sample 1f: *Question 19 of the 2019 BEC Agriculture Paper*

32. Which of the following breeds of chicken is kept for egg production?

A Plymouth Rock
B Cornish game
C Cobb 500
D Isa brown

Item Sample 1g: *Question 32 of the 2018 BEC Agriculture Paper*

Item Sample 6, for example, focused on crop diseases, with a difficulty index beyond the appropriate threshold (*b*=3.104) and poor discrimination (*a*=0.316) (see Table 8). Similar indicators were recorded for sample items *1f* (*b*=-0.995, *a*=0.158, see Table 8) and *1g* (*b*=-3.888, *a*=0.265, see Table 7). These findings suggest that items 6, 19 and 32 did not function as expected and thus, they failed to accurately differentiate between examinees with higher ability and those with lower ability. Meanwhile, scholars have found that items with lower discrimination power provide less accurate information (Hambleton et al., 1991; Zanon et al. 2016). Given this understanding, it is obvious that the ability estimations for male and female examinees using these items are inaccurate, and thus, any gender DIF identified is merely due to error.

A point worthy of highlighting is that some of the items that showed gender DIF in favour of male examinees showed negative discrimination. A classic example is question 37 of the 2020 Agriculture paper (see *Item Sample 1h*) which had a negative discrimination value of -0.380 (see Table 9). This index implies that examinees with lower ability have higher chances of getting the item right compared with those with higher ability (Bichi & Talib, 2018). This suggests that although the gender DIF showed that the item favoured male examinees, this might not be the reality since this phenomenon is caused by the negative discrimination indicator. Hence, this item may favour the examinees with weaker abilities and will be a disadvantage to examinees with higher abilities.

37. Which of the following is **not** a function of management?

A Control
B Implementation
C Packaging
D Planning

Item Sample 1h: *Question 37 of the 2020 BEC Agriculture Paper*

Gender DIF items favouring female examinees

Eight items demonstrated differential functioning in favour of females and thus, they were scrutinised to identify the sources of DIF. Unlike the male examinees being favoured on items that were related to livestock management in animal production, and other agricultural activities that are male-dominated, items that favoured female examinees reflected economics in livestock and crop diseases, chicken production and management, and horticulture farming and related activities.

Looking at Q2 (in 2020, *Item Sample 2a*) and Q8 (in 2019, *Item Sample 2b*), these questions were focused on crop diseases and horticulture farming, field crops to be specific. With Botswana's cultural farming arrangement where females are custodians of crop husbandry and horticultural farming (FAO, 2018; UNDP, 2016; Ihechu & Madu, 2016), it is plausible that this inevitably gives female students practical exposure and advantage over male examinees who are mostly limited in these areas (Satyavathi, Bharadwaj, & Brahmanand, 2010; Zewdu et al., 2016). This would have given the females an undue advantage over their male counterparts.

2. Which of the following nutrients is correctly matched with its deficiency symptom?

	Nutrient	Deficiency symptom
A	Nitrogen	Yellow leaves
B	Phosphorus	Formation of nodules
C	Zinc	Poor root development
D	Potassium	Reddish, purple leaves

Item Sample 2a: *Question 2 of the 2020 BEC Agriculture Paper*

Use the information below to answer question 8.

Ms Lelo is a farmer specialising in the production of oranges, carrots, grapes and cabbages.

8. What type of a farmer is Ms Lelo?

A Agronomist
B Florist
C Horticulturalist
D Posologist

Item Sample 2b: *Question 8 of the 2019 BEC Agriculture Paper*

Other items that showed gender DIF in favour of female examinees covered agriculture economics and milk products (i.e., Q38 in 2020; Q40 in 2018; Q31 in 2020, see *Item Samples 2c – 2e*). This finding can be explained by the fact that activities of agricultural marketing (economics) and milk products in Botswana are female-dominated activities (Ministry of Agriculture [Botswana], 2008). This perspective has been corroborated in several previous studies that have found women being at the centre of producing crops, selling farm products, and raising milk from animals, meats and eggs (Rao, 2006; Raney et al., 2011; Sachs, & Alston, 2010). These experiences of female examinees give them an advantage over their male counterparts.

38. A business owned by an individual is described as a

A business partnership.
B sole proprietorship.
C limited company.
D public company.

Item Sample 2c: *Question 38 of the 2020 BEC Agriculture Paper*

40. Which of the following factors **cannot** influence the demand of meat as a product in a market?

A Weather changes
B Taste of consumers
C Price of other goods
D Supply of other goods

Item Sample 2d: *Question 40 of the 2018 BEC Agriculture Paper*

In addition to the rich experiences aforementioned, females also have added advantage on some questions (e.g., Q31 in 2020, *Item Sample* 2e) as the key (yoghurt) is a product commonly used in home economics classes that are female dominated. Females may know about the cheese or yogurt-making process more than male examinees who mostly choose Design and Technology.

31. Which of the following is a milk product?

A Margarine
B Oil
C Tallow
D Yoghurt

Item Sample 2e: *Question 31 of the 2020 BEC Agriculture Paper*

It must be noted that items that covered areas such as livestock, operation of petrol chainsaws, and equipment used for testing mastitis in milk, that showed gender DIF in favour of female examinees had poor psychometric features (*Item Sample 2f, 2g, 2h*). For example, *Item Sample 2g* (Q14 in 2020) exhibited a high guessing factor (*c*=0.303). Similar indicators were also reported in *Item Sample 2h* (*c*=0.303, probability of getting an item right=.554), Given these characteristics, these items may not accurately estimate the abilities of the examinees and thus, the gender DIF observed is possibly erroneous.

21. Which pair of livestock diseases is caused by bacteria?

	Diseases	
A	Anthrax	Heart water
B	Foot and mouth	Anthrax
C	Heart water	Trypanosomiasis
D	Foot and mouth	Heart water

Item Sample 2f: Question 21 of the 2018 BEC Agriculture Paper

14. What impact does the operation of a petrol chainsaw have on the environment?

A Air pollution
B Soil erosion
C Deforestation
D Desertification

Item Sample 2g: Question 14 of the 2020 BEC Agriculture Paper

30. Which of the following equipment is used to test for mastitis in milk?

A Churn
B Indicator
C Strip cup
D Separator

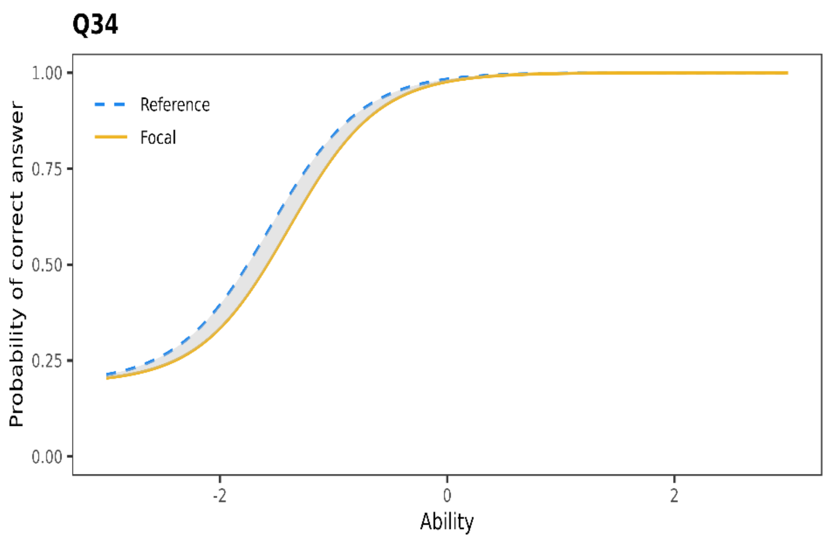
Item Sample 2h: Question 30 of the 2020 BEC Agriculture Paper

The diagram below shows the head of a hen before and after a management practice was carried out. Use it to answer question 34.

34. What management practice has been carried out on the hen?

A Debeaking
B Evisceration
C Scalding
D Vaccination

Item Sample 3a: Question 34 of the 2020 Agriculture Paper

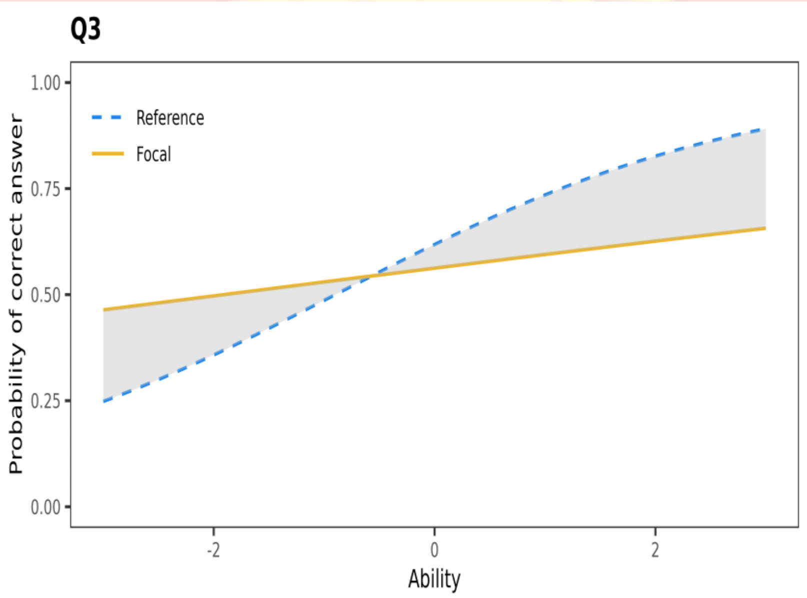


Out of all items identified gender DIF, 2 of them exhibited crossing DIF also referred to as non-uniform. For example, *Item Sample 3a* (Q 34, 2020), shows that at lower ability scale the reference group (females) performed better than males. However, as ability increased the probability of the focal group (males) getting the item increased and the performance gap narrowed. As highlighted, poultry farming is dominated by females which gives them the advantage of exposure (FAO, 2018). However, the decrease in the performance gap could be attributed to the fact males with higher abilities acquired knowledge during lessons at school and became comparable to their female counterparts,

3. Which of the following is a positive effect of the use of artificial fertilizers on the environment?

A Reduced soil erosion
B Increased water retention
C Increased microbial multiplication
D Enhanced healthy growth of vegetation

Item Sample 3b: Question 3 of the 2018 Agricultural Paper



Item Sample 3b (Q 3, 2018) shows it favoured focal group (males) at lower ability but as ability increased it favoured reference group (female). Though this item has stable parameter estimates, (*c*=0.010, *b*=.-0.638, and *a*=0.612), a close look shows some challenges with answer options. It is plausible that "*Enhanced healthy growth of vegetation*" which is key can lead to *reduced soil erosion* (distractor A) and *increased water retention* (distractor B). For this reason, examinees at higher ability may have seen these two distracters as possible answers. This may have led to false DIF flagging.

Location DIF in the JCE Agriculture Science Multiple Choice Items

The findings from this study showed that location DIF was present in the 2018-2020 BEC Agriculture examinations. Specifically, 12 items across the 3-year data exhibited location DIF with 3 items favouring examinees in rural schools, 8 items favouring those in urban schools, and 1 item being non-uniform DIF. Just like the findings of this study, existing literature have also revealed the prevalence of location DIF in nationwide and high-stakes examination in subjects like Economics, Social Studies, Mathematic,s and Agricultural Science (see Abedalaziz, 2012; Ikeh et al., 2020; Mokobi & Adedoyin, 2014; Obiebi-Uyoyou, 2023; Ogbebor, & Onuka, 2012; Osadebe & Agbure, 2020; Yohanna & Muhammad, 2022). What is common among these studies is that the majority of the items identified as location DIF favoured examinees in the urban schools; a similar finding was also found in this research. This differential item functioning could be attributed to variations in infrastructure, teacher allocation, teaching and learning materials, good communication networks, access to educational resources, and conducive teaching and learning environment (Alokan, 2010; Chime, 2012; Ikeh et al., 2020).

Items favouring Urban based examinees

Closely examining the items flagged as location DIF favouring examinees in urban schools, it was revealed that these items covered milk products, chicken rearing and management practices, and agriculture economics. For question 26 of the 2020 BEC Agriculture paper (see *Sample Item 1d*), for example, the tasks required an extra ability in mathematical computation in addition to an ability in chicken management practices. These traits required by the question for a candidate to be successful in responding are quite predominant in students found in schools located in urban areas (Calhoun et al., 2014; Obiebi-Uyoyou, 2023). It is not surprising that this same item favoured male examinees as have been indicated earlier on.

26. A layer feeds on 60 g of layers mash a day.

How much layers mash will be fed to 100 layers in 5 days?

A 5 kg
B 30 kg
C 60 kg
D 100 kg

Item Sample 1d: *Question 26 of the 2020 BEC Agriculture Paper*

Furthermore, question 31 (*Item Sample 2e*), which entailed the identification of milk products, favoured urban school examinees. This is not surprising as studies have shown that the key (i.e., Yoghurt) is a product that is associated with the affluent members of society who are predominantly in urban areas of Botswana (Kasimba, Motswagole, Covic, & Claasen, 2018; Legwegoh & Hovorka, 2016). Examinees in urban areas become familiar with such products and this could become a source of differential functioning of the task. Additionally, it becomes very uncommon for teachers in rural schools to adopt instructional delivery without a sample of the product and this situation is likely to occur since resources like electricity for refrigeration are a challenge (Legwegoh & Hovorka, 2016).

31. Which of the following is a milk product?

A Margarine
B Oil
C Tallow
D Yoghurt

Item Sample 2e: *Question 31 of the 2020 BEC Agriculture Paper*

For item 34 of the 2018 paper (*Item Sample 3a*), the management practice illustrated is carried out to prevent bad habits like egg pecking in chickens (debeaking). It is important to note that debeaking is done at 12 weeks of age to avoid carrying over into the egg production period. On the contrary schools usually receive chickens at 17-18 weeks. This invariably means debeaking is hardly carried out in schools as part of practicals to enhance learning. It is therefore mostly taught theoretically and the item favouring urban examinees may be due to the superior levels of ability of examinees in the urban areas (Alokan, 2010; Chime, 2012; Ikeh et al., 2020).

The diagram below shows the head of a hen before and after a management practice was carried out. Use it to answer question 34.

34. What management practice has been carried out on the hen?

A Debeaking
B Evisceration
C Scalding
D Vaccination

Item Sample 3a: Question 34 of the 2020 Agriculture Paper

Examining the items on agricultural economics (Question 37 [*Item Sample 1h*] & Question 40 [*Item Sample 3c*]) points to a key factor accounting for why these items favoured examinees in urban schools. These items are basically “recall” questions and thus, differential access to quality teachers, good materials, excursion sites, and good infrastructure by examinees in the urban areas may give them a better chance of performing well on the item (Chime, 2012; Mokobi & Adedoyin, 2014; Ikeh et al., 2020).

37. Which of the following is **not** a function of management?

A Control
B Implementation
C Packaging
D Planning

Item Sample 1h: *Question 37 of the 2020 BEC Agriculture Paper*

40. Which of the following best describes budgeting?

A Keeping records
B Buying commodities
C Making arrangements for selling
D Making financial plans for a business

Item Sample 3c: Question 40 of the 2020 Agriculture Paper

Notably, three items on chicken production and related activities (Q19 in 2019, Q21 in 2019 & Q32 in 2018) were flagged as location DIF in favour of urban school examinees. These items, however, had poor item properties (i.e., poor discrimination, faulty difficulty index) and consequently, may not have the capacity to accurately estimate students' ability within each group membership (Barker, 2001). Hence, it would not be appropriate for such a comparison to be made. This concern applies to question 6 of the 2019 BEC Agriculture paper which showed a non-uniform DIF; meanwhile, the item has a poor discrimination index.

Items favouring rural-based examinees

There are three items (Q2 in 2019, Q21 in 2018, Q24 in 2018) which favoured rural examinees. These three items were also fagged for DIF on the other two demographic areas of gender and school type. The content areas these items covered include planting tree seedlings, and livestock management and practices. For question 2, examinees are expected to give reasons for separating sub-soil from topsoil when digging holes for planting tree seedlings. Most farming activities are done in rural villages where there is sufficient land (Ministry of Agriculture [Bostwana], 2006). Due to population growth town schools have lost some of the land designated for agricultural practicals to construction of new classrooms (Drescher, 2002). This is likely to leave the remaining space insufficient for individual activities (in this case digging holes for tree planting). Consequently, examinees from rural schools would have practical exposure to reinforce theory and thus give them an advantage over town students.

In *question 21* of 2019, examinees were required to have knowledge about livestock diseases. Since communities in rural areas are allowed to keep up to 40 herds of cattle while those in towns are barred (Ministry of Agriculture [Bostwana], 2006), it is most likely students from rural areas know about animal diseases due to direct experience. More often animal disease outbreaks in Botswana are reported in those rural areas where livestock is kept in shared communal grazing areas (Darkoh & Mbaiwa, 2002; Gabalebatse et al., 2013). Though schools in rural areas like their counterparts in towns do not keep cattle, they have the advantage of proximity and access to those kept by students parents within the community.

Similarly, *question 24* of 2019 covers the livestock identification method (branding) and it favours examinees in rural schools. This is not at all surprising as this identification procedure in Bostwana is commonly carried out by parents with their children particularly boys (Ben et al., 2018; Oladele, 2011). This exposure and experience will most likely remain engraved on their minds and help them understand the theoretical part when attending agricultural lessons in schools.

DIF School Type in the JCE Agriculture Science Multiple Choice Items

For the 2018 to 2020 period under review, 6% of items showed school-type DIF with a breakdown of 3 items favouring examinees from public schools, none favouring private schools, and 4 showing non-uniform DIF. Although 23 items were flagged as DIF, only 7 are considered true DIF as others had poor parameter estimates. All previous studies that examined school-type DIF revealed its presence in BECE, SSCE, and NECO examinations (Dogan, & Ögretmen, 2008; Scarpati, Wells, Lewis, & Jirka, 2011; Ogbebor & Onuka, 2012; Ogbebor & Onuka, 2012; Ayva Yörü, & Atar, 2019; Osadebe & Agbure, 2020).

Items favouring Public school examinees

The results from the analysis showed that 11 items from the 3-year data (i.e., 2018-2020) were flagged as DIF in favour of public school examinees. However, only three items were considered true DIF while the remaining 8 were found to have poor parameter estimates. Closely evaluating these DIF items, it was found that the items reflected either practical contents or mathematical computations. This observation can be accounted for by the notable discrepancies in the structure and practices of the private and public school management.

Both examinees from public and private schools take three pieces of asssesments. However, private schools mostly take the written form of examination while their public school counterparts take practical school based examination. Consequently, the public schools engage their students in more practical lessons compared to the private school counterparts who do not usually have agricultural facilities for such exercise. An example is question 3 of the 2018 paper (see *Item Sample 3b*) which requested examinees to identify the positive effect of artificial fertilizers. The fact that examinees from public schools are mandated to present a vegetable plot with mature crops as part of their assessment for paper 3 gives them direct experience and use of artificial fertilizers. On the contrary, most students from private schools opt for written practical examinations (paper 4). The selection of paper 4 by most private schools limits experiential learning by examinees and puts the public schools at an advantage.

3. Which of the following is a positive effect of the use of artificial fertilizers on the environment?

A Reduced soil erosion
B Increased water retention
C Increased microbial multiplication
D Enhanced healthy growth of vegetation

Item Sample 3b: Question 3 of the 2018 Agricultural Paper

Furthermore, *Item Sample 4a Item Sample 4b* which is (Q 18) for both 2018 and 2019 BEC Agriculture papers also deepens the disparities regarding the item functioning between examinees in public and private schools. For (Q 18) of the 2018 paper, students were required to identify a proppagation method which happens to be grafting. This item requires practical experience or exposure to the many propgation methods. This is mostly done in public schools while due to a lack of land for agricultural practicals it is not done in most private schools. This supported by Isaacs, (2016) who states agricultural land is pushed out of towns and pre-urban areas for human settlement.

The diagram below illustrates stages of a method of asexual propagation. Use it to answer questions 18 and 19.

18. Which method of asexual propagation is this?

A Budding
B Cutting
C Grafting
D Layering

Item Sample 4a: Question 18 of the 2018 Agricultural Paper

Item Sample 4b (Q 18 of 2019) requires knowledge on the identification of protective clothes used when spraying pesticides on a farm. Vegetable production is carried out by all public school students. They have to prepare at least two different vegetable crops as part of the assessment for components for paper 3. For a similar reason of the lack of agriculture practicals in private schools, this is mostly not done. Students in public schools have the advantage of using or learning the protective clothes during management practice (pest control) of their vegetable crops.

The diagram below shows a farmer using a sprayer. The farmer is wearing protective clothing. Use it to answer question 18.

18. How is the protective clothing labelled **P** useful?

A It helps the farmer to breath.
B It helps the farmer to see far.
C It prevents the farmer from inhaling chemicals.
D It prevents the farmer from swallowing chemicals.

Item Sample 4b: Question 18 of the 2019 Agricultural Paper

Other items (i.e., Q 14, Q 26, and Q 30 all of 2020) which also exhibited DIF were observed to have poor parameter estimates. For example, their random guessing factors were *c*=0.305, *c*=0.253, and *c*=0.296 for Q14, Q16, and Q30 respectively. Given the fact the guessing factors are above the cut-off of .25, the items are not measuring the construct of interest well (Baker, 2001). Therefore, the identified DIF may be due to error.

Items with nonuniform DIF for school type examinees

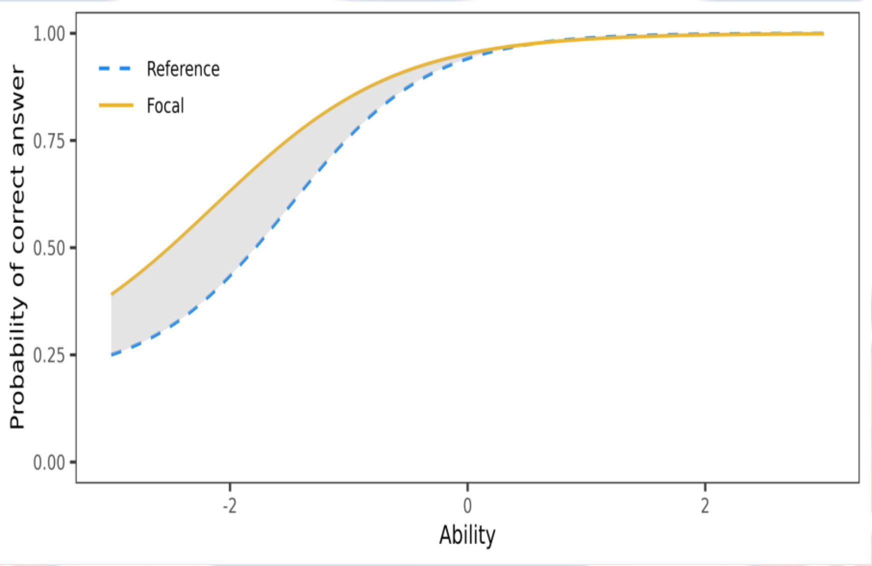
The results from the analysis showed that 10 items from the 3-year data (i.e., 2018-2020) were flagged as showing nonuniform. Notwithstanding, other items (i.e., Q 11 , and Q 24 all of 2018; Q 32, Q 37, and Q 40) were observed to have poor parameter estimates. For example, discrimination index were *a*=0.489, *a*=-1.283, and *a*=-380 for Q11, Q32 and Q37 respectively. For guessing factor it was *c*=0.489 and *c*=0.257 for Q24 and Q40 for 2018 and 2020 respectively. Futhermore items Q40 of 2018, Q2 of 2019 and Q34 of

2020 showed confounding DIF. This items had earlier been flagged for either gender or location DIF. Consequently, the identified DIF is due to intercation between two or more demographic varibales, making it to difficult to account for how each demogaphic variable contributed. The comfounding requires the use of an intersectional approach to DIF analyses (Russell etal, 2021; Mislevy etal, 2012).

Eventually, items Q23 and Q40 of 2019 were the ones showing true nonuniform DIF by school type.

Item Sample 4c: Question 23 of the 2019 Agricultural Paper

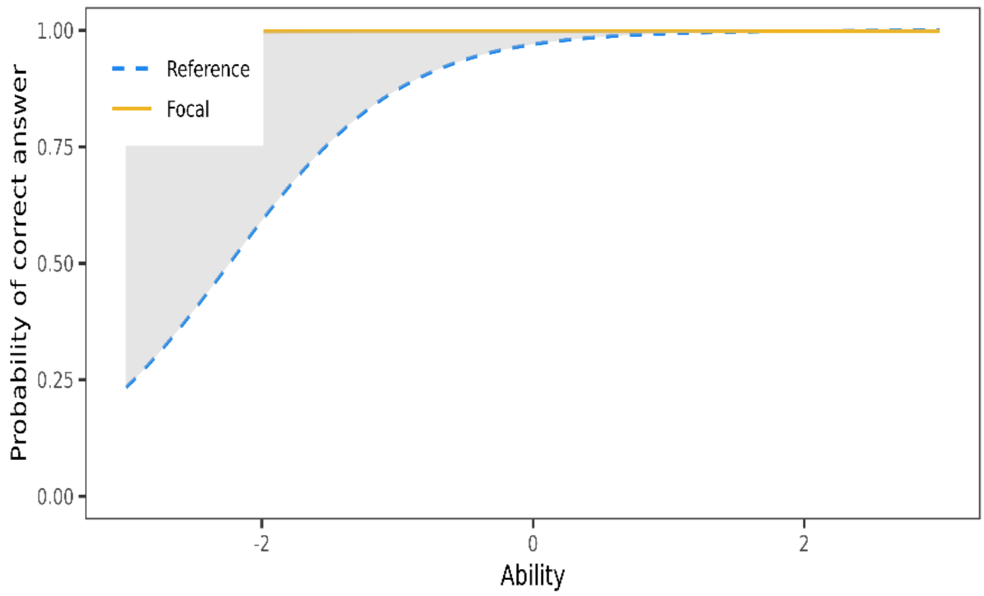
Q23



This item (Q23, 2019) is a graphic presentation of monthly egg production record. The item calls for secondary ability (mathematics). The item shows that at lower ability there is a peformance gap with private school examinees showing lower probabilty of getting the item right, Howver, as ability inreases the probality of endorsing the correct answer for private and public school examinees levelled up. This could be explained by the fact as part of chicken production practical lessons, examinees in public schools are required to keep records of daily to monthly egg collection. This would give them exposure and advantage over their private school conterparts. However, examinees in private schools with high ability were able to apply secondary ability (mathematics) to correctly interpret the graph regardless of their lack of expoure in record keeping in school.

Item Sample 4d: Question 40 of the 2019 Agricultural Paper)

Q40



Item Sample 4d (Q40 of 2019*)* is agricultural economics related. The item requires examinees ability to identify activities that can predict how products react to price change. The focal group (public school examinees) probability is consistent through out the ability scale while with private school examinees probality increased with increase in ability. Form the researcher's observation both public and private school examinees do not participate in farm products pricing and marketing. Its plausible that examinee response across all school types could be as result actual ability rather than construct irrelevant content.

Chapter Summary

items had varied levels of difficulty, 15% to 20% of the items exhibited poor discrimination It was discovered that 48 out of the 120 agricultural multiple-choice items for the years 2018-2020 functioned differentially by gender, location, and school type. Notwithstanding the 48 items showing DIF, some items showed DIF across all three demographic characteristics. This could be attributed to interaction effects as gender remained the same for examinees across school locations and types of schools. It was further revealed that DIF items emanated from crop science (field crops and horticulture), animal husbandry (cattle and chicken production), agricultural economics, and farm implements.

For cattle farming-related items, particularly those dealing with management practices, they favoured male, rural, and public school examinees. On the contrary, most items dealing with chicken production showed non-uniform DIF across the gender, location, and school type. For crop science, most items favoured female, rural, and public school examinees. However, items on horticulture favoured females and urban based examinees. School type had the most DIF items that were non-uniform. The study further found out that items on agricultural economics favoured females and urban examinees. When it comes to school type, most items showed crossing DIF (non-uniform). Most of these non-uniform items were observed to show DIF across at least two demographic characteristics.

CHAPTER FIVE

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

This section presents information in three main areas. The first section provides an overview of the research methodology and key findings. This is followed by how research contributes to knowledge and decision-making, and finally a summary of recommendations for policy and practice and recommendations for future research.

Summary

Overview of the study

The main purpose of this study is to examine the quality of Agriculture science examination items for the 2018-2020 review period using IRT 3PLM. The study was grounded in the positivists' paradigm using the descriptive quantitative approach. The study covered Agriculture science candidates for JCE. The study targeted 123,218 examinees who wrote 40-item test, and this translates to 4,928,720 cases of data point obtained from the Botswana Junior Certificate Examination from 2018, 2019, and 2020 years. The study used secondary data on students' responses to Agriculture Science JCE obtained from BEC. The study used all candidates' responses, thus it was based on the census approach.

Ethical approval was received from UCC's Ethics Review Board after the proposal was approved. This was followed by an introductory letter received from the Department of Education and Psychology. A letter was also obtained from one of the supervisors.

The researcher then wrote another letter and attached it to the first two together with a copy of the proposal. This package was then sent to BEC to officially seek permission to access and use the data on student responses. This was done after an initial meeting with the Director of DRPD-BEC to initially brief the Board and discuss with them the scope of my research work and its significance to BEC (Creswell, 2012). Key ethical issues of anonymity, data de-identification, and confidentiality were maintained. The data were analysed by conducting psychometric analyses.

Key findings

The findings of the study are summarised as follows:

1. Few JCE agriculture science multiple choice items for 2018-2020 had parameter estimates (difficulty, discrimination, and guessing indices) outside the thresholds ranges indicating that these items were faulty.
2. Most JCE agriculture science multiple choice items were functional and contributed accurately to the measuring of student achievement in agriculture science for 2018-2020.
3. There was differential item functioning in JCE agriculture science multiple choice items for 2018-2020 by sex. There were an equal number of DIF items for both female and male examinees. Non-uniform DIF was present in the the JCE Agriculture Science multiple-choice test by sex. Items on livestock management-related content favoured male students while those on crop husbandry favoured female students
4. There was differential item functioning in JCE agriculture science multiple choice items for 2018-2020 by school location (urban and rural). Differential item functioning concerning location was mostly in

favour of examinees in urban examinees with few in favour of rural examinees.

5. There was differential item functioning in JCE agriculture science multiple choice items for 2018-2020 by school type (public and private schools). Differential item functioning about school type was mostly in favour of public schools. Few items had confounding DIF which cuts across sex, location, and school type.

Conclusions

The study concluded that the contribution of items to student achievement in agriculture science was good with few items having validity and reliability concerns. It further concluded DIF exist across all three demographic variables of gender, location and school type. Items on livestock management-related content favours male students while those on crop husbandry favour female students. The observed difference appears to result from gender-based differences in exposure to and participation in agricultural activities at the family or societal level. The study further concluded that resource disparities, for example the lack of agriculture practical lessons in private schools disadvantages students in relation to answering items that are highly dependent on direct engagement.

Contribution to Knowledge

This study makes three main novel contributions. First, this appears to be the only research that has looked at sources of DIF in Agriculture science. It must be stated that the few studies that looked at DIF in Agriculture science did not go beyond identifying the origin of DIF items (Moyo 2015; Akanwa, Ihechu, & Nkwocha, 2020; Ihechu & Madu, 2016). This is a critical contribution in that it directly points item developers and curriculum specialists to content contributing to observed performance differences across demographic factors.

Secondly, this study appears to be the only research in the area of psychometric analysis that has investigated how items across the ability continuum contribute to the achievement construct. It must be acknowledged that most studies focused on fairness by identifying DIF items (Moyo 2015; Akanwa, Ihechu, & Nkwocha, 2020; Ihechu & Madu, 2016; Anan-Brew, 2020). What these studies failed to do is to assess the quality of the items in relation contribution to measuring the construct of interest.

Finally, the outcome of the study makes a significant contribution to knowledge for the validation of items by BEC. This research will draw the attention of BEC to an area that appears to have received less focus over the years. Studies on the quality of BEC items mainly focused on fairness and dimensionality (Moyo, 2017; Siamisang & Nenty, Motshabi, Kgosi & Nenty, 2012; Adedoyin, & Mokobi, 2013). This study will first draw the attention of item developers to parameter estimates and their contributions to validity, reliability, and fairness. This gives the BEC a more detailed and comprehensive picture about the overall quality of their items (Rust, & Golombok, 2014).

Recommendations

The following policy and practice recommendations were made based on the findings and conclusions of this study:

1. The quality of the items reflects the test development skills of the examiners/BEC or the entire assessment system. Based on the

observation that some items had faulty parameter estimates BEC must focus on exploring ways in which test items developed by experts have minimal problems. Training workshops should be rolled out to the examiners or subject experts during item writing.

2. Detailed psychometric analysis of the multiple-choice items by BEC during the pre-testing of items must be carried out. The BEC should engage in regular revision and validation of the items in terms of the structure of items, quality and functioning of the items. Faulty items must revisited or deleted thus ensuring that only items with good psychometric properties go into the BEC item bank.
3. Due to the findings there is differential item functioning in terms of sex, school location and school ownership this can be associated to different pedagogical practices. Continuous professional development should be rolled out to Agriculture teachers by the Ministry of Education and Skills Development through the Department of Curricum Development. The training workshops should guide teachers on how to narrow exposure difference between the various groups within different school settings. The study further recommends that schools must introduce field trips to makeup for agriculture teaching resources which may not be vailable in their schools.
4. The study recommends that the Ministry of Education and Skills Development must enact a policy that every school must have an operational farm and laboratory.

Suggestions for Further Studies

Although random guessing is a common practice among test-takers when responding to multiple-choice questions the current study did not look at the possible influence of distractors on response patterns. Implausible distractors are reported to increase the complexity of the cognitive task and this invariably increases guessing (Testa, Toscano, & Rosato, 2018; Rush, Rankin, & White, 2016; Jin, Siu, & Huang, 2022)

In addition, some of the items showed poor parameter estimates values. This is worrying because the contribution of items to achievement construct and DIF determination are estimated based on difficulty and discrimination. Future research on psychometric properties should incorporate distractor efficiency. Further research should also explore confounding DIF through an intersectional approach.

REFERENCES

Abdelaziz, M. M. A. (2012). A novel and generalized three-phase power flow algorithm for islanded microgrids using a newton trust region method. *IEEE Transactions on Power Systems*, *28*(1), 190-201.

Abdullahi, I. (2016). Analyses of psychometric properties of 2016 Mathematics Basic Education Certificate Examination Questions (BECEQ) in Gombe State. *Nigerian Journal of Educational Research and Evaluation*, *19*, 67-77

Abedalaziz, N., Leng, Â. C. H., & Alahmadi, Â. A. (2018). Detecting a gender-related differential item functioning using transformed item difficulty. *MOJES: Malaysian Online Journal of Educational Sciences*, *2*(1), 16-22.

Abedlaziz, N. (2011). Detecting a gender-related dif using logistic regression and transformed item difficulty. *Online Submission*.

Abonyi, S. O. (2011). *Instrumentation in behavioural research: A practical approach*. Timex Publisher.

Aborisade, O. J., & Fajobi, O. O. (2020). Comparative analysis of psychometric properties of mathematics items constructed by WAEC and NECO in nigeria using item response theory approach. *Educational Research and Reviews*, *15*(1), 1-7.

Acar, T. (2011). Sample Size in Differential Item Functioning: An application of hierarchical linear modeling. *Educational Sciences: Theory and Practice*, *11*(1), 284-288.

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.

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.

Adedoyin, O. O., Nenty, H. J., & Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Reviews*, *3*(3), 83-93.

Adedoyin, R. A., Mbada, C. E., Balogun, M. O., Martins, T., Adebayo, R. A., Akintomide, A., & Akinwusi, P. O. (2008). Prevalence and pattern of hypertension in a semiurban community in Nigeria. *European Journal of Preventive Cardiology*, *15*(6), 683-687.

Ahmad, H., & Mokshein, S. E. (2016). Is 3PL item response theory an appropriate model for dichotomous item analysis of the anatomy and physiology final examination (13-23). *Jurnal Pendidikan Sains dan Matematik Malaysia*, *6*(1), 13-23.

Ajayi, L. (2009). English as a second language learners' exploration of multimodal texts in a junior high school. *Journal of Adolescent & Adult Literacy*, *52*(7), 585-595.

Akanwa, U. N., Ihechu, K. J. P., & Nkwocha, P. C. (2020). Effect of language manipulation on differential item functioning of agricultural science multiple choice test items. *Journal of the Nigerian Academy of Education*, *14*(1), 211-224.

Akindele, S. T. (2003). The Theory and Practice of Federalism as a Structural Mechanism of Governance: How Adequate for Gender Struggle and Representation in Nigeria? *The Anthropologist*, *5*(3), 169-178.

Akour, M., & AL-Omari, H. (2013). Empirical Investigation of the Stability of IRT Item-Parameters Estimation. *International Online Journal of Educational Sciences*, *5*(2), 291-301.

Alade, O. M., Aletan, S., & Sokenu, B. S. (2021). Assessing the differential item functioning of 2018 WASSCE mathematics achievement tests in Lagos State, Nigeria. *The African Journal of Behavioural and Scale Development Research*, *2*(2), 8-24.

Alagoz, C. (2005). *Scoring tests with dichotomous and polytomous items* [unpublished doctoral dissertation, University of Georgia].

Ali, S. K. (2006). Evaluation of multiple choice and short essay question items in basic medical sciences. *Pakistan Journal of Medical Sciences*, *30*(1), 3-6.

Alokan, F.B. (2010). Influence of sex and location on relationship between student problems and academic performance. *The Social Sciences (TSS), 5(4),* 340 – 345

Alordiah, C. (2015). Comparison of index of differential item functioning under the methods of item response theory and classical test theory in mathematics. [Unpublished doctoral thesis of Delta State University, , Nigeria].

Amad, A., & Moksham, P. (2016). The age-dependent plasticity highlights the conceptual interface between borderline personality disorder and PTSD. *European Archives of Psychiatry and Clinical Neuroscience*, *266*, 373-375.

Amaechi, C. E., & Onah, F. E. (2020). Detection of uniform and non-uniform gender differential item functioning in Economics multiple choice standardized test in Nigeria. *Journal of the Nigerian Academy of Education*, *15*(2), 224-233.

Amao, S. R., Adewuyi, A. S., Gbadamosi, J., Salami, T. B., & Ogunjinmi, O. O. (2016). Gender bias and achievement in agricultural science among public and private secondary school students in Oyo-west local government area of Oyo-state, Nigeria. *International Journal of the Guild of Contemporary Academic Researchers*, *2*(1), 102-108.

Amedahe, F. K., & Gyimah, E. A. (2015). *Introduction to educational research for centre for continuing education.* UCC Press.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education, (1999). *Standards for educational and psychological testing*. American Educational Research Association.

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

Amie-Ogan, O. T., & Friday, E. E. (2020). Influence of teachers' competence on students' academic performance in public senior secondary schools in ikot ekpene and essien udim local government areas of Akwa Ibom State. *International Journal of Innovative Social Sciences & Humanities Research 8*(4), 118-128.

An, X., & Yung, Y. F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *SAS Institute Inc.*, *10*(4), 1-14.

Anastasia, O., Urbina, S. (2002). Test preparation and performance: A self-regulatory analysis. *The Journal of Experimental Education*, *70*(2), 101-113.

Andersson, B., & Xin, T. (2018). Large sample confidence intervals for item response theory reliability coefficients. *Educational and Psychological Measurement*, *78*(1), 32-45.

Anene, G. U., & Ndubuisi, O. G. (2003). *Test development processes: Educational measurement and evaluation.* Theory and practice. University Trust Publishers.

Angoff, G. M. (1993). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*(4), 584-601.

Anhwere, Y. M. (2009). *Assessment practices of teacher training college tutors in Ghana* [unpublished doctoral dissertation, University of Cape Coast].

Anigbo, L. C. (2006). *Development and Standardization of Mathematics Achievement Test Batteries for primary four pupils in Nigeria.* [Unpublished doctoral dissertation, University of Nigeria Nsukka].

Annan-Brew, R. (2020). *Differential item functioning of West African Senior School Certificate Examination in core subjects in Southern Ghana* [unpublished doctoral dissertation, University of Cape Coast]

Ansari, S., Panhwar, A. H., & Shah, A. A. (2017). Post-positivism: An effective paradigm for social and educational research. *International Research Journal of Arts and Humanities, 45*(45), 253-259.

APA, AERA, & NCME (American Psychological Association, American Education Research Association, and National Council on Measurement in Education). (1999). *Standards for educational and psychological testing.* American Psychological Association.

Applegate, G. M., Sutherland, K. A., Becker, K. A., & Luo, X. (2019). The effect of option homogeneity in multiple-choice items. *Applied Psychological Measurement*, *43*(2), 113-124.

Apuke, O. D. (2017). Quantitative research methods: A synopsis approach. *Kuwait Chapter of Arabian Journal of Business and Management Review*, *33*(5471), 1-8.

Ashikhia, D. A. (2010). Students and teachers' perception of the causes of poor academic performance in Ogun State secondary schools in Nigeria: Implications for counselingg for national development. *European Journal of Social Sciences*, *13*(2), 229-242.

Asuquo, U. M., Ekon, J. S., & Chigbu, N. N. (2022). Guessing parameters and multiple-choice test items of 2016 - 2017 National Examination Council (NECO) mathematics examination instruments in Calabar Education Zone of Cross River State. *J Adv Educ Philos*, *6*(9), 492-497.

Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection: An IRT-Based Monte Carlo study with SIBTEST and Mantel-Haenszel procedures* [unpublished doctoral dissertation, Virginia Tech].

Ayanwale, M. A. (2019). *Efficacy of item response theory in the validation and score ranking of dichotomous and polytomous response mathematics achievement tests in Osun State, Nigeria*. [unpublished doctoral dissertation, University of Ibadan].

Ayanwale, M. A., Adeleke, J. O., & Mamadelo, T. I. (2018). An assessment of item statistics estimates of basic education certificate examination through classical test theory and item response theory approach. *International Journal of Educational Research Review*, *3*(4), 55-67.

Ayeni, M. F. (2022). Effects of experiential and problem-solving strategies on academic performance of biology students in senior secondary schools in Nigeria. *International Journal of Education, Learning and Development*, *10*(4), 1-8.

Ayva Yörü, and Atar, (2019) Ayva Yörü, F. G., & Atar, H. Y. (2019). Determination of Differential item functioning according to SIBTEST, Lord's [Chi-squared], Raju's Area Measurement and Breslow-Day Methods. *Journal of Pedagogical Research*, *3*(3), 139-150.

Bachman, L. F. (1995). *An investigation into the comparability of two tests of English as a foreign language* (Vol. 1). Cambridge University Press.

Baharloo, A. (2013). Test fairness in traditional and dynamic assessment. *Theory and Practice in Language Studies*, *3*(10), 1930-1938.

Baker, F. B. (2001). *The basics of item response theory*. http://ericae.net/irt/baker.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Baker, F. B., & Kim, S. H. (2017). *The basics of item response theory using R* (pp. 17-34). Springer.

Baliyan, S. P., & Nenty, H. J. (2015). Demographic factors influencing senior secondary school students' attitude towards agriculture in Botswana. *International Journal of Education and Research*, *3*(10), 457-470.

Baliyan, S. P., Malebalwa, K., Keregero, K. J., & Mabusa, K. (2021). Poor academic performance of students in agriculture at primary schools in Botswana: Analysis of causes and ways to improve. *International Journal of Learning, Teaching and Educational Research*, *20*(9), 116-135.

Bandalos, D. L., & Finney, S. J. (2018). Factor analysis: Exploratory and confirmatory. In *The reviewer's guide to quantitative methods in the social sciences* (pp. 98-122). Routledge.

Bean, G. J., & Bowen, N. K. (2021). Item response theory and confirmatory factor analysis: complementary approaches for scale development. *Journal of Evidence-Based Social Work, 18*(6), 597-618.

Beckstead, J. W., Swisher, L., & Bebeau, M. J. (2004). Factor analysis as a tool for survey analysis using a professional role orientation inventory as an example. *Physical Therapy*, *84*(9), 784-799.

Ben, B., Moreki, J. C., Wame, B., Tlotleng, K., & Lesaba, K. (2018). A survey of livestock theft at Mogonono village in Kweneng District of Botswana. *Journal of Animal Science and Veterinary Medicine*, *3*(4), 94-10.

Bichi, A. A., & Talib, R. (2018). Item response theory: an introduction to latent trait models to test and item development. *International Journal of Evaluation and Research in Education*, *7*(2), *142-151*.

Bland, L. C. (2017). *Applying psychometric theory to the examination of learning within informal engineering learning environments.* George Mason University, Fairfax.

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Bonsu, H. D. (2016). A comparative analysis of academic performance of public and private junior high schools in the basic education certificate in Sekondi/Takoradi. *European Journal of Basic and Applied Sciences*, *3*(1).

Boone, S. E., & Canterbury, S. K. (2005). A 3D virtual desktop. *Proceedings of the 4th International Conference on Mobile and Ubiquitous Multimedia*.

Botswana Examination Council (2020), *Summary of results*. www.bec.co.bw /assessment-tools/schemes...assessment/.../agric-jcesyllabus-2020-final

Botswana, S. (2015). *Annual agricultural survey report 2013.* Published by the Department of Printing and Publishing Services Gaborone, 24-26.

Botswana, S. (2018). *Botswana demographic survey report 2017*. Retrieved from: https://statsbots.org.bw/

Statistics Botswana (2023) Secondary Education Stats Brief 2023. Gaborone, Botswana: Statistics Botswana.

Bulala, T., & Malima, K. (2019). Gender as a factor in predicting academic performance in agriculture in primary school leaving examination in Gaborone. *Asian Journal of Education and Social Studies, 5*(2), 1-8.

Bulala, T., & Mbisana, M. (2019). Validity of standard four attainment scores in predicting agriculture primary school leaving examination results. *Journal of Agriculture and Crops*, *5*(9), 172-177.

Bulala, T., Ramatlala, M., & Nenty, H. J. (2014). Location as a factor in the prediction of performance in Botswana junior school certificate agriculture examinations by continuous assessment scores. *Creative Education*, *5*(1), 11-14.

Bundsgaard, J. (2019). DIF as a pedagogical tool: Analysis of item characteristics in ICILS to understand what students are struggling with. *Large-scale Assessments in Education*, *7*(1), 9.

Butakor P.K. (2022). Using classical test and item response theories to evaluate psychometric quality of teacher-made test in Ghana. *European Scientific Journal, ESJ, 18* (1), 139. https://doi.org/10.19044/esj.2022.v18n1p139.

Buzick, H., & Stone, E. (2011). Recommendations for conducting differential item functioning (DIF) analyses for students with disabilities based on previous DIF studies. *ETS Research Report Series, 2011*(2), i-26.

Calhoun, A., Goeman, M., & Tsethlikai, M. (2014). Gender Equity for American Indians. In *Handbook for Achieving Gender Equity Through Education* (pp. 555-582). Routledge.

Camilli, G. (2006). Test fairness. *Educational Measurement*, *4*, 221-256.

Camilli, G., & Fox, J. P. (2013). An aggregate IRT procedure for exploratory factor analysis. *Journal of educational and behavioral statistics*, *40*(4), 377-401.

Cao, F., Yan, X., Wang, Z., Liu, Y., Wang, J., Spray, G. J., & Deng, Y. (2017). Neural signatures of phonological deficits in Chinese developmental dyslexia. *Neuroimage*, *146*, 301-311.

Chan, E. K. (2014). Standards and guidelines for validation practices: Development and evaluation of measurement instruments. In *Validity and validation in social, behavioral, and health sciences* (pp. 9-24). Cham: Springer International Publishing.

Chime, U. M. (2012). Development and validation of economics achievement test for senior secondary school students. *Unpublished M. Ed. [Unpublished masters dissertation, University of Nigeria, Nsukka]*.

Cobbinah, A., Annan-Brew, R. K., & Quansah, F. (2022). Item difficulty as a source of variability in student achievement in the West African Secondary School Certificate Examination (WASSCE): Application of generalizability Theory. *J Adv Educ Philos*, *6*(3), 136-142.

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*(1), 15-26.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum.

Cohen, L., Manion, L., & Morrison, K. (2003). *Research methods in education*. Routledge.

Cole, J. S., Bergin, D. A., & Summers, J. (2018). A lottery improves performance on a low-stakes test for males but not females. *Assessment in Education: Principles, Policy & Practice*, *25*(5), 488-503.

Creswell, J. W. (2012). *Research design: Qualitative, quantitative, and mixed methods approaches.* Sage Publications.

Crocker, L. (2012). Introduction to measurement theory. *Handbook of complementary methods in education research* (pp. 371-384). Routledge.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Culligan, B. (2008). A comparison of three test formats to assess word difficulty. *Language Testing*, *32*(4), 503-520.

Darkoh, M. B., & Mbaiwa, J. E. (2002). Globalisation and the livestock industry in Botswana. *Singapore Journal of Tropical Geography*, *23*(2), 149-166.

Dayioğlu, M., & Türüt-Aşik, S. (2007). Gender differences in academic performance in a large public university in Turkey. *Higher Education*, *53*, 255-277.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Deborah, A. C., Temitope, B., & Peter, A. O. (2020). Comparism of item difficulty and discrimination of pre and post university entrance examinations in Nigeria. *The Universal Academic Research Journal*, *3*(1), 1-9.

DeMars C. (2008, April). *Scoring multiple choice items: A comparison of IRT and classical polytomous and dichotomous methods* [Paper presentation]. Annual meeting of the National Council on Measurement in Education, New York, NY, United States of America.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Denga, E. (2003). Relationship between electron-scattering grand total and ionization total cross sections. *Radiation Physics and Chemistry*, *68*(1-2), 169-174.

Dixit, P. (2019). Diversification of economy–an insight into economic development with special reference to Kurdistan's oil economy and agriculture economy. *Russian Journal of Agricultural and Socio-Economic Sciences*, *85*(1), 395-404.

Dogan, N., & Ögretmen, T. (2008). The comparison of Mantel-Haenszel, Chi-square and logistic regression techniques for identifying differential item functioning. *Egitim ve Bilim*, *33*(148), 100.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education*, *38*(9), 1006-1012.

Drescher, A. W. (2002, September). Food for the cities: urban agriculture in developing countries. In *International Conference on Urban Horticulture 643* (pp. 227-231).

Duong, D. (2004). Measuring client-perceived quality of maternity services in rural Vietnam. *International Journal for Quality in Health Care*, *16*(6), 447-452.

Effiom, A. P. (2021). Test fairness and assessment of differential item functioning of mathematics achievement test for senior secondary students in Cross River state, Nigeria using item response theory. *Global Journal of Educational Research*, *20*(1), 55-62.

Ekong, E. M., Ubi, I. O., & Eni, E. I. C. (2020). Differential item functioning of 2018 Basic education certificate examination (BECE) in mathematics: A comparative study of male and female candidates. *International Journal of Educational Administration, Planning & Research*, *12*(1&2), 57-65. https://doi.org/10.5281/zenodo.5892613

Eleje, L. I., & Esomonu, N. P. (2018). Test of achievement in quantitative economics for secondary schools: construction and validation using item response theory. *Asian Journal of Education and Training*, *4*(1), 18-28.

Embretson, S. E. (1996). Item response theory models and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*(3), 201-212.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*(4), 341-349.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Enunwah, R. D. C., Akwa, A. M., & Okon, C. E. (2014). Differential items and group functions of secondary students' achievements in mathematics in cross-river state. *African Education Indices*, *7*(1), 76 – 82.

Ezeh, C. O., & Onah, P. C. (2005). *Measurement and evaluation in education.* Computer Edge Publishers.

Falani, I., Nisraeni, N., & Irdiyansyah, I. (2017, October). The ability of estimation stability and item parameter characteristics reviewed by item response theory model. In *International Conference on Education in Muslim Society (ICEMS 2017)* (pp. 175-178). Atlantis Press.

FAO. (2018). *National gender profile of agriculture and rural livelihoods – Botswana: Country gender assessment series..* http://www. fao.org/3/i8639en/I8639EN.pdf.

Fechter, T., Dai, T., Cromley, J. G., Nelson, F. E., Van Boekel, M., & Du, Y. (2021, September). Developing a validity argument for an inference-making and reasoning measure for use in higher education. In *Frontiers in Education* (Vol. 6, p. 727539). Frontiers Media SA.

Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4*(2), 113-148.

Ford, L. R., & Scandura, T. A. (2018). A typology of threats to construct validity in item generation. *American Journal of Management*, *18*(2), 132-142.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315-332.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessmen* Routledge.

Gabalebatse, M., Ngwenya, B. N., Teketay, D., & Kolawole, O. D. (2013). Ethno-veterinary practices amongst livestock farmers in Ngamiland District, Botswana. *African Journal of Traditional, Complementary and Alternative Medicines*, *10*(3), 490-502.

Gana, M. (2007). *Forest strategy: Strategic management and sustainable development for the forest sector*. Springer Science & Business Media.

Gao, H. (2019). Reliability modelling for degradation-shock dependence systems with multiple species of shocks. *Reliability Engineering & System Safety*, *185*, 133-143.

Gbadamosi, A. (2007). Cognitive dissonance: The implicit explication in low-income consumers' shopping behaviour for "low-involvement" grocery products. *International Journal of Retail & Distribution Management*, *37*(12), 1077-1095.

Ghazvini, S. D., & Khajehpour, M. (2011). Gender differences in factors affecting academic performance of high school students. *Procedia-Social and Behavioral Sciences*, *15*, 1040-1045.

Goni, M. D., Naing, N. N., Hasan, H., Wan-Arfah, N., Deris, Z. Z., Arifin, W. N., ... & Arshad, M. R. (2015). Development and validation of knowledge, attitude and practice questionnaire for prevention of respiratory tract infections among Malaysian Hajj pilgrims. *BMC Public Health*, *20*, 1-10.

Government of Botswana. (2022). *Ministry of Basic Education.* https://www.gov.bw/ministries/ministry-basic-education

Guyer, R., & Thompson, N. A. (2011). User's manual for Xcalibre 4.1. *St. Paul MN: Assessment Systems Corporation*.

Hachey, K. K. (2008). *Using item response theory to examine the psychometric properties of the job content questionnaire* (Doctoral dissertation, University of Saskatchewan).

Hagell, P. (2014). Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: the primacy of theory over statistics. *Open Journal of Statistics*, *4*(6), 456-465.

Hagquist, C. (2019). Explaining differential item functioning focusing on the crucial role of external information–an example from the measurement of adolescent mental health. *BMC Medical Research Methodology*, *19*(1), 1-9.

Hair, J. F., Sarstedt, M., Pieper, T. M., & Ringle, C. M. (2012). The use of partial least squares structural equation modeling in strategic management research: A review of past practices and recommendations for future applications. *Long Range Planning*, *45*(5-6), 320-340.

Hambleton, R. K. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38-47.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Haris, A. (1989). Influencing Of Policy Challenges of Urban Agriculture and Economic Viability on Urban Agriculture in Libya. *International Journal of Business, Economics and Law*, *13*(3),1-10.

Harris, A.M., & Carlton, S.T. (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. *Applied Measurement in Education*, 6(2), 137-151.

Hasan, A. (2021). Ethical considerations in the use of secondary data for built environment research. In E. Manu & J.Akotia (Ed) *Secondary Research Methods in the Built Environment* (pp. 26-39). Routledge.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological measurement*, *66*(3), 393-416.

Herrera, A. N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity*, *42*, 739-755.

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

Hulela, K., & Miller, W. W. (2003). The development of agricultural education in the education system of Botswana. *Proceedings of the 19th annual conference of the Association for International Agricultural and Extension Education Conference*.

Hulela, K., Mukuni, J., Abreh, M. K., Kasozi, J., & Kraybill, D. S. (2021). Transformative curricula and teaching practices to meet labour market needs in tertiary agricultural education in africa.. Transforming Tertiary Agricultural Education in Africa, 126-134. https://doi.org/10.1079/9781789246544.0007.

Ibrahim, B. M. (2023). Psychometric properties analysis of mathematics West African Senior School Certificate Examination in Dala Education Zone, Kano State, Nigeria (2020-2022). *International Journal of Advanced Academic Research*, *9*(1), 199-213.

Ifeakor, A. C. (2012). Assessment of the optimization of e-learning facilities to lecturers and students in nigeria federal colleges of education. *AFRREV IJAH: An International Journal of Arts and Humanities*, *1*(4), 319-334.

Ihechu, K. J. P., & Madu, A. O. (2016). Differential item functioning of national examinations council (NECO) agricultural science questions for (2015). *Journal of Educational and Leadership Development December*, *8*(2), 1-17.

Ikeh, E. F., Ugwu, F. C., Mfon, T. E., Omosowon, O. V., Iketaku, I. R., Opa, F. A., Eze, B. A., Kalu, I. A., Ikwueze, C. C., & Ani, M. E. (2020). Analysis of differential item functioning in economics multiple choice items administered by West African Examination Council using logistic regression procedure. *Journal of Critical Reviews*, *8*(1), 980-986.

Ikeh, F. E., Ene, C. C., Ojobo, B., Ani, I. M., Metu, C. I., Ugwu, F. C., Owolawi, O., Omosowon, V. C., Oguguo, B. C., Ezugwe, J. I., & Agugoesi, J. O. (2021). Assessment of differential item functioning to detect gender biased items in Economics multiple choice questions in Senior School Certificate Examination. *Journal of Critical Reviews*, *8*(1), 546-552.

Iñarrairaegui, M., Fernández-Ros, N., Lucena, F., Landecho, M. F., García, N., Quiroga, J., & Herrero, J. I. (2022). Evaluation of the quality of multiple-choice questions according to the students' academic level. *BMC Medical Education*, *22*. https://doi.org/10.1186/s12909-022-03844-3

Ivanova, V. (2014). Use of Achievement Tests in the English Classes at the Distributed E-Learning Centre. *Proceedings of the International Conference FROM DELC TO VELSPACE, Plovdiv* (pp. 26-28).

Jia, B., He, D., & Zhu, Z. (2020). Quality and feature of multiple-choice questions in education. *Problems of Education in the 21st Century*, *78*(4), 576-594.

Jimoh, M. I., Daramola, D. S., Oladele, J. I., & Sheu, A. L. (2020). Assessment of items prone to guessing in SSCE economics multiple-choice tests among students in Kwara State, Nigeria. *Anatolian Journal of Education*, 5(1), 17-28. https://doi.org/10.29333/aje.2020.512a

Jin, K. Y., Siu, W. L., & Huang, X. (2022). Exploring the impact of random guessing in distractor analysis. *Journal of Educational Measurement, 59*(1), 43-61.

Jol, G., & Stommel, W. (2016). Ethical considerations of secondary data use What about informed consent? *Dutch Journal of Applied Linguistics, 5*(2), 180-195.

Jolliffe, I. T. (2002). *Principal component analysis for special types of data* (pp. 338-372). Springer.

Kalaycioğlu, D. B., & Berberoğlu, 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.

Karami, H, & Nodoushan M.A. (2016). Differential item functioning: Current problems and future directions. *International Journal of Language Studies.5*(3), 133- 142.

Karami, H., & Salmani Nodoushan, M. A. (2011). Differential iem functioning (DIF): current problems and future directions. *Online Submission*, *5*(3), 133-142.

Kasimba, S. N., Motswagole, B. S., Covic, N. M., & Claasen, N. (2018). Household access to traditional and indigenous foods positively associated with food security and dietary diversity in Botswana. *Public Health Nutrition*, *21*(6), 1200-1208.

Kean, J., Bisson, E. F., Brodke, D. S., Biber, J., & Gross, P. H. (2018). An introduction to item response theory and Rasch analysis: Application using the eating assessment tool (EAT-10). *Brain Impairment*, *19*(1), 91-102.

Kgoboge, K. (2020, November 6). Moeng College introduces multiple pathways. *Mmegi News*, p. 2.

Khwaileh, F. M., & Zaza, H. I. (2011). Gender differences in academic performance among undergraduates at the University of Jordan: Are they real or stereotyping. *College Student Journal*, *45*(3), 633-648.

Kilmen, S. (2016). Effect of DIF magnitudes, focal group sample size, and DIF ratio on the performance of SIBTEST. *International Journal of Social Sciences and Education*, *6*(1), 91-98.

Kim, S. H., Cohen, A. S., Alagoz, C., & Kim, S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement*, *44*(2), 93-116.

Koyuncu, I., & KILIÇ, A. (2019). The use of exploratory and confirmatory factor analyses: A document analysis. *Egitim Ve Bilim-Education and Science*, *44*(198), 361-388.

Kunwar, R. (2021). Construction and validation of self-assessment instrument for students' mathematics classroom learning behaviour. *International Journal of Innovation in Science and Mathematics Education*, *31*(2), 1-18.

Lai, J. S., Stucky, B. D., Thissen, D., Varni, J. W., DeWitt, E. M., Irwin, D. E., & DeWalt, D. A. (2005). Development and psychometric properties of the pediatric fatigue item banks. *Quality of Life Research*, *22*, 2417-2427.

Langer, M. R. (2008). Assessing the contribution of foraminiferan protists to global ocean carbonate production 1. *Journal of Eukaryotic Microbiology*, *55*(3), 163-169.

Latterell, C. M., & Regal, R. R. (2003). Are placement tests for incoming undergraduate mathematics students worth the expense of administration?. *Problems, Resources, and Issues in Mathematics Undergraduate Studies*, *13*(2), 152-164.

Leavy, P. (2017). *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. Guilford Publications.

Legwegoh, A. F., & Hovorka, A. J. (2016). Exploring food choices within the context of nutritional security in Gaborone, Botswana. *Singapore Journal of Tropical Geography*, *37*(1), 76-93.

Liao, L., & Yao, D. (2021). Grade-related differential item functioning in general English proficiency test-kids listening. *Frontiers in Psychology*, *12*, (5), 1-8.

Liu, R., & Bradley, K. D. (2021). Differential item functioning among English language learners on a large-scale mathematics assessment. *Frontiers in Psychology*, *12* (5), 1-12.

López, N., Erwin, C., Binder, M., & Chavez, M. J. (2018). Making the invisible visible: Advancing quantitative methods in higher education using critical race theory and intersectionality. *Race Ethnicity and Education*, *21*(2), 180-207.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

Lord, F. M., & White, D. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lu, L., & Xu, X. (2022). Research on Improving the Efficiency of School Enterprise Joint Training Mode. *Creative Education*, *13*(4), 1359-1367.

Machisu, V., Opondo, V., Nakhumicha, A., & Mosi, R. O. (2022). Influence of school agricultural farms on academic performance in agriculture in secondary schools. *Journal of Agriculture, Science and Technology*, *21*(3), 49-60.

Madu, B. C. (2012). Analysis of gender-related differential item functioning in mathematics multiple choice items administered by West African Examination Council (WAEC). *Journal of Education and Practice*, *3*(8), 71-78.

Makwinja, V. M. (2017). Rethinking education in Botswana: A need to overhaul the Botswana education system. *Journal of International Education Research (JIER)*, *13*(2), 45-58.

Masole, T.M. (2011). *Enhancing the quality of performance assessment in agriculture in Botswana.* [Unpublished doctoral dissertation). University of Pretoria, Pretoria].

Mathema, K. B. (2007). Crisis in education and future challenges for Nepal. *European Bulletin of Himalayan Research*, *31*(46), 46-66.

Matlock-Hetzel, S. (2011). *Basic concepts in item and test analysis*. A paper presented at the annual meeting of the southwest Educational Research Association, Austin. http://ericae.net/ft/tamu.Espy.htm. 1997.

McCarty, W. (2005). *Humanities computing* (p. 110). Palgrave Macmillan.

Medina, J. E. C., Poveda, D., & Ortíz, D. A. R. (2019). Education quality: Reflections on its evaluation through standardized testing. *Saber, Ciencia y Libertad*, *14*(2), 247-255.

Messick, D. M. (1995). Equality, fairness, and social conflict. *Social Justice Research*, *8*, 153-173.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Mills, A. J. (2017). Organization, gender, and culture. A.J, Mills (Ed). *Insights and Research on the Study of Gender and Intersectionality in International Airline Cultures* (pp. 15-33). Emerald Publishing Limited.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of Educational Data Mining*, *4*(1), 11-48.

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.

Monseur, C., Baye, A., Lafontaine, D., & Quittre, V. (2011). PISA test format assessment and the local independence assumption. *IERI Monographs Series. Issues and Methodologies in Large-Scale Assessments.*, *4*, 131-158.https://doi.org/10.12973/eu-jer.3.3.111

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.

Moyo, S. E. (2017). *Dimensionality of students' 2013 BGCSE Agricultural Examination: Implications for differential item functioning.* [Unpublished masters thesis, University of Botswana].

Mphale, L. M. & Mhlauli, M. B. (2014). An investigation on students' academic performance for Junior Secondary Schools in Botswana. *European Journal of Educational Research*, *3*(3), 111-127.

Mudekunye, J., Manwa, L., Manwa, L., & Mpofu, M. (2020). Educational goals for practical subjects: factors that determine implementation of physical education and home economics curriculum in Masvingo primary schools, Zimbabwe. *International Journal of Advanced Research and Publications*, *4*(3), 201-206

Muijs, D. (2004). *Doing qualitative research in education with SPSS.* SAGE Publication.

Murphy, J., Elliott, S. N. & Goldring, E. (2006). *Learning-centered leadership: A conceptual foundation.* Learning Sciences Institute, Vanderbilt University (NJ1).

Naite, I. (2021). Impact of parental involvement on children's academic performance at crescent international school, Bangkok, Thailand. I.Naite (Ed). *Conference Series: Earth and Environmental Science* (Vol. 690, No. 1, p. 012064). IOP Publishing.

Ndlovu, N. (2018). School Resources and Student Achievement: A Study of Primary Schools in Zimbabwe. *Educational Research and Reviews*, *13*(7), 236-248.

Nenty, H. J. (2004). Trends in graduate studies: Lessons for African new and emerging universities. *Journal African Research Network*, *4*(2), 42-58.

Nenty, H. J. (2008). 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). Institute of Education, University of Ibadan, Nigeria.

Nitko, J. A. (2001). *Educational assessment of students*. Prentice Hall.

Nkpone, H. L. (2001). The use of latent trait modes in the development and validation of physics achievement test for senor secondary students. *Journal of Science Teachers' Association of Nigeria, 36 (1&2) 72, 80*.

Nworgu, B. G., & Agah, J.J. (2011). Application of three-parameter logistic model in the calibration of a mathematics achievement test. *Journal of Educational Assessment in Africa, 7,* 162 - 172.

Obiebi-Uyoyou, O. (2023). Assessment of differential item functioning in mathematics multiple choice test questions in senior secondary school certificate examination in Delta Central Senatorial District. *Assessment*, *3*(2), 34-57.

Obinne, A. D. E. (2012). Using IRT in Determining Test Item Prone to Guessing. *World Journal of Education*, *2*(1), 91-95.

Obinne, A.D.E. (2008). *Psychometric properties of senior certificate biology examinations conducted by West African Examinations council: Application of item response theory.*[Unpublished doctoral dissertation, University of Nigeria, Nsukka].

Obon, A. M., & Rey, K. A. M. (2019). Analysis of multiple-choice questions (MCQs): Item and test statistics from the 2nd year nursing qualifying exam in a university in Cavite, Philippines. *Abstract Proceedings International Scholars Conference* (Vol. 7, No. 1, pp. 499-511).

Odili, J. N. (2003). Effect of language manipulation on differential item functioning of biology multiple choice test. [unpublished thesis, University of Nigeria, Nsukka]

Ogbebor, U., & Onuka, A. O. U. (2012). Differential item functioning methods as an item bias indicator. *Educational Research*, *4*(4), 367-373.

Oguguo, B., & Lotobi, R. A. (2019). Parameters of basic science test item's of 2011 basic education certificate examination using item response theory (IRT) approach in Delta State, Nigeria. *European Journal of Educational Sciences*, *6*(1), 22-36.

Ogunbamowo, A. O., Adediwura, A. A. & Diyan, R. O. (2019). Psychometric properties of 2017 West African examination council and national examinations council's economic senior school certificate examination items. *Journal of Advances in Education and Philosophy*, *3*(1), 9-18.

Ogunyemi, O. (2000). GLIF3: the evolution of a guideline representation format. In *Proceedings of the AMIA Symposium* (p. 645). American Medical Informatics Association.

Oitsile, N., & Oats, R. A. (2020). Comparative study on the performance of students in PSLE and JCE agriculture. *Lonaka JoLT*, *11*(2), 80-90.

Okolo, S. O. (2006). *Ascension to power in Nigeria: A case study of leadership in a developing country* [unpublished doctoral dissertation, Capella University].

Okoye, K. R., & Udoudo, E. S. (2015). Vocationalisation of secondary education in Nigeria: Issues, challenges and prospects. *Journal of Education and Practice*, *6*(30), 71-76.

Oladele, O. I. (2011). Determinants of constraints to livestock Identification and Trace-back System use for Disease monitoring among cattle farmers in Botswana. *International Journal of Applied Research in Veterinary Medicine*, *9*(2), 143-53.

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education*, *24*(4), 349-366.

Oliveri, M. E., & Lawless, R. (2018). The validity of inferences from locally developed assessments administered globally. *ETS Research Report Series*, *2018*(1), 1-12.

Omorogiuwa, O. K., & Iro-Aghedo, E. P. (2016). Determination of differential item functioning by gender in the national business and technical examinations board (NABTEB) 2015 mathematics multiple-choice examination. *International Journal of Education, Learning and Development*, 4(10) 25-35.

Onah, F. E., & Jiwueze, C. C. (2015). Psychometric properties of west african examination council (WAEC) o'level multiple chioce past question paper on economics for the year 2013. *Journal of Science & Computer Education*, *3*(3) 23-34.

Onunkwo, G. I. N. (2002). *Fundamentals of educational measurement and evaluation*. Cape Publishers.

Opara, O. S. D., & Margaret, I. (2019). Differential item functioning in english language test using item response theory for ethnic groups. *Journal of Science and Computer Education*, *3*(3)24-54.

Opoku-Asare, N. A. A., & Siaw, A. O. (2015). Rural–urban disparity in students' academic performance in visual arts education: evidence from six senior high schools in Kumasi, Ghana. *Sage Open*, *5*(4), 2158244015612523.

Orabi, I. (2007, June). *Gender differences in student academic performance and attitudes*. *Annual Conference & Exposition* (pp. 12-777).

Orangi, A., & Dorani, K. (2010). Developing a social studies achievement test for high school students based on item-response theory (IRT). *Psychological Methods and Models*, *1*(1), 1-14.

Orluwene, G. W., & Emekene, C. O. (2017). Standardization and validation of the advanced progressive matrices (APM) Scale for use in Nigeria. *International Journal of Novel Research in Education and Learning*, *4*(2), 191-207.

Osadebe, P. U. (2014). Standardization of Test for Assessment and Comparing of Students' Measurement. *International Education Studies*, *7*(5), 94-103.

Osadebe, P. U. (2015). Construction of Valid and Reliable Test for Assessment of Students. *Journal of Education and Practice*, *6*(1), 51-56.

Osadebe, P. U., & Agbure, B. (2020). Assessment of differential item functioning in social studies multiple choice questions in basic education certificate examination. *European Journal of Education Studies*. https://oapub.org/edu/index.php/ejes/article/view/2863

Osadebe, P. U., & Agbure, B. (2020). Assessment of differential item functioning in social studies multiple choice questions in basic education certificate examination. *European Journal of Education Studies*, *6*(8), 312-344.

Osei-Mensah, F. (2012). *Factors that influence the performance in general knowledge in Art of senior high school students in Abura Asebu Kwamankese district in the Central Region* [Unpublished doctoral dissertation, Kwame Nkrumah University of Science and Technology].

Oshima, T. C., & Morris, S. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement: Issues and Practice, 27*(3), 43-50.

Oshima, T. C., Raju, N. S., & Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*(1), 1-17.

Otemuyiwa, B. I. (2017) gender influence on students'performance in agricultural science in basic education certificate examination (bece) in kwali area council, federal capital territory. *African Journal of Theory and Practice of Educational Assessment*, *3*, 43-50.

Özkan, Y. Ö., & Güvendir, M. A. (2021). Differential item functioning analysis of a high-stake test in terms of statistical regions of Turkey. *Journal of Pedagogical Research*, *5*(3), 122-134.

Pande, S. S., Pande, S. R., Parate, V. R., Nikam, A. P., & Agrekar, S. H. (2013). Correlation between difficulty and discrimination indices of MCQs in formative exam in physiology. *South-East Asian Journal of Medical Education*, *7*(1), 45-50.

Panhwar, A. H., Ansari, S., & Shah, A. A. (2017). Post-positivism: An effective paradigm for social and educational research. *International Research Journal of Arts & Humanities (IRJAH)*, *45*(45) 45-51.

Park, S. K. (2006). Development of high efficiency nanofilters made of nanofibers. *Current Applied Physics*, *6*(6), 1030-1035.

Penfield, R., & Lee, O. (2010). Test-based accountability: potential benefits and pitfalls of science assessment with student diversity. *Journal of Research in Science Teaching*, *47*(1), 6-24. doi:10.1002/tea.20307

Perez, R. (2020). Effects of cell cycle variability on lineage and population measurements of messenger RNA abundance. *Journal of the Royal Society Interface*, *17*(168), 20200360.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Sage.

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.

Polit, D. F., & Hungler, C. T. (2006). Generalization in quantitative and qualitative research: Myths and strategies. *International Journal of Nursing Studies*, *47*(11), 1451-1458.

Price, C. P. (2014). Existing and emerging technologies for point-of-care testing. *The Clinical Biochemist Reviews*, *35*(3), 155-162.

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1301013.

Quansah, F., & Cobbinah, A. (2021). Equivalence of parallel tests in a basic statistics course in higher education using classical measurement theory. *Canadian Journal of Educational and Social Studies*, *1*(2), 13-28.

Quansah, F., Amoako, I., & Ankomah, F. (2019). Teachers' test construction skills in senior high schools in Ghana: Document analysis. *International Journal of Assessment Tools in Education*, *6*(1), 1-8.

Rahman, N. ve Chajewski, M. (2014). A Comparison and Validation of 2- and 3-PL IRT Calibrations in BILOG, PARSCALE, IRTPPRO, flexMIRT, and LTM (R). National Council of Measurement in Education, at Philadephia. Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, 53, 495-502.

Rammolai, M. (2009). *The image of agriculture education in Botswana* (unpublished doctoral dissertation, Murdoch University).

Raney, T., Anríquez, G., Croppenstedt, A., Gerosa, S., Lowder, S. K., Matuschke, I., & Skoet, J. (2011). *The role of women in agriculture.* http://dx.doi.org/10.22004/ag.econ.289018

Rao, E. K. (2006). Role of women in agriculture: A micro level study. *Journal of Global Economy*, *2*(2), 109-120.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The Relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades. *Educational Researcher*. https://doi.org/10.3102/0013189X1 8762105

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230.

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. *Assessing Quality of Life in Clinical Trials: Methods of Practice*, *2*, 55-73.

Republic of Botswana, 1994. *The Revised National Policy on Education.* Government Paper No 2 of 1994. Government Printers, Gaborone.

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*. Government printer

Rivera, J. (2011). Test item construction and validation: Developing a statewide assessment for agricultural science education. *Career and Technical Education Research*, *36*(2), 69-80.

Robin, F., Zenisky, A. L., & Hambleton, R. K. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*(1), 51-64.

Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, *48*(2), 268-302.

Rodrigues, L. D. O., Costa, E. M., Silva, V. H. M. C., Mariano, F. Z., & Jesus Filho, J. D. (2021). A note on performance differences between urban and rural schools in Brazil. *Revista Brasileira de Economia*, *74*, 494-507.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*(1), 1-10.

Russell, M., Szendey, O., & Kaplan, L. (2021). An intersectional approach to DIF: Do initial findings hold across tests? *Educational Assessment*, *26*(4), 284-298.

Rust, J., & Golombok, S. (2014). *Modern psychometrics: The science of psychological assessment*. Routledge.

Sabri, S. (2013). Item analysis of student comprehensive test for research in teaching beginner string ensemble using model-based teaching among music students in public universities. *International Journal of Education and Research*, *1*(12), 1-14.

Sachs, C., & Alston, M. (2010). Global shifts, sedimentations, and imaginaries: An introduction to the special issue on women and agriculture. *Signs: Journal of Women in Culture and Society*, *35*(2), 277-287.

Şahin, M. G., Yildirim, Y., & Öztürk, N. B. (2022). Examining the Achievement Test Development Process in the Educational Studies. *Participatory Educational Research*, *10*(1), 251-274.

Satyavathi, C. T., Bharadwaj, C., & Brahmanand, P. S. (2010). Role of farm women in agriculture: Lessons learned. *Gender, Technology and Development*, *14*(3), 441-449.

Scarpati, S. E., Wells, C. S., Lewis, C., & Jirka, S. (2011). Accommodations and item-level analyses using mixture differential item functioning models. *The Journal of Special Education*, *45*(1), 54-62.

Schuster, C., & Martiny, S. E. (2017). Not feeling good in STEM: Effects of stereotype activation and anticipated affect on women's career aspirations. *Sex Roles*, *76*(1-2), 40-55.

Seheryeli, M. Y., & Şeref, T. A. N. (2019). Examination of the reliability of the measurements regng the written expression skills according to different test theories. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(3), 327-347.

Sharma Poudyal, C. (2012). Nepali private schools and tolerated illegality: A Foucauldian analysis of privatisation of education in Nepal. *Policy Futures in Education*, *15*(4), 537-549.

Sheppard, R., Han, K., Colarelli, S. M., Dai, G., & King, D. W. (2006). Differential item functioning by sex and race in the Hogan Personality Inventory. *Assessment*, *13*(4), 442-453.

Sheybani, E., & Zeraatpishe, M. (2018). On the dimensionality of reading comprehension tests composed of text comprehension items and cloze test items. *International Journal of Language Testing*, *8*(1), 12-26.

Sheytanova, T. (2015). *A Monte Carlo study comparing three methods for determining the number of principal components and factors*. [unpublished master's thesis, Örebro University].

Shogbesan, Y. O. (2017). *Effect of test facets on the construct validity of economics achievement tests in osun state secondary schools*. [Unpublished master's thesis, Obafemi Awolowo University, Ile-Ife, Nigeria].

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.

Sibanda, C., Hulela, K., & Tselaesele, N. (2016). Perceived influencers of the decline on performance of students in botswana general certificate of secondary education's agriculture examination results. *Journal of Education and Learning*, *5*(1), 199-209.

Sohn, K. (2012). A new insight into the gender gap in math. *Bulletin of Economic Research*, *64*(1), 135–155.

Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, *19*(2-3), 245-263.

Solly, M., & Koloi-Keaikitse, S. (2019). Practical performance as a predictor of theory in Botswana junior certificate agriculture. *Mosenodi Journal*, *22*(2), 89-101.

Stemler, S. E., & Naples, A. (2021). Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. *Practical Assessment, Research & Evaluation*, *26*, 11 45-54.

Stiggins, R. J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement: Issues and Practice*, 20(3), 5-15.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617.

Subedi, G., Shrestha, M. G., & Suvedi, M. (2014, May). Dimensions and implications of privatisation of education in Nepal: The case of primary and secondary schools. I. Macpherson, S. Robertson & G. Walford (Ed) *Education, privatisation and social justice: Case studies from Africa, South Asia and South East Asia. Oxford, United Kingdom: Symposium Books*.

Sullivan, G. M. (2011). A primer on the validity of assessment instruments. *Journal of graduate medical education*, *3*(2), 119-120.

Sumida, S., & Kawata, K. (2021). An analysis of the learning performance gap between urban and rural areas in sub-Saharan Africa. *South African Journal of Education*, *41*(2), 1-17.

Suping, K. (2022). Political Spectacle and the Decline of Public Education in Botswana. *Journal of Asian and African Studies, 0*(0). https://doi.org/10.1177/00219096221117077.the

Tavakol, M., & Dennick, R. (2012). Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. *Medical Teacher*, *34*(3), 161-175.

Tay, A. K., Rees, S., Chen, J., Kareth, M., Mohsin, M., & Silove, D. (2015). The refugee-mental health assessment package (R-MHAP); rationale, development and first-stage testing amongst West Papuan refugees. *International Journal of Mental Health Systems*, *9*, 1-13.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*(11), 39-49.

Testa, S., Toscano, A., & Rosato, R. (2018). Distractor efficiency in an item pool for a statistics classroom exam: assessing its relation with item cognitive level classified according to Bloom's taxonomy. *Frontiers in Psychology*, *9*, (15), 85-90.

Thapa, A. (2012). Public and private school performance in Nepal: An analysis using the SLC. *Examination Education Economics*, *23*(1), 47-62.

Thissen, D. (1991). *Multilog users' guide: Multiple categorical item analysis and test scoring using item response theory.*: Scientific Software International.

Thissen, D., & Steinberg, L. (2009). Item response theory. *The Sage Handbook of Quantitative Methods in Psychology*, *3*(4),148-177.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397-412.

Thobega, M., & Masole, T. M. (2008). Predicting students' performance on agricultural science examination from forecast grades, *US-China Education Review*, *5*(10), 45-52.

Thomas, D. R., & Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics*, *21*(2), 110-130.

Thompson, B. (2002). *Score reliability: Contemporary thinking on reliability issues*. Sage publications.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. *Seventeenth yearbook of the national society for the study of education*, 16-24.

Tisdell, C & Moepeng, P (2010). *The pattern of livelihoods in a typical rural village: The importance of subjective poverty and inequality measures in Botswana: New perspectives.* Koln, Germany: LAP Lambert Academic Publishing.

Tommy, U. E., & Udo, E. M. (2019). Examining item difficulty and student ability parameters of national examinations council's Biology examinations using the Rasch measurement model in Nigeria. *British Journal of Education*, *7*(8), 66-80.

Ubi, I. O. (2006). *Item local independence, dimensionality, and trend candidates' mathematics performance on University Matriculation Examination in Nigeria*. [Unpublished doctoral dissertation, University of Calabar, Nigeria.

Ugodulunwa, C. A. (2014). *Quality assurance in research, assessment and evaluation in Nigeria. A Paper presented at 16th Annual Conference of the Association of Educational Researchers and Evaluators of Nigeria held at International Conference Centre University of Calabar, Cross River State, Nigeria 14th-18th July*.

Ul Hassan, M., & Miller, F. (2019). Optimal item calibration for computerized achievement tests. *Psychometrika*, *84*(4), 1101-1128.

Umar, U.S. & Samuel, R.I (2018). School location as correlate of students' achievement in basic science. *International Journal of Innovative Education Research, 6*(3), 14-17.

Umoinyang, I. E. (2011). *The challenging of removing consistent errors in achievement test using differential item functioning (DIF) detecting methods.* Paper presented at the 30th conference of the Association of Educational Assessment in Africa, Nairobi, Kenya.

UNESCO. (2002). *Forum on the impact of open courseware for higher education in developing countries: Final report.* http://unesdoc.unesco.org/images/0012/001285/128515e.pdf

United Nations Development Programme, (2016). *Botswana gender assessment and mainstreaming strategy.*

United Nations Educational, Scientific and Cultural Organisation - International Bureau of Education [UNESCO-IBE] (2010). *World data on education* (7th ed.). Retrieved from http://www.ibe.unesco.org/

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2004). *The UN decade of education for sustainable development (2005-2014): International implementation scheme.* http://unedoc.unesco.org/images/0012/001285/128515e.pdf

Uwhekadom, E. E. (2014). *Influence of language manipulation on differential item functioning in chemistry multiple choice test items.* [Unpublished doctoral thesis, University of PortHarcourt].

Uzokwe, U. N. (2009). Gender roles in agricultural production in the Seychelles. *Nigeria Agricultural Journal*, *40*(2), 34-40.

Veldkamp, B. P. (2013). Application of robust optimization to automated test assembly. *Annals of Operations Research*, *206*(1), 595-610.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*(3), 255-275.

Wang, W. C., & Su, Y. H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113-144.

Ward, J. T., Ray, J. V., & Fox, K. A. (2018). Exploring differences in self-control across sex, race, age, education, and language: Considering a bifactor MIMIC model. *Journal of Criminal Justice*, *56*, 29-42.

Wetzel, A. P. (2012). Factor analysis methods and validity evidence: a review of instrument development across the medical education continuum. *Academic Medicine*, *87*(8), 1060-1069.

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion referenced licensing test. *Educational Measurement*, *60*, 1-33.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*(3), 60-66.

Woitschach, P., Zumbo, B. D., & Fernández Alonso, R. (2019). An ecological view of measurement: Focus on multilevel model explanation of differential item functioning. *Psicothema*, *31* (2), 194-203. doi: 10.733 4/psicothema2018.303

Woods, A. J., Göksun, T., Chatterjee, A., Zelonis, S., Mehta, A., & Smith, S. E. (2013). The development of organized visual search. *Acta Psychologica*, *143*(2), 191-199.

Woods, M. (2009). Rural geography: blurring boundaries and making connections. *Progress in Human Geography*, *33*(6), 849-858.

Woods, N. (2008). Whose aid? Whose influence? China, emerging donors and the silent revolution in development assistance. *International Affairs*, *84*(6), 1205-1221.

Wrenn, J., & Wrenn, B. (2009). Enhancing learning by integrating theory and practice. *International Journal of Teaching and Learning in Higher Education*, *21*(2), 258-265.

Wright, K. D., & Oshima, T. C. (2015). An effect size measure for Raju's differential functioning for items and tests. *Educational and Psychological Measurement*, *75*(2), 338-358.

Xi, X. (2010). How do we go about investigating test fairness. *Language Testing*, *27*(2), 147-170.

Yadidi, D.C., & Banda, A. C. (2008). Making assessment for the promotion of teaching and learning in school-the case of the Malawi School Certificate of Education Exams (MSCE). *Journal of Educational Assessment in Africa*, *2*(1), 44-52.

Yang, F. M., & Kao, S. T. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171–177. https://doi.org10.3969/j.issn.1002-0829.2014.03.010

Yao, D., & Chen, K. (2020). Gender-related differential item functioning analysis on an ESL test. *Journal of Language Testing & Assessment*, *3*(1), 5-19.

Yohanna, J., & Muhammad, H. B. (2022). Location and Gender as Determinants of Students' Academic Performance in Agricultural Science in Zariia Education Zone, Kaduna State. *International Journal of Innovative Science and Research Technology*, *7*(2), 95-99.

Yoon, S. Y. (2011). *Psychometric properties of the revised purdue spatial visualization tests: Visualization of rotations (The Revised PSVT: R)*. Purdue University.

Yörü, F. G. A., & Atar, H. Y. (2019). Determination of differantial item functioning (DIF) according to SIBTEST, Lord's χ^ 2, Raju's area measurement and Breslow-Day methods. *Journal of Pedagogical Research*, *3*(3), 139-150.

Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*(1), 65-79.

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessments: A study evaluating a two-stage approach. *Educational and Psychological Measurement*, *63*(1), 51-64.

Zewdu, A., Zenebe, G., Abraha, B., Abadi, T., & Gidey, N. (2016). Assessment of the Gender Role in Agricultural Activities at Damota Kebele of Haramaya District, Eastern Hararghe Zone, Ethiopia. *Assessment*, *26*(4), 11-19.

Zheng, M. (2017). Faculty professional development in emergent pedagogies for instructional innovation in dental education. *European Journal of Dental Education*, *21*(2), 67-78.

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*, *7*(3), 65-70.

Zieky, M. (2003). *A DIF primer*. Princeton, NJ: Educational Testing Service.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF).* National Defense Headquarters, 160.

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

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223-233.

APPENDICES

APPENDIX A

EXAMINATION QUESTIONS

2018

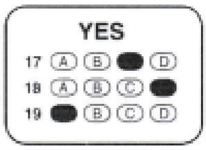
AGRICULTURE **16/1**

Paper 1 **October/November 2018**

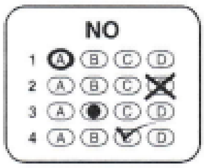
Marks: 40 **Time: 1 Hour**

INSTRUCTIONS

1. Answer **ALL** questions on the answer sheet provided.
2. Four possible answers are given for each question. Select the correct answer and fill the oval for that answer on your answer sheet.
   Be sure to fill the ovals like this:



3. If more than **one** oval is filled for a question it will be marked wrong. Erase completely answers that you change. Keep your answer sheet clean.
4. **DO NOT mark the oval like this:**



This question paper contains ***12*** *printed pages.*

DO NOT TURN THE PAGE UNTIL YOU ARE TOLD TO DO SO.

198161

1. Which set of minerals is made up of major elements only?

A	Potassium	Magnesium	Phosphorus
B	Magnesium	Copper	Nitrogen
C	Nitrogen	Phosphorus	Zinc
D	Copper	Iron	Cobalt

2. Which of the following activities would increase the pH value of soil from 3 to 5?

A Adding organic manure
B Adding inorganic fertilizers
C Ploughing back crop residues
D Flushing soil with a lot of water

3. Which of the following is a positive effect of the use of artificial fertilizers on the environment?

A Reduced soil erosion
B Increased water retention
C Increased microbial multiplication
D Enhanced healthy growth of vegetation

4. Which of the following activities can improve crop production in a flooding area?

A Adding inorganic fertilizers to the soil
B Planting trees around the location
C Growing crops on raised beds
D Mulching the soil with plastic

5. Which of the following pests affects stored grains?

A

B

C

D

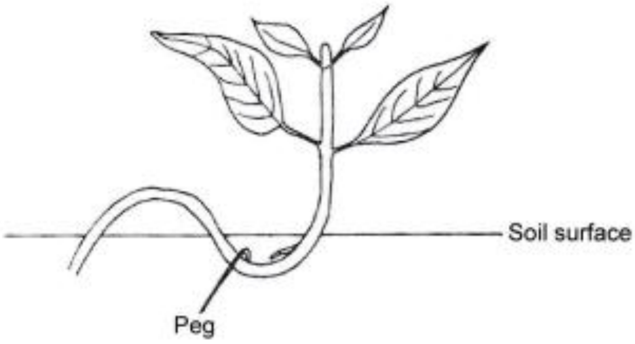
198161

The diagram below shows a farming operation on a field.
Use it to answer question 6.

6. Which activities are carried out before the operation?

A Disking and harrowing
B Clearing and stumping
C Stumping and planting
D Harrowing and clearing

The diagram below shows a method of propagating some fruit trees.
Use it to answer question 7.



7. Which of the following fruit trees can be propagated using the method shown?

A Apple
B Peach
C Banana
D Paw paw

The diagram below shows a farmer carrying out an activity on a fruit tree. Use it to answer question 8.

8. What is the **best** time for carrying out the activity shown?

A Late morning
B Early morning
C Late afternoon
D Early afternoon

9. What is the percentage by mass of phosphorus in a compound fertilizer 1:3:2 (20)?

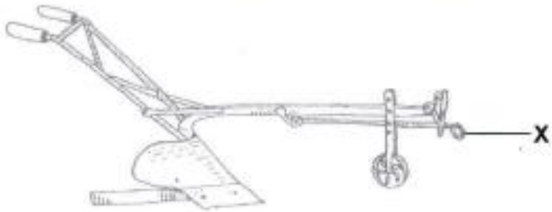
A 3.3%
B 6.6%
C 10.0%
D 20.0%

1:3:2 (20)
20:60:40
100

10. Which of the following classes of vegetables requires earthing up?

A Flower
B Leaf
C Legume
D Root

11. What is the function of the part labelled **X** on the implement shown below?



A To cut the soil deep
B To make pulling easier
C To turn the furrow slice
D To balance the implement

198161

12. The recommended application rate of super phosphate is 50 kg/ha.
1 ha = 10 000 m^2

How much fertilizer should be applied to a 20 m^2 plot?

A 0.001 kg
B 0.01 kg
C 0.1 kg
D 1.0 kg

Use the information below to answer question 13.

A student was given malathion powder to spray pests. The powder was not in its original container. She did not use all the powder.

13. Which of the following precautions should the student take before storing away the remaining powder for future use?

A Label the container.
B Find a cool place to store the powder.
C Wrap remaining powder in a plastic bag.
D Place the container on a high shelve in the store room.

14. What should be the maximum planting depth of a seed with a diameter of 3 mm?

A 3 mm
B 9 mm
C 15 mm
D 21 mm

15. Which of the following is an exotic fruit to Botswana?

A Mulberry
B Wild berry
C Snot apple
D Monkey orange

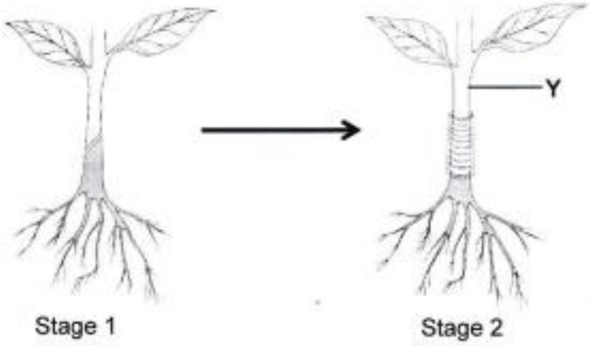
16. Which of the following is a suitable period for carrying out artificial insemination in Botswana?

A September to November
B November to January
C January to March
D April to June

17. Which of the following is a suitable temperature for storing semen used in artificial insemination?

A −196°C
B −96°C
C 96°C
D 196°C

The diagram below illustrates stages of a method of asexual propagation. Use it to answer questions 18 and 19.



18. Which method of asexual propagation is this?

A Budding
B Cutting
C Grafting
D Layering

19. What is the name of the part labelled **Y**?

A Bud
B Scion
C Union
D Root stock

20. Which of the following is **not** a requirement for establishing a vegetable business?

A Skilled personnel
B Reliable market
C Human labour
D Planting soil

21. Which pair of livestock diseases is caused by bacteria?

	Diseases	
A	Anthrax	Heart water
B	Foot and mouth	Anthrax
C	Heart water	Trypanosomiasis
D	Foot and mouth	Heart water

22. Which of the following is a concentrate feed?

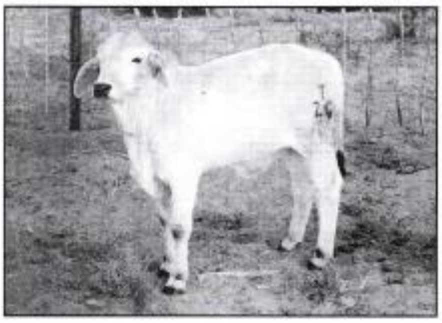
A Hay
B Lucern
C Rumevite
D Silage

The diagram below shows an animal handling structure. Use it to answer question 23.

23. What is the name of the structure shown?

A Dip tank
B Spray race
C Cattle crush
D Holding parlour

The diagram below shows a calf that can be easily identified by the owner. Use it to answer question 24.



24. Which method of identification makes the calf easily identifiable?

A Branding
B Notching
C Tagging
D Tattooing

25. At about what age do heifers reach puberty?

A 3 – 4 months
B 6 – 8 months
C 9 – 15 months
D 17 – 21 months

26. How long approximately, is the gestation period of a cow?

A 155 days
B 186 days
C 283 days
D 365 days

27. Which of the following is the **most** commonly used type of cattle housing in Botswana?

A Kraal
B Crush
C Calf pan
D Milk parlour

28. Why is it important to allow calves to suckle colostrum immediately after birth?

A It helps the calf gain strength after birth.
B It makes the cow to produce more milk.
C It relieves the cow from stress of retained milk.
D It helps the calf to expel its first dung after birth.

Use the information below to answer question 29.

Javu keeps his animals on a farm measuring 800 hectares. The farm is divided into 10 grazing areas, where 110 animals are kept per grazing area.

29. Javu sold 80 animals.
How many animals will remain in Javu's whole farm?

A 80
B 720
C 800
D 1020

30. Which of the following parts of an egg protects an embryo?

A Yolk
B Chalaza
C Albumen
D Membrane

31. Which of the following chicken diseases is caused by a virus?

A Bronchitis
B Newcastle
C Coccidiosis
D Fowl typhoid

32. Which of the following breeds of chicken is kept for egg production?

A Plymouth Rock
B Cornish game
C Cobb 500
D Isa brown

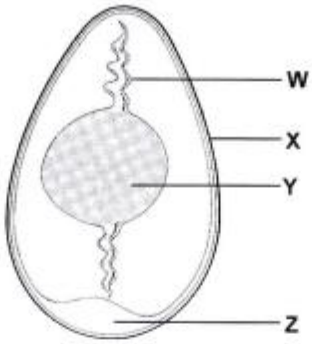
Use the information below to answer question 33.

Mr Osa is a poultry farmer and has 25 layers. After 3 weeks he had collected 430 eggs.

33. What percentage of the expected eggs was collected?

A 17%
B 22%
C 75%
D 82%

The diagram below shows an egg with parts labelled W, X, Y and Z. Use it to answer question 34.



34. Which part holds the egg yolk in position?

A **W**
B **X**
C **Y**
D **Z**

The table below shows characteristics of layers L, M, N and O on some given criteria. Use the table to answer question 35.

	Layers			
Criterion	**L**	**M**	**N**	**O**
Distance between pelvic bone and abdomen	2 fingers	2 fingers	4 fingers	4 fingers
Size of comb	large	small	large	small
Condition of eyes	bright	dull	bright	dull
Colour of shanks	light yellow	pale	light yellow	pale

35. Which is the correct sequence in which the layers **L**, **M**, **N** and **O** should be culled?

A **L, M, N, O**
B **N, O, L, M**
C **O, N, L, M**
D **M, O, L, N**

198161

The table below shows an extract of a record kept on a farm.
Use it to answer question 36.

Return / Income			Cost / Expenses		
Date	**Item sold**	**Amount**	**Date**	**Item bought**	**Amount**
14/6/12	10 rabbits	P500	11/4/12	2 bunnies	P160.00
20/6/12	10 bags of manure	P1000.00	15/5/12	1 buck	P100.00
27/6/12	10 kg fur	P1000.00	16/6/12	50 kg bag of feed	P350.00

36. What is the net income of the business?

A − P2500.00
B − P1190.00
C P610.00
D P1800.00

37. Information and communication technology promotes trade in the agricultural sector by allowing

A buyers to buy more.
B buyers to compare prices.
C producers to control prices.
D producers to produce more.

38. Which of the following is an example of a financial record kept on a poultry farm?

A Amount from the sale of chickens
B Cost of building a poultry house
C Number of chickens bought
D Number of chickens sold

39. Which of the following business names and legal type of business organisations are correctly matched?

	Business name	**Legal type of business organisation**
A	Triple X Investment	Corporation
B	Chicken City (Pty) Ltd	Partnership
C	G G Sharps and Associates	Limited company
D	J Giblets, Small General Dealer	Sole proprietor

40. Which of the following factors **cannot** influence the demand of meat as a product in a market?

A Weather changes
B Taste of consumers
C Price of other goods
D Supply of other goods

198161

2019

AGRICULTURE 16/1

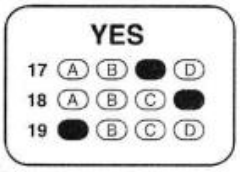
Paper 1 October/November 2019

Marks: 40 Time: 1 Hour

INSTRUCTIONS

1. Answer **ALL** questions on the answer sheet provided.

2. Four possible answers are given for each question. Select the correct answer and fill the oval for that answer on your answer sheet.
   Be sure to fill the ovals like this:



3. If more than **one** oval is filled for a question it will be marked wrong. Erase completely answers that you change. Keep your answer sheet clean.

4. **DO NOT mark the oval like this:**

*This question paper contains **11** printed pages.*
DO NOT TURN THE PAGE UNTIL YOU ARE TOLD TO DO SO.

199115

1. What does the value 13 on a pH meter indicate?

A Weak acid
B Strong acid
C Weak alkaline
D Strong alkaline

2. When digging holes for planting trees, top soil is separated from sub-soil.

This is done so that top soil is used for

A filling the bottom of the hole.
B making a basin around young plants.
C filling the space around the growing plants.
D making new seedbeds for raising seedlings.

3. Which of the following is **not** a method of preventing soil erosion?

A Contouring
B Intercropping
C Mulching
D Weeding

4. Which of the following factors determines the spacing of seedlings?

A Time to maturity
B Presence of weeds
C The likely prevalence of pests
D Machinery to use in management activities

5. To which class of crops does sunflower belong?

A Cereal
B Fruit
C Oil
D Root

Use the information below to answer question 6.

Mr Vavi planted a sorghum crop. At maturity stage, grains turned into small bag-like structures containing a dark powder.

6. Which disease is likely to have attacked the crop?

A Rust
B Smut
C Streak virus
D Mosaic virus

Use the information below to answer question 7.

A farmer planted maize on a plot. He observed that the maize has big soft and dark green leaves. The leaves easily break.

7. How can the farmer control the condition?

A Add lime
B Add phosphorus fertilizer
C Avoid adding nitrogen fertilizer
D Avoid adding potassium fertilizer

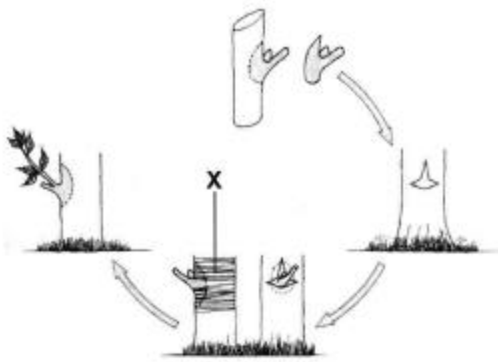
Use the information below to answer question 8.

Ms Lelo is a farmer specialising in the production of oranges, carrots, grapes and cabbages.

8. What type of a farmer is Ms Lelo?

A Agronomist
B Florist
C Horticulturalist
D Posologist

The diagram below shows a method of vegetative propagation. Use it to answer question 9.



9. What is the name of the binding material labelled **X**?

A Cloth
B Wire
C Rubber band
D Grafting tape

10. Which of the following propagation methods is suitable for the production of peaches?

A Budding
B Cutting
C Layering
D Sucking

11. Which of the following pests cause damage by sucking juices from vegetable crops?

A American ball worm
B Bargrada bug
C Grasshopper
D Corn cricket

The diagram below shows a pest feeding on a crop.
Use it to answer question 12.

12. What type of damage is caused by the pest shown?

A Biting
B Boring
C Chewing
D Sucking

The diagram below shows an activity carried out on a field.
Use it to answer question 13.

13. What is the environmental impact of the activity being carried out?

A Decreased rate of soil erosion
B Reduced exposure of soil to heat
C Reduced habitat for soil organisms
D Decreased amount of carbon dioxide

199115

14. Which of the following precautions should be taken before operating a petrol fuelled chainsaw?

A Check oil level
B Check age of the saw
C Check weight of the saw
D Check length of the blade

The diagram below shows a method of weed control being carried out on a field. Use it to answer question 15.

15. Which method of weed control is being used on the field?

A Cultural control
B Chemical control
C Biological control
D Mechanical control

16. Which of the following are recommended dimensions of a hole dug to plant a fruit tree?

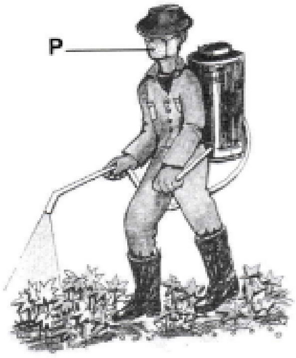
A 30 cm × 10 cm × 80 cm
B 10 cm × 10 cm × 100 cm
C 100 cm × 80 cm × 100 cm
D 100 cm × 130 cm × 180 cm

17. A student is interested in studying how to take proper care of machines used on a farm.

Which of the following subjects will the student find most useful?

A Biology
B Chemistry
C Mathematics
D Physics

The diagram below shows a farmer using a sprayer. The farmer is wearing protective clothing. Use it to answer question 18.



18. How is the protective clothing labelled **P** useful?

A It helps the farmer to breath.
B It helps the farmer to see far.
C It prevents the farmer from inhaling chemicals.
D It prevents the farmer from swallowing chemicals.

Use the information below to answer question 19.

Mr Tin's chickens are observed to have greenish diarrhoea, thick liquid running out of their nostrils as well as coughing and sneezing.

19. Which disease is likely to have attacked the chickens?

A Mareks
B Newcastle
C Coccidiosis
D Fowl typhoid

20. Which statement describes a brooding unit?

A It is where eggs are hatched.
B It is where hens sit on their eggs.
C It is where young chicks are raised.
D It is where broilers are slaughtered.

199115

The information below shows a record of activities carried out in a poultry farm. Use it to answer questions 21 and 22.

Date	Activities
14/07/2017	5 bags of feeds bought
14/07/2017	8 casual labourers hired
16/07/2017	Arrival of pullets
25/07/2017	150 eggs collected

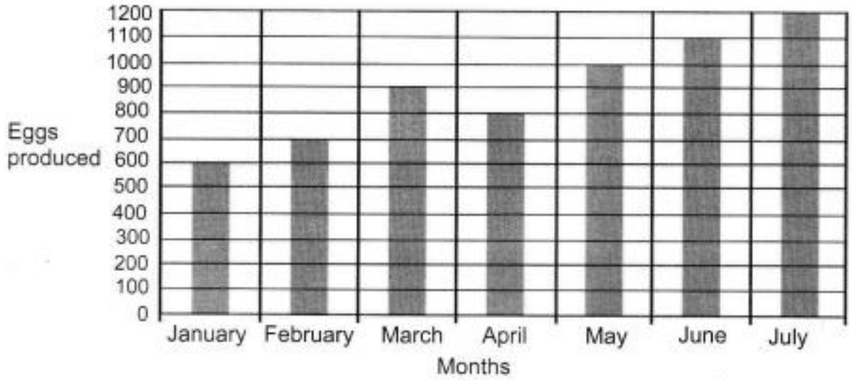
21. What can the farmer do to increase the number of eggs produced?

A Hire more casual labourers
B Buy more pullets
C Restrict feeding
D Ad-lib feeding

22. What type of enterprise is the farmer undertaking?

A Pullet production
B Feed production
C Layer production
D Broiler production

The chart below shows a record of egg production in a farm from January to July 2017. Use it to answer questions 23 and 24.



23. In which month was the lowest number of eggs produced?

A January
B February
C June
D July

199115

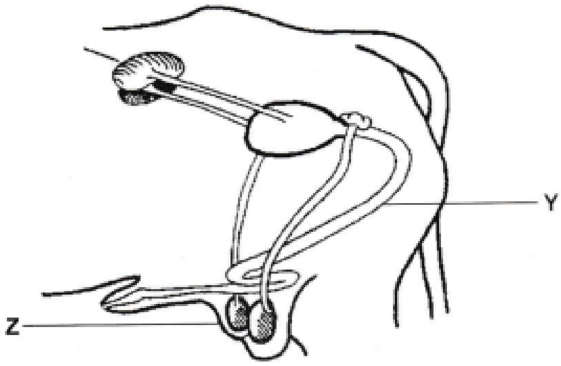
24. What is the total number of eggs produced in May and June?

A 1000
B 1100
C 2100
D 2200

25. Which of the following is a sign of good health in animals?

A Shiny fur
B Lameness
C Overweight
D Restlessness

The diagram below shows the reproductive system of a bull. Use it to answer questions 26 and 27.



26. What is the name of the part labelled **Z**?

A Epididymis
B Penis
C Scrotum
D Testis

27. What is the function of the part labelled **Y**?

A It produces hormones.
B It produces urine.
C It carries semen.
D It carries blood.

28. During which months are farmers allowed to send cows to Artificial Insemination camps in Botswana?

A October to December
B July to September
C January to March
D April to June

29. Which method of treating milk does **not** change its taste?

A Fermenting
B Pasteurizing
C Sieving
D Sterilizing

30. Which of the following factors about an animal does **not** affect meat quality?

A Breed
B Health
C Height
D Weight

Use the information below to answer question 31.

A farmer owns 24 Tswana cows. He bought an exotic bull to mate with his Tswana cows.

31. Which method of breeding is the farmer practising?

A Cross breeding
B Pure breeding
C Inbreeding
D Upgrading

32. It is important to observe signs of heat in animals for timely

A collection of semen.
B control of diseases.
C insemination.
D milking.

33. Which of the following animal feeds is appropriate for maximising returns in beef animals?

A Flakes
B Winter lick
C Grower mash
D Finisher mash

34. Which term describes the fusion of a male sex cell and a female sex cell?

A Fertilisation
B Gestation
C Ovulation
D Parturition

The diagram below shows a method of milking in progress. Use it to answer question 35.

35. What is the disadvantage of using the method of milking shown?

A It is expensive.
B Milk is contaminated.
C Animal teats are damaged.
D Milk temperature is reduced.

36. Which state of supply and demand are likely to result in lowest prices?

	Supply	Demand
A	High	High
B	High	Low
C	Low	High
D	Low	Low

199115

The information below was obtained from a farm record. Use it to answer questions 37 and 38.

	P
Cost of seeds	50
Cost of pesticides	60
Cost of 2:3:2 (22) fertiliser	150
Salaries for workers	2 040
Amount raised from selling onions	2 500

37. What type of record will contain the information shown?

A Financial
B Production
C Balance sheet
D Profit and loss account

38. What is the total cost incurred by the farm?

A P200
B P260
C P2 240
D P2 300

39. Which of the following activities is **not** a function of marketing?

A Advertising
B Importing
C Pricing
D Researching

40. Which of the following activities can help a farmer to know how a product reacts towards a change in price?

A Advertising
B Processing
C Researching
D Storing

2020

AGRICULTURE **16/1**

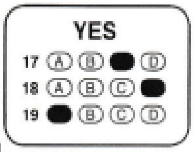
Paper 1 October/November 2020

Marks: 40 Time: 1 Hour

INSTRUCTIONS

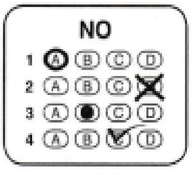
1. Answer **ALL** questions on the answer sheet provided.

2. Four possible answers are given for each question. Select the correct answer and fill the oval for that answer on your answer sheet.
Be sure to fill the ovals like this:



3. If more than **one** oval is filled for a question it will be marked wrong.
Erase completely answers that you change.
Keep your answer sheet clean.

4. **DO NOT mark the oval like this:**



---

*This question paper contains **11** printed pages.*
DO NOT TURN THE PAGE UNTIL YOU ARE TOLD TO DO SO.

1. Which of the following management activities is carried out to improve soil fertility?

   A Irrigating a field
   B Removing weeds
   C Burying crop residues
   D Ploughing along a slope

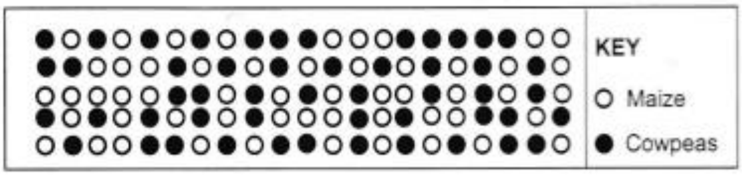
2. Which of the following nutrients is correctly matched with its deficiency symptom?

	Nutrient	Deficiency symptom
A	Nitrogen	Yellow leaves
B	Phosphorus	Formation of nodules
C	Zinc	Poor root development
D	Potassium	Reddish, purple leaves

3. Why is sorghum a recommended cereal crop for growing in Botswana?

   A It is the easiest crop to manage.
   B It cannot be affected by pests.
   C It is more drought resistant.
   D It matures earlier.

The diagram below illustrates a cropping system used by a farmer. Use it to answer questions 4 and 5.



4. Which cropping system is illustrated in the diagram?

   A Monoculture
   B Crop rotation
   C Mixed cropping
   D Continuous cropping

5. Which of the following is an advantage of the illustrated cropping system?

   A It makes harvesting easy.
   B It allows easy use of machinery.
   C It reduces competition for nutrients.
   D If one crop fails the other may succeed.

Use the information below to answer question 6.

Tao stored apples on a shelf. After two weeks he found that they were all brown and soft.

6. What could have Tao done to prevent the apples from turning brown and soft?

A Refrigerated them
B Wrapped them
C Smoked them
D Boiled them

7. Which of the following propagation methods is suitable for the production of bananas?

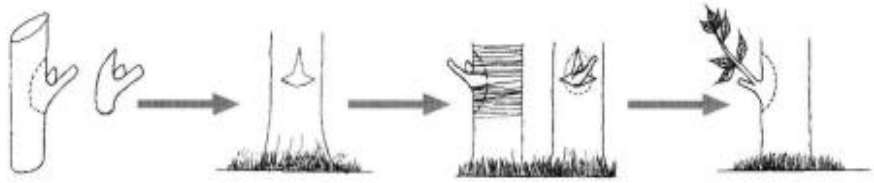
A Budding
B Cutting
C Layering
D Sucking

The diagram below shows a farm activity in progress.
Use it to answer question 8.

8. Which branches of Agriculture are represented in the farm activity?

A Animal husbandry, engineering and economics
B Agronomy, engineering and animal husbandry
C Agronomy, horticulture and engineering
D Horticulture, agronomy and economics

The diagram below shows a method of vegetative propagation.
Use it to answer question 9.



9. Which method of vegetative propagation is shown?

A Budding
B Cutting
C Grafting
D Layering

Use the information below to answer questions 10 and 11.

A farmer planted 20 certified seeds on a plot. Only 12 seedlings emerged.

10. What was the percentage of germination?

A 8%
B 20%
C 32%
D 60%

11. Which of the following factors could have led to the emergence of only 12 seedlings?

A Pest attack
B Low viability
C Deep planting
D Disease attack

The table below describes how a pest, represented by letter G, causes damage to crops and how it is controlled. Use it to answer question 12.

Pest	How damage is caused	How pest is controlled
G	Defoliates leaf vegetables	Use of scare-crows

12. At what stage of development is the pest represented by letter **G**?

A Adult
B Egg
C Larva
D Nymph

The table below shows the number of cattle kept in four different paddocks of the same size. Each paddock has fresh palatable grasses for animals to graze. Use it to answer question 13.

Paddocks	U	V	W	X
Number of cattle	50	120	80	200

13. Which paddock will experience **fast** pasture deterioration?

A **U**
B **V**
C **W**
D **X**

14. What impact does the operation of a petrol chainsaw have on the environment?

A Air pollution
B Soil erosion
C Deforestation
D Desertification

15. Which of the following crops is a leaf vegetable?

A Beetroot
B Broccoli
C Lettuce
D Pepper

16. Which of the following trees is indigenous to Botswana?

A Baobab
B Eucalyptus
C Pepper
D Syringa

17. The Devil's claw is used

A as food.
B as medicine.
C for dying clothes.
D for making baskets.

18. Which of the following management activities should be carried out on a mouldboard plough immediately after use?

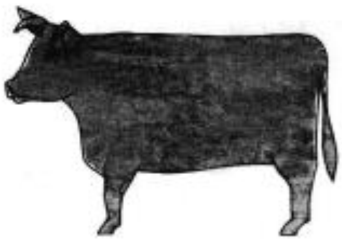
A Adjusting
B Oiling
C Painting
D Scraping

19. A farmer kept milk in an air tight container.

Which of the following methods of preservation will increase the shelf-life of the milk?

A Fermentation
B Pasteurization
C Refrigeration
D Sterilization

The diagram below illustrates characteristics of a beef breed. Use it to answer question 20.



20. Which characteristic of a good beef breed is shown?

A Blocky body
B Short horns
C Small head
D Long tail

21. At what age are dairy calves weaned?

A 7 days
B 7 weeks
C 7 months
D 7 years

22. Which of the following chemicals is used to control external parasites in beef cattle?

A Acaricides
B Avicide
C Fungicide
D Insecticide

23. Which of the following sets of organisms are all internal parasites?

A	Liver fluke	roundworm	tapeworm
B	Liver fluke	tapeworm	tsetsefly
C	Tsetsefly	liver fluke	tick
D	Roundworm	tick	tsetsefly

24. It is important to observe signs of heat in animals in order to

A inseminate.
B collect semen.
C control diseases.
D milk them properly.

25. Which of the following signs show ill-health in animals?

A Alertness
B Glossy fur
C Lameness
D Bright eyes

26. A layer feeds on 60 g of layers mash a day.

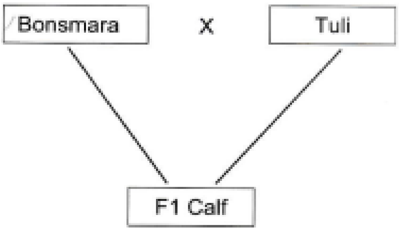
How much layers mash will be fed to 100 layers in 5 days?

A 5 kg
B 30 kg
C 60 kg
D 100 kg

27. Which of the following may cause a layer to produce soft-shelled eggs?

A Shortage of calcium
B Lack of exercising
C Laying at night
D Parasite attack

The diagram below illustrates a method of breeding in cattle production. Use it to answer questions 28 and 29.



28. Which breeding method is illustrated in the diagram?

A Upgrading
B Inbreeding
C Pure breeding
D Cross breeding

29. Which of the following terms describes the F1 calf?

A Hybrid
B Pedigree
C Line breed
D Pure breed

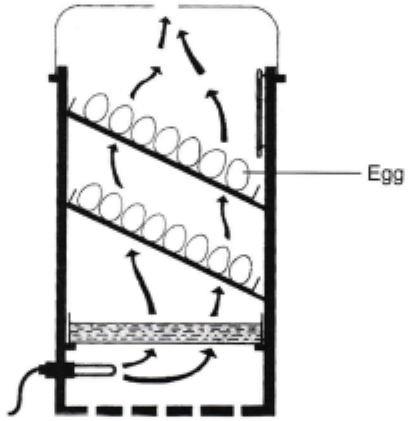
30. Which of the following equipment is used to test for mastitis in milk?

A Churn
B Indicator
C Strip cup
D Separator

31. Which of the following is a milk product?

A Margarine
B Oil
C Tallow
D Yoghurt

The diagram below shows an equipment used in poultry production. Use it to answer questions 32 and 33.



32. What is the use of the equipment?

A Brooding
B Cleaning
C Incubation
D Packaging

33. For how long are chicken eggs kept in the equipment?

A 14 days
B 21 days
C 24 days
D 30 days

The diagram below shows the head of a hen before and after a management practice was carried out. Use it to answer question 34.

34. What management practice has been carried out on the hen?

A Debeaking
B Evisceration
C Scalding
D Vaccination

The diagram below shows animal feed harvested on a farm. Use it to answer question 35.

35. Which type of animal feed is shown in the diagram?

A Succulent roughage
B Dry roughage
C Concentrate
D Silage

Use the information below to answer question 36.

Two dairy products **Q** and **R** were available to customers at the same price, but customers bought more of **R** than **Q**.

36. Which determinant could be used to explain the consumer's actions?

A Consumer's preference
B Price of related product
C Price of the commodity
D Consumer's income

37. Which of the following is **not** a function of management?

A Control
B Implementation
C Packaging
D Planning

38. A business owned by an individual is described as a

A business partnership.
B sole proprietorship.
C limited company.
D public company.

Use the information below to answer question 39.

Data showed that when the price of watermelons was P50 villagers bought 1 000 watermelons, when the price was P100 only 500 watermelons were bought.

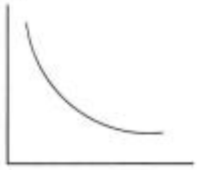
39. Which graph represents the demand of watermelons in the village?

A

B

C

D



40. Which of the following best describes budgeting?

A Keeping records
B Buying commodities
C Making arrangements for selling
D Making financial plans for a business

APPENDIX B

INTRODUCTORY LETTER

UNIVERSITY OF CAPE COAST
COLLEGE OF EDUCATION STUDIES
FACULTY OF EDUCATIONAL FOUNDATIONS
DEPARTMENT OF EDUCATION AND PSYCHOLOGY

Telephone: 0332091697
Email: dep@ucc.edu.gh

UNIVERSITY POST OFFICE
CAPE COAST, GHANA

Our Ref: DEP/72^D/Vol. 1

Your Ref:

3rd October, 2022

The Chairman
Institutional Review Board
University of Cape Coast
Cape Coast

Dear Sir/Madam,

INTRODUCTION

I, Professor Mark Owusu Amponsah is the Head of Department of Education and Psychology, University of Cape Coast.

I wish to humbly indicate that, Tapela Bulala with registration number EF/MEE/20/0008 is a PhD student studying Measurement and Evaluation in the Department of Education and Psychology, University of Cape Coast.

I humbly request that you provide him with the necessary assistance. I hope this letter will be considered favourably.

Thank you.

Yours faithfully

Professor Mark Owusu Amponsah
HEAD

APPENDIX C

UCC IRB ETHICAL CLEARANCE

UNIVERSITY OF CAPE COAST

INSTITUTIONAL REVIEW BOARD SECRETARIAT

TEL: 0558093143 / 0508878309
E-MAIL: irb@ucc.edu.gh
OUR REF: IRB/C3/Vol.1/0046
YOUR REF:
OMB NO: 0990-0279
IORG #: IORG0011497

17TH FEBRUARY 2023

Mr Tapela Bulala
Department of Educational Foundations
University of Cape Coast

Dear Mr Bulala,

ETHICAL CLEARANCE – ID (UCCIRB/CES/2022/112)

The University of Cape Coast Institutional Review Board (UCCIRB) has granted Provisional Approval for the implementation of your research on ***Assessing the Psychometric Properties of the Botswana Junior Secondary Agriculture Multiple-choice Items using Item Response Theory.*** This approval is valid from 17th February 2023 to 16th February 2024. You may apply for a renewal subject to the submission of all the required documents that will be prescribed by the UCCIRB.

Please note that any modification to the project must be submitted to the UCCIRB for review and approval before its implementation. You are required to submit a periodic review of the protocol to the Board and a final full review to the UCCIRB on completion of the research. The UCCIRB may observe or cause to be observed procedures and records of the research during and after implementation.

You are also required to report all serious adverse events related to this study to the UCCIRB within seven days verbally and fourteen days in writing.

Always quote the protocol identification number in all future correspondence with us in relation to this protocol.

Yours faithfully,

Kofi F. Amuquandoh

Ag. UCCIRB Administrator

ADMINISTRATOR
INSTITUTIONAL REVIEW BOARD
UNIVERSITY OF CAPE COAST

APPENDIX D

BEC IRB CLEARANCE

BOTSWANA
EXAMINATIONS
COUNCIL

Reference No: BEC/10/1/2/ VII (126)

24 February 2023

Tapela Bulala (Mr)
Private Bag 0027
Gaborone

Dear Sir

RE: REQUEST FOR JCE AGRICULTURE MULTIPLE CHOICE DATA

Reference is made to your letter dated 29th November 2021, on the subject matter.

This letter serves to grant you permission to use the 2018, 2019 and 2020 JCE Agriculture data on your PhD research project, titled: **Assessing the Psychometric Properties of Junior Certificate Examination for Agriculture Multiple Choice Items: The Case of Botswana.**

The data has been sent to you as an attachment through an e-mail. We thank you for seeking permission to use the data. You are reminded that it is your responsibility to ensure that all research ethics are adhered to, during the use of the data. Your research should observe among others, data security, confidentiality and oath of secrecy followed by Botswana Examinations Council (BEC).

BEC requires that you submit a copy of your report after completing the study.

Thank you.

Yours faithfully

Chawangwa Mudongo
for/**Chief Executive Officer**

Plot 54864, KT Motsete Rd, Private Bag 0070, Gaborone
Tel: +267 365 0700 Fax: +267 318 5011

APPENDIX E

LOCAL INDEPENDENCE

LOCAL INDEPENDENCE FOR 2018 DATA SET

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39
Q2	.040																																						
Q3	.121	.013																																					
Q4	.166	.007	.136																																				
Q5	.095	.025	.088	.163																																			
Q6	.137	.029	.121	.195	.099																																		
Q7	.009	.008	.007	.011	.037	.012																																	
Q8	.068	.021	.047	.075	.039	.075	.004																																
Q9	.062	.020	.062	.040	.045	.073	.015	.004																															
Q10	.115	.017	.086	.165	.107	.157	.007	.069	.071																														
Q11	.066	.000	.060	.098	.110	.071	.034	.025	.033	.070																													
Q12	.113	.032	.089	.122	.087	.129	.000	.077	.117	.110	.077																												
Q13	.175	.012	.147	.276	.171	.225	.019	.063	.052	.174	.130	.132																											
Q14	.095	.023	.072	.106	.076	.120	.007	.069	.103	.102	.040	.150	.104																										
Q15	.139	.013	.110	.197	.160	.161	.013	.059	.057	.150	.103	.133	.223	.111																									
Q16	.028	.004	.026	.035	.032	.027	.003	.001	.038	.024	.019	.030	.025	.014	.049																								
Q17	.126	.028	.096	.160	.112	.150	.004	.099	.096	.128	.062	.163	.147	.162	.134	.075																							
Q18	.157	.027	.120	.195	.131	.180	.003	.095	.106	.147	.090	.175	.191	.149	.178	.045	.205																						
Q19	.156	.020	.126	.213	.148	.191	.007	.081	.079	.165	.107	.156	.238	.140	.204	.043	.186	.335																					
Q20	.012	.019	.015	.016	.003	.011	.000	.006	.011	.007	.002	.019	.032	.028	.022	.016	.011	.017	.024																				
Q21	.074	.008	.054	.088	.060	.079	.009	.033	.034	.076	.050	.076	.094	.057	.090	.024	.087	.089	.097	.022																			
Q22	.051	.021	.045	.035	.046	.066	.009	.057	.093	.060	.002	.097	.009	.082	.051	.040	.118	.091	.038	.004	.041																		
Q23	.131	.010	.103	.199	.134	.155	.019	.069	.061	.147	.119	.104	.258	.068	.170	.025	.131	.167	.187	.003	.071	.027																	
Q24	.129	.012	.101	.180	.139	.142	.009	.064	.055	.144	.128	.114	.176	.097	.175	.039	.157	.172	.191	.013	.092	.051	.162																
Q25	.003	.001	.013	.003	.013	.002	.003	.001	.030	.005	.018	.007	.010	.002	.012	.007	.006	.006	.002	.010	.009	.008	.011	.012															
Q26	.136	.031	.112	.167	.103	.167	.011	.094	.110	.142	.061	.170	.171	.161	.158	.049	.214	.205	.198	.016	.101	.101	.151	.145	.028														
Q27	.152	.005	.145	.241	.144	.174	.019	.033	.036	.144	.140	.093	.364	.080	.214	.030	.109	.163	.224	.049	.090	.057	.218	.170	.012	.135													
Q28	.013	.012	.018	.013	.020	.012	.000	.013	.019	.012	.012	.029	.005	.035	.019	.017	.037	.027	.016	.013	.010	.033	.001	.009	.019	.029	.029												
Q29	.171	.038	.144	.258	.159	.218	.014	.097	.095	.179	.110	.173	.277	.120	.218	.025	.204	.226	.228	.016	.097	.085	.209	.191	.009	.214	.210	.001											
Q30	.070	.014	.059	.092	.053	.075	.010	.036	.042	.066	.044	.060	.096	.058	.070	.009	.083	.079	.076	.003	.032	.040	.074	.065	.006	.089	.074	.008	.088										
Q31	.092	.012	.064	.109	.068	.092	.011	.051	.056	.089	.061	.104	.107	.089	.111	.024	.109	.118	.123	.021	.113	.039	.095	.105	.007	.119	.107	.007	.114	.054									
Q32	.030	.017	.036	.048	.035	.039	.009	.011	.006	.030	.063	.015	.100	.012	.049	.011	.004	.026	.057	.026	.028	.058	.052	.037	.011	.014	.146	.000	.028	.010	.068								
Q33	.019	.018	.018	.029	.023	.025	.004	.023	.033	.028	.004	.062	.016	.037	.023	.004	.045	.037	.026	.006	.021	.044	.020	.009	.016	.047	.016	.006	.055	.014	.002	.032							
Q34	.175	.026	.140	.264	.178	.222	.014	.083	.078	.179	.141	.145	.325	.115	.215	.032	.177	.228	.254	.003	.089	.037	.242	.201	.011	.193	.269	.008	.289	.103	.126	.059	.020						
Q35	.115	.020	.089	.160	.097	.157	.004	.091	.091	.127	.065	.137	.155	.107	.138	.030	.151	.176	.158	.009	.070	.089	.133	.129	.013	.167	.118	.018	.196	.066	.094	.013	.028	.188					
Q36	.064	.022	.056	.081	.049	.078	.003	.053	.077	.063	.046	.097	.061	.075	.064	.014	.083	.091	.074	.021	.035	.076	.059	.067	.012	.095	.043	.013	.127	.045	.039	.012	.025	.085	.058				
Q37	.062	.016	.056	.093	.059	.085	.008	.061	.063	.070	.039	.083	.083	.058	.077	.020	.096	.104	.094	.013	.025	.074	.068	.079	.009	.101	.049	.023	.136	.052	.041	.010	.018	.102	.105	.076			
Q38	.105	.034	.075	.117	.062	.109	.001	.052	.063	.096	.030	.108	.114	.094	.113	.016	.104	.116	.109	.011	.067	.073	.092	.093	.005	.126	.093	.004	.152	.051	.073	.013	.029	.113	.108	.079	.065		
Q39	.135	.026	.119	.186	.098	.173	.010	.083	.105	.142	.050	.163	.174	.146	.155	.031	.188	.193	.182	.014	.087	.107	.141	.142	.008	.197	.139	.027	.220	.088	.092	.002	.035	.183	.173	.093	.116	.160	
Q40	.116	.012	.094	.165	.085	.144	.009	.058	.063	.116	.068	.093	.191	.083	.126	.028	.127	.136	.146	.047	.063	.048	.138	.117	.008	.125	.161	.006	.167	.074	.082	.047	.009	.184	.127	.054	.085	.091	.153

LOCAL INDEPENDENCE FOR 2019 DATA SET

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39
Q2	.129																																						
Q3	.194	.191																																					
Q4	.156	.131	.190																																				
Q5	.103	.108	.114	.054																																			
Q6	.066	.050	.084	.045	.062																																		
Q7	.081	.051	.081	.089	.005	.035																																	
Q8	.142	.109	.148	.107	.130	.040	.034																																
Q9	.119	.094	.133	.086	.127	.037	.026	.127																															
Q10	.058	.041	.072	.063	.011	.015	.036	.037	.019																														
Q11	.083	.061	.109	.071	.032	.048	.038	.039	.053	.015																													
Q12	.103	.128	.183	.113	.151	.043	.041	.113	.122	.025	.068																												
Q13	.194	.161	.246	.150	.178	.064	.060	.187	.158	.042	.081	.193																											
Q14	.108	.101	.121	.073	.162	.039	.020	.121	.111	.021	.024	.119	.174																										
Q15	.142	.088	.151	.170	.028	.029	.093	.096	.056	.068	.061	.043	.120	.065																									
Q16	.065	.050	.078	.074	.007	.031	.052	.014	.022	.030	.051	.032	.005	.018	.066																								
Q17	.236	.157	.251	.179	.140	.068	.081	.201	.155	.063	.090	.136	.263	.148	.183	.062																							
Q18	.164	.152	.189	.081	.209	.076	.028	.174	.184	.010	.066	.167	.260	.197	.055	.002	.234																						
Q19	.022	.014	.011	.022	.040	.006	.018	.041	.021	.002	.001	.030	.027	.026	.008	.011	.030	.036																					
Q20	.168	.131	.194	.142	.129	.053	.068	.160	.128	.049	.070	.135	.199	.119	.114	.042	.223	.164	.042																				
Q21	.066	.048	.071	.051	.051	.021	.017	.060	.063	.023	.014	.042	.080	.059	.043	.019	.093	.076	.008	.077																			
Q22	.182	.156	.236	.141	.177	.082	.057	.178	.165	.051	.071	.180	.258	.154	.099	.034	.258	.250	.026	.216	.075																		
Q23	.104	.104	.096	.039	.263	.045	.005	.143	.170	.019	.019	.168	.184	.181	.017	.007	.133	.285	.061	.122	.063	.192																	
Q24	.102	.088	.098	.048	.202	.035	.013	.124	.130	.006	.023	.131	.154	.132	.014	.001	.123	.204	.035	.110	.051	.154	.323																
Q25	.164	.166	.222	.119	.186	.067	.039	.171	.178	.037	.075	.183	.250	.168	.086	.026	.240	.272	.018	.192	.070	.251	.207	.182															
Q26	.178	.143	.217	.130	.168	.072	.051	.158	.159	.038	.076	.214	.238	.148	.083	.040	.212	.225	.023	.187	.068	.242	.191	.158	.237														
Q27	.194	.147	.244	.161	.149	.064	.056	.154	.144	.036	.076	.163	.233	.147	.113	.050	.229	.221	.023	.188	.081	.233	.159	.134	.225	.216													
Q28	.010	.001	.012	.013	.023	.009	.011	.009	.002	.003	.008	.002	.015	.010	.010	.012	.011	.028	.010	.008	.003	.011	.036	.021	.023	.015	.009												
Q29	.002	.002	.001	.004	.006	.005	.022	.021	.008	.011	.009	.004	.025	.003	.021	.001	.022	.002	.012	.016	.001	.009	.019	.005	.014	.004	.005	.023											
Q30	.191	.143	.236	.165	.135	.064	.072	.157	.144	.050	.080	.146	.237	.142	.127	.054	.247	.224	.016	.185	.079	.223	.148	.125	.231	.195	.212	.006	.014										
Q31	.035	.035	.049	.013	.092	.019	.010	.044	.054	.001	.001	.083	.059	.077	.002	.001	.032	.087	.020	.032	.016	.064	.142	.099	.083	.080	.060	.004	.020	.051									
Q32	.140	.095	.163	.132	.078	.048	.050	.091	.099	.025	.076	.112	.151	.086	.105	.060	.158	.130	.016	.125	.048	.141	.089	.074	.129	.140	.136	.014	.003	.144	.047								
Q33	.001	.008	.010	.018	.032	.005	.005	.014	.011	.007	.014	.013	.018	.024	.008	.009	.004	.033	.013	.007	.001	.010	.038	.034	.017	.014	.014	.006	.003	.003	.039	.038							
Q34	.178	.136	.200	.143	.161	.054	.059	.156	.145	.044	.059	.152	.225	.161	.102	.023	.233	.222	.047	.177	.078	.229	.190	.154	.222	.215	.221	.001	.001	.202	.063	.129	.044						
Q35	.141	.138	.196	.118	.133	.055	.043	.131	.134	.033	.066	.143	.212	.140	.077	.028	.197	.228	.033	.159	.055	.207	.158	.133	.217	.179	.192	.016	.012	.186	.060	.124	.001	.175					
Q36	.080	.057	.102	.078	.044	.014	.036	.057	.040	.023	.036	.049	.077	.048	.061	.032	.076	.065	.000	.065	.033	.073	.042	.045	.068	.061	.080	.002	.001	.080	.023	.064	.009	.077	.029				
Q37	.161	.140	.201	.117	.141	.069	.050	.170	.149	.038	.062	.139	.236	.150	.102	.027	.230	.231	.025	.172	.070	.228	.162	.140	.228	.193	.208	.006	.023	.219	.055	.131	.015	.194	.187	.071			
Q38	.149	.113	.187	.149	.082	.037	.063	.104	.094	.053	.055	.103	.150	.093	.114	.053	.175	.125	.012	.130	.055	.148	.081	.078	.140	.128	.144	.005	.001	.160	.046	.108	.009	.143	.126	.079	.148		
Q39	.096	.068	.100	.080	.064	.029	.035	.077	.065	.034	.032	.060	.097	.071	.083	.032	.130	.081	.021	.085	.028	.099	.078	.060	.099	.077	.098	.007	.010	.109	.021	.079	.006	.097	.071	.042	.086	.107	
Q40	.127	.105	.171	.127	.086	.047	.044	.103	.096	.037	.062	.083	.162	.091	.107	.045	.168	.137	.012	.124	.048	.153	.095	.078	.153	.121	.138	.007	.003	.155	.040	.092	.001	.137	.130	.060	.135	.130	.220

LOCAL INDEPENDENCE FOR 2020 DATA SET

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	Q21	Q22	Q23	Q24	Q25	Q26	Q27	Q28	Q29	Q30	Q31	Q32	Q33	Q34	Q35	Q36	Q37	Q38	Q39
Q2	.191																																						
Q3	.346	.186																																					
Q4	.099	.068	.076																																				
Q5	.294	.150	.280	.170																																			
Q6	.174	.097	.195	.112	.207																																		
Q7	.195	.135	.178	.071	.130	.091																																	
Q8	.251	.142	.239	.090	.206	.193	.151																																
Q9	.272	.157	.250	.099	.217	.151	.229	.201																															
Q10	.309	.171	.301	.121	.264	.234	.167	.271	.252																														
Q11	.150	.051	.139	.055	.117	.094	.067	.105	.118	.128																													
Q12	.278	.138	.267	.096	.224	.174	.137	.194	.196	.245	.150																												
Q13	.178	.111	.171	.081	.138	.110	.113	.139	.137	.175	.061	.169																											
Q14	.105	.062	.093	.052	.077	.070	.053	.075	.059	.097	.036	.071	.087																										
Q15	.238	.129	.249	.099	.235	.235	.126	.217	.206	.276	.096	.208	.137	.081																									
Q16	.083	.066	.094	.072	.075	.105	.070	.094	.085	.116	.046	.083	.073	.029	.100																								
Q17	.259	.146	.267	.098	.222	.220	.139	.223	.205	.278	.102	.218	.148	.091	.259	.116																							
Q18	.202	.094	.164	.046	.139	.051	.130	.110	.127	.131	.070	.127	.097	.036	.099	.035	.120																						
Q19	.048	.003	.043	.006	.036	.049	.011	.043	.022	.051	.020	.032	.032	.010	.044	.002	.032	.038																					
Q20	.226	.110	.230	.079	.196	.177	.120	.179	.180	.222	.096	.181	.113	.057	.193	.087	.199	.110	.024																				
Q21	.121	.082	.109	.040	.094	.047	.096	.080	.090	.098	.050	.094	.060	.040	.071	.030	.078	.073	.011	.059																			
Q22	.218	.123	.194	.043	.153	.084	.150	.135	.157	.170	.066	.145	.106	.057	.126	.050	.147	.114	.001	.129	.061																		
Q23	.208	.130	.239	.133	.220	.284	.113	.221	.202	.299	.113	.218	.148	.076	.277	.145	.281	.068	.055	.218	.055	.092																	
Q24	.236	.118	.231	.079	.193	.154	.139	.176	.183	.214	.098	.183	.113	.062	.178	.084	.181	.114	.016	.174	.080	.110	.172																
Q25	.301	.167	.296	.078	.240	.162	.157	.205	.217	.254	.100	.234	.157	.090	.237	.065	.228	.158	.031	.185	.102	.178	.193	.196															
Q26	.136	.087	.132	.059	.111	.083	.109	.117	.117	.156	.054	.106	.085	.056	.112	.052	.113	.087	.002	.075	.057	.081	.100	.112	.138														
Q27	.244	.123	.247	.094	.213	.215	.114	.190	.190	.251	.114	.203	.123	.073	.231	.106	.224	.103	.046	.199	.065	.113	.265	.187	.200	.098													
Q28	.034	.026	.036	.057	.051	.062	.011	.031	.030	.049	.023	.052	.031	.021	.040	.036	.035	.032	.001	.047	.011	.001	.075	.033	.023	.035	.071												
Q29	.270	.141	.252	.077	.202	.155	.153	.186	.204	.226	.096	.205	.132	.077	.174	.076	.211	.133	.022	.179	.104	.165	.175	.198	.227	.116	.165	.091											
Q30	.153	.086	.155	.058	.123	.087	.118	.106	.124	.122	.058	.115	.088	.048	.112	.042	.124	.075	.012	.116	.101	.137	.112	.097	.124	.064	.097	.014	.144										
Q31	.148	.094	.180	.116	.168	.225	.086	.167	.140	.209	.084	.163	.110	.066	.250	.121	.207	.063	.040	.139	.047	.070	.273	.132	.165	.079	.197	.040	.129	.094									
Q32	.115	.059	.117	.047	.096	.119	.058	.100	.098	.132	.041	.085	.063	.036	.106	.055	.109	.042	.019	.097	.027	.056	.137	.094	.092	.051	.119	.027	.092	.061	.116								
Q33	.145	.092	.139	.056	.115	.097	.124	.124	.121	.159	.051	.103	.087	.048	.112	.053	.119	.076	.011	.099	.042	.097	.122	.105	.114	.083	.109	.025	.127	.088	.093	.120							
Q34	.231	.136	.253	.124	.226	.271	.124	.234	.220	.302	.107	.209	.141	.075	.287	.133	.289	.076	.062	.210	.065	.130	.357	.189	.213	.108	.252	.053	.201	.124	.278	.169	.166						
Q35	.088	.078	.095	.063	.092	.077	.056	.072	.092	.092	.054	.101	.059	.033	.106	.048	.085	.057	.010	.062	.033	.042	.102	.070	.093	.047	.089	.025	.063	.056	.100	.044	.032	.103					
Q36	.309	.168	.303	.107	.276	.208	.162	.233	.225	.309	.117	.241	.164	.092	.257	.083	.245	.149	.043	.202	.104	.166	.246	.221	.282	.129	.238	.043	.236	.115	.193	.113	.131	.224	.097				
Q37	.130	.063	.107	.013	.070	.023	.081	.072	.078	.068	.039	.083	.061	.036	.060	.018	.056	.092	.005	.049	.050	.083	.025	.076	.109	.057	.049	.000	.089	.055	.024	.016	.052	.034	.042	.073			
Q38	.308	.179	.312	.102	.283	.232	.177	.263	.247	.340	.122	.248	.157	.082	.281	.096	.286	.142	.051	.211	.107	.188	.272	.212	.307	.137	.252	.022	.251	.148	.216	.128	.151	.298	.084	.345	.112		
Q39	.173	.098	.194	.070	.161	.145	.105	.152	.150	.202	.076	.157	.102	.053	.169	.065	.161	.082	.017	.123	.051	.094	.173	.127	.161	.084	.148	.023	.136	.084	.134	.075	.096	.180	.078	.192	.042	.209	
Q40	.241	.126	.239	.085	.225	.186	.120	.200	.181	.254	.091	.190	.133	.083	.229	.093	.219	.116	.038	.171	.078	.126	.223	.170	.230	.104	.208	.037	.186	.110	.190	.103	.103	.244	.083	.244	.070	.284	.143

APPENDIX F

MEASUREMENT INVARIANCE

GENDER (MALE vs. FEMALE)

Year	Models	χ^2/df	*p*-value
2018	Measurement weights	1.631	0.094
	Structural covariances	1.078	0.073
	Measurement residuals	16.735	0.000
2019	Measurement weights	2.143	0.056
	Structural covariances	2.094	0.055
	Measurement residuals	12.359	0.000
2020	Measurement weights	1.956	0.133
	Structural covariances	1.764	0.149
	Measurement residuals	21.134	0.000

SCHOOL LOCATION (RURAL vs. URBAN)

Year	Models	χ^2/df	*p*-value
2018	Measurement weights	1.872	0.118
	Structural covariances	1.885	0.127
	Measurement residuals	11.083	0.000
2019	Measurement weights	2.087	0.063
	Structural covariances	2.183	0.072
	Measurement residuals	27.569	0.000
2020	Measurement weights	2.870	0.052
	Structural covariances	2.262	0.063
	Measurement residuals	23.002	0.000

SCHOOL TYPE (PUBLIC vs. PRIVATE)

Year	Models	χ^2/df	*p*-value
2018	Measurement weights	1.844	0.109
	Structural covariances	1.986	0.113
	Measurement residuals	19.334	0.000
2019	Measurement weights	2.304	0.073
	Structural covariances	2.099	0.056
	Measurement residuals	21.052	0.000
2020	Measurement weights	2.871	0.057
	Structural covariances	2.901	0.058
	Measurement residuals	31.093	0.000

APPENDIX G

GENDER 3-PARAMETER ESTIMATES

2018 FEMALE 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.081	-0.115	0.780	0.561
Q2	0.014	-2.968	-0.141	0.405
Q3	**0.028**	**-0.790**	**0.549**	**0.618**
Q4	0.000	-0.971	1.160	0.755
Q5	0.000	-0.183	0.602	0.528
Q6	0.001	-0.276	0.867	0.560
Q7	0.199	41.281	0.046	0.302
Q8	0.075	2.312	0.759	0.211
Q9	0.191	1.563	2.191	0.217
Q10	0.031	-0.576	0.773	0.622
Q11	**0.001**	**-0.514**	**0.490**	**0.563**
Q12	0.181	0.820	1.586	0.356
Q13	0.000	-1.484	1.865	0.941
Q14	0.325	1.007	1.878	0.414
Q15	0.000	-0.757	0.887	0.662
Q16	0.293	2.465	1.931	0.299
Q17	0.086	0.875	1.339	0.302
Q18	0.275	0.259	2.217	0.536
Q19	0.316	-0.257	1.737	0.733
Q20	0.233	6.638	0.067	0.533
Q21	**0.008**	**1.506**	**0.452**	**0.342**
Q22	0.076	1.884	2.042	0.096
Q23	**0.175**	**-1.026**	**1.268**	**0.824**
Q24	**0.335**	**-0.375**	**1.623**	**0.766**
Q25	0.164	61.539	0.019	0.360
Q26	0.218	0.523	1.840	0.434
Q27	0.000	-2.073	1.749	0.974
Q28	0.234	2.783	2.008	0.237
Q29	0.130	k-0.096	1.778	0.602
Q30	0.114	1.951	0.645	0.310
Q31	0.463	0.728	1.415	0.604
Q32	**0.001**	**-8.489**	**0.124**	**0.741**
Q33	0.002	-1.795	-0.221	0.404
Q34	0.090	-0.712	1.714	0.792
Q35	0.143	0.669	1.213	0.406
Q36	0.142	2.010	1.149	0.219
Q37	0.181	1.737	1.108	0.285
Q38	0.265	0.930	1.054	0.466
Q39	0.147	0.451	1.345	0.448
Q40	**0.048**	**-0.631**	**0.712**	**0.629**

2018 MALE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.001	-0.080	0.772	0.516
Q2	0.074	-9.972	-0.059	0.405
Q3	**0.007**	**-0.473**	**0.672**	**0.582**
Q4	0.080	-0.484	1.381	0.688
Q5	0.000	-0.055	0.697	0.510
Q6	0.052	0.186	1.101	0.478
Q7	0.017	6.668	0.099	0.352
Q8	0.150	2.500	1.195	0.191
Q9	0.220	1.869	2.397	0.229
Q10	0.205	0.297	1.086	0.539
Q11	**0.000**	**-1.660**	**0.700**	**0.762**
Q12	0.231	1.276	2.205	0.275
Q13	0.000	-0.966	1.953	0.868
Q14	0.299	1.257	2.716	0.321
Q15	0.000	-0.400	1.087	0.607
Q16	0.270	2.458	1.534	0.286
Q17	0.131	1.142	1.703	0.239
Q18	0.210	0.400	1.974	0.457
Q19	0.144	-0.199	1.785	0.647
Q20	0.014	1.092	0.114	0.476
Q21	**0.007**	**1.514**	**0.464**	**0.336**
Q22	0.116	2.011	2.933	0.119
Q23	**0.000**	**-1.388**	**1.131**	**0.828**
Q24	**0.459**	**-0.418**	**1.861**	**0.830**
Q25	0.183	7.492	0.153	0.380
Q26	0.191	0.845	1.971	0.320
Q27	0.000	-1.380	2.211	0.955
Q28	0.196	2.746	2.328	0.197
Q29	0.142	0.019	1.842	0.564
Q30	0.027	1.931	0.439	0.319
Q31	0.335	0.228	0.975	0.631
Q32	**0.000**	**-2.525**	**0.402**	**0.734**
Q33	0.053	-9.271	-0.038	0.444
Q34	0.071	-0.595	1.750	0.758
Q35	0.157	1.030	1.309	0.330
Q36	0.204	1.929	1.534	0.243
Q37	0.192	1.867	1.191	0.271
Q38	0.277	1.184	1.364	0.397
Q39	0.159	0.910	1.788	0.297
Q40	**0.047**	**-0.239**	**0.856**	**0.572**

2019 FEMALE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.281	0.627	1.829	0.454
Q2	**0.059**	**0.394**	**0.775**	**0.458**
Q3	0.090	0.473	1.805	0.362
Q4	0.237	0.915	2.993	0.283
Q5	0.151	-2.423	1.456	0.976
Q6	**0.011**	**3.401**	**0.318**	**0.261**
Q7	0.198	1.981	2.080	0.211
Q8	**0.435**	**-0.441**	**1.352**	**0.799**
Q9	0.000	-1.939	0.713	0.800
Q10	0.196	2.576	0.942	0.262
Q11	0.136	2.741	1.054	0.182
Q12	0.000	-1.169	0.845	0.729
Q13	0.155	-0.588	1.650	0.768
Q14	0.000	-1.950	0.836	0.836
Q15	0.076	1.609	2.104	0.106
Q16	0.094	2.392	1.977	0.102
Q17	0.092	0.039	1.726	0.531
Q18	0.003	-1.647	1.707	0.943
Q19	**0.065**	**-0.319**	**0.152**	**0.544**
Q20	0.246	0.008	1.301	0.621
Q21	**0.013**	**1.342**	**0.308**	**0.406**
Q22	0.210	-0.536	1.685	0.772
Q23	0.000	-2.424	1.909	0.990
Q24	0.000	-2.845	1.020	0.948
Q25	0.000	-1.081	1.349	0.811
Q26	0.137	-0.500	1.323	0.706
Q27	0.242	-0.191	1.716	0.682
Q29	0.174	-29.560	-0.076	0.253
Q30	0.091	-0.173	1.293	0.596
Q31	0.001	-3.978	0.281	0.753
Q32	0.257	1.418	1.323	0.356
Q33	0.122	-3.114	-2.300	0.123
Q34	0.457	-0.305	2.615	0.832
Q35	0.001	-1.110	0.948	0.742
Q36	0.422	1.738	1.221	0.484
Q37	0.000	-0.828	1.032	0.702
Q38	0.237	0.830	1.566	0.401
Q39	0.305	1.429	1.024	0.435
Q40	0.135	0.532	0.878	0.468

2019 MALE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.229	0.603	1.858	0.419
Q2	**0.137**	**0.680**	**0.976**	**0.430**
Q3	0.169	0.476	1.959	0.404
Q4	0.298	1.003	2.741	0.340
Q5	0.006	-2.084	1.449	0.954
Q6	**0.003**	**2.786**	**0.336**	**0.283**
Q7	0.202	2.055	2.125	0.212
Q8	**0.243**	**-0.233**	**1.116**	**0.670**
Q9	0.000	-1.172	0.868	0.734
Q10	0.234	2.492	1.687	0.245
Q11	0.140	2.226	1.235	0.192
Q12	0.000	-1.404	1.076	0.819
Q13	0.190	-0.239	1.987	0.690
Q14	0.000	-1.592	0.853	0.795
Q15	0.074	1.721	2.922	0.080
Q16	0.143	2.241	2.436	0.147
Q17	0.130	0.468	2.009	0.374
Q18	0.132	-1.023	1.980	0.899
Q19	**0.009**	**-0.694**	**0.172**	**0.534**
Q20	0.230	0.193	1.550	0.558
Q21	**0.009**	**1.055**	**0.399**	**0.401**
Q22	0.158	-0.443	1.920	0.748
Q23	0.000	-1.724	2.833	0.992
Q24	0.000	-1.796	1.387	0.924
Q25	0.000	-0.570	1.373	0.686
Q26	0.070	-0.886	1.463	0.800
Q27	0.269	-0.234	1.987	0.718
Q28	0.270	-4.012	-1.083	0.279
Q29	0.099	-580.335	-0.002	0.297
Q30	0.119	0.203	1.554	0.491
Q31	0.000	-2.634	0.493	0.786
Q32	0.158	0.708	1.287	0.400
Q33	0.125	-3.214	-1.541	0.131
Q34	0.262	-0.521	1.841	0.796
Q35	0.183	-0.312	1.215	0.668
Q36	0.317	1.193	0.881	0.493
Q37	0.103	-0.162	1.354	0.601
Q38	0.242	1.004	1.413	0.390
Q39	0.238	1.281	1.058	0.394
Q40	0.218	0.871	1.362	0.401

2020 FEMALE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.136	0.111	2.667	0.505
Q2	0.279	0.577	1.293	0.511
Q3	0.174	-0.181	2.203	0.668
Q4	0.785	0.747	1.970	0.825
Q5	0.204	-0.373	1.618	0.719
Q6	0.000	-2.356	1.019	0.917
Q7	0.203	0.982	1.623	0.337
Q8	0.228	-0.729	1.213	0.774
Q9	0.192	-0.050	1.549	0.611
Q10	0.137	-0.868	1.925	0.863
Q11	0.000	-0.473	0.465	0.555
Q12	0.239	-0.002	1.635	0.620
Q13	0.555	0.398	2.157	0.687
Q14	**0.354**	**1.040**	**0.556**	**0.586**
Q15	0.000	-1.588	1.225	0.875
Q16	0.000	-3.098	0.370	0.759
Q17	0.211	-1.059	1.536	0.870
Q18	0.102	1.515	1.339	0.206
Q19	0.131	-3.916	-1.214	0.139
Q20	0.309	-0.708	1.341	0.807
Q21	0.188	1.804	1.295	0.259
Q22	0.131	0.860	1.095	0.375
Q23	0.079	-1.758	1.980	0.973
Q24	0.072	0.067	0.869	0.523
Q25	0.106	-0.094	1.740	0.589
Q26	**0.292**	**1.130**	**0.984**	**0.467**
Q27	0.000	-1.514	1.142	0.849
Q28	**0.009**	**-1.294**	**0.186**	**0.564**
Q29	0.160	0.024	1.413	0.573
Q30	**0.304**	**0.372**	**0.941**	**0.592**
Q31	**0.000**	**-2.607**	**1.006**	**0.932**
Q32	0.000	-3.127	-1.304	0.017
Q33	0.000	-0.800	0.596	0.617
Q34	**0.105**	**-1.668**	**2.321**	**0.982**
Q35	0.349	0.264	0.536	0.651
Q36	0.000	-0.628	1.481	0.717
Q37	**0.000**	**-1.436**	**-0.470**	**0.337**
Q38	**0.000**	**-0.954**	**1.832**	**0.852**
Q39	0.169	-1.110	0.739	0.746
Q40	0.374	-0.732	1.548	0.848

2020 MALE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.183	0.280	2.886	0.435
Q2	0.248	0.999	1.201	0.422
Q3	0.159	0.010	2.109	0.575
Q4	0.000	-2.378	0.729	0.850
Q5	0.170	-0.166	1.542	0.638
Q6	0.000	-1.468	1.509	0.902
Q7	0.199	1.212	1.970	0.266
Q8	0.164	-0.245	1.482	0.657
Q9	0.215	0.165	1.594	0.556
Q10	0.141	-0.350	2.048	0.718
Q11	0.000	-0.650	0.528	0.585
Q12	0.111	-0.209	1.390	0.620
Q13	0.401	0.240	1.390	0.651
Q14	**0.289**	**1.157**	**0.628**	**0.521**
Q15	0.000	-0.702	1.375	0.724
Q16	0.000	-1.848	0.657	0.771
Q17	0.149	-0.602	1.644	0.769
Q18	0.134	1.548	1.686	0.194
Q19	0.133	-3.961	-0.878	0.159
Q20	0.163	-0.915	1.310	0.806
Q21	0.175	2.149	1.207	0.232
Q22	0.205	1.325	1.578	0.293
Q23	0.072	-1.282	2.532	0.965
Q24	0.081	0.038	1.324	0.529
Q25	0.181	0.544	1.920	0.394
Q26	**0.257**	**1.125**	**1.282**	**0.399**
Q27	0.118	-0.857	1.629	0.825
Q28	0.001	-1.656	0.300	0.622
Q29	0.216	0.322	1.805	0.497
Q30	**0.301**	**0.738**	**1.063**	**0.520**
Q31	**0.000**	**-1.342**	**1.364**	**0.862**
Q32	0.000	-2.886	-1.277	0.024
Q33	0.003	-0.458	0.618	0.572
Q34	**0.226**	**-0.891**	**2.869**	**0.944**
Q35	0.000	-1.178	0.498	0.643
Q36	0.153	0.101	2.018	0.533
Q37	**0.000**	**-1.996**	**-0.304**	**0.353**
Q38	**0.157**	**-0.058**	**2.361**	**0.608**
Q39	0.131	-0.437	0.922	0.652
Q40	0.246	-0.324	1.711	0.725

APPENDIX H

LOCATION 3-PARAMETER ESTIMATES

2018 RURAL PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.052	-0.056	0.752	0.536
Q2	0.150	-6.777	-0.129	0.400
Q3	0.036	-0.582	0.664	0.610
Q4	0.003	-0.836	1.263	0.743
Q5	0.001	-0.162	0.707	0.529
Q6	0.009	-0.121	0.955	0.533
Q7	0.235	64.643	0.030	0.333
Q8	0.115	2.480	0.781	0.226
Q9	0.203	1.623	2.162	0.226
Q10	0.142	0.005	0.954	0.570
Q11	0.003	-1.405	0.498	0.669
Q12	0.227	0.975	1.924	0.330
Q13	0.000	-1.315	1.850	0.919
Q14	0.303	1.061	2.014	0.377
Q15	0.000	-0.688	1.061	0.675
Q16	0.266	2.643	1.135	0.300
Q17	0.100	0.918	1.489	0.283
Q18	0.224	0.260	2.098	0.509
Q19	0.187	-0.321	1.696	0.702
Q20	0.149	3.084	0.113	0.501
Q21	**0.096**	**1.820**	**0.605**	**0.322**
Q22	0.098	1.882	2.488	0.107
Q23	0.129	-1.202	1.116	0.819
Q24	**0.391**	**-0.439**	**1.812**	**0.810**
Q25	0.157	15.873	0.069	0.368
Q26	0.191	0.608	1.933	0.381
Q27	0.000	-1.666	2.044	0.968
Q28	0.208	2.774	2.055	0.211
Q29	0.114	-0.136	1.701	0.608
Q30	0.105	2.050	0.546	0.325
Q31	0.419	0.562	1.202	0.615
Q32	**0.004**	**-4.186**	**0.255**	**0.745**
Q33	0.048	-2.542	-0.165	0.426
Q34	0.084	-0.688	1.695	0.782
Q35	0.148	0.887	1.131	0.377
Q36	0.158	1.963	1.158	0.237
Q37	0.185	1.761	1.107	0.286
Q38	0.277	0.989	1.312	0.432
Q39	0.173	0.699	1.560	0.381
Q40	0.010	-0.537	0.761	0.605

2018 URBAN PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.003	-0.349	0.768	0.568
Q2	0.017	-3.362	-0.127	0.405
Q3	0.003	-0.826	0.552	0.613
Q4	0.008	-0.989	1.173	0.763
Q5	0.001	-0.048	0.512	0.507
Q6	0.000	-0.243	0.919	0.556
Q7	0.074	52.066	0.019	0.327
Q8	0.120	2.411	0.953	0.200
Q9	0.210	1.627	2.271	0.229
Q10	0.166	0.026	0.862	0.578
Q11	0.001	-1.475	0.411	0.647
Q12	0.197	0.979	1.865	0.309
Q13	0.000	-1.407	1.914	0.937
Q14	0.304	1.086	2.572	0.344
Q15	0.000	-0.591	0.830	0.620
Q16	0.275	2.521	1.710	0.285
Q17	0.106	0.946	1.438	0.289
Q18	0.211	0.219	1.909	0.524
Q19	0.200	-0.328	1.693	0.708
Q20	0.083	2.074	0.079	0.504
Q21	**0.040**	**1.874**	**0.436**	**0.334**
Q22	0.093	1.917	2.242	0.106
Q23	0.000	-1.570	1.097	0.849
Q24	**0.364**	**-0.370**	**1.604**	**0.774**
Q25	0.138	10.635	0.094	0.370
Q26	0.211	0.582	1.825	0.414
Q27	0.000	-1.837	1.932	0.972
Q28	0.022	12.339	0.112	0.217
Q29	0.133	-0.195	1.751	0.640
Q30	0.021	1.669	0.452	0.334
Q31	0.483	0.795	1.449	0.607
Q32	**0.001**	**-5.725**	**0.180**	**0.737**
Q33	0.011	-2.132	-0.160	0.422
Q34	0.084	-0.883	1.647	0.827
Q35	0.146	0.685	1.268	0.398
Q36	0.165	1.836	1.164	0.253
Q37	0.163	1.662	0.985	0.299
Q38	0.246	0.940	1.049	0.451
Q39	0.141	0.534	1.493	0.408
Q40	0.012	-0.776	0.766	0.649

2019 RURAL PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.228	0.550	1.709	0.445
Q2	**0.094**	**0.421**	**0.860**	**0.466**
Q3	0.149	0.405	1.937	0.416
Q4	0.267	0.892	2.678	0.328
Q5	0.005	-2.439	1.384	0.967
Q6	**0.046**	**3.252**	**0.324**	**0.293**
Q7	0.212	1.930	2.359	0.220
Q8	0.334	-0.448	1.269	0.759
Q9	0.001	-1.440	0.866	0.777
Q10	0.231	2.425	1.250	0.267
Q11	0.130	2.413	1.113	0.185
Q12	0.001	-1.569	0.922	0.810
Q13	0.216	-0.349	1.906	0.734
Q14	0.004	-1.831	0.825	0.820
Q15	0.079	1.687	2.292	0.098
Q16	0.117	2.312	1.973	0.126
Q17	0.105	0.157	1.850	0.488
Q18	0.048	-1.439	1.850	0.938
Q19	**0.042**	**-0.243**	**0.125**	**0.528**
Q20	0.277	0.140	1.532	0.600
Q21	**0.048**	**1.407**	**0.381**	**0.400**
Q22	0.130	-0.631	1.700	0.778
Q23	0.000	-2.049	2.337	0.992
Q24	0.000	-2.248	1.240	0.942
Q25	0.000	-0.874	1.367	0.768
Q26	0.120	-0.738	1.371	0.765
Q27	0.287	-0.234	1.893	0.721
Q28	0.140	48.389	0.033	0.284
Q29	0.142	-24.504	-0.072	0.268
Q30	0.108	-0.034	1.404	0.565
Q31	0.001	-3.482	0.385	0.793
Q32	0.155	0.875	1.022	0.400
Q33	0.125	-3.190	-1.891	0.127
Q34	0.385	-0.379	2.247	0.816
Q35	0.070	-0.737	1.064	0.709
Q36	0.394	1.467	1.075	0.498
Q37	0.059	-0.537	1.193	0.675
Q38	0.217	0.792	1.365	0.415
Q39	0.274	1.306	1.132	0.409
Q40	0.161	0.698	1.042	0.434

2019 URBAN PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.265	0.473	1.847	0.482
Q2	**0.149**	**0.773**	**0.851**	**0.439**
Q3	0.145	0.469	1.790	0.403
Q4	0.251	0.883	2.796	0.310
Q5	0.070	-2.423	1.371	0.968
Q6	**0.006**	**3.122**	**0.312**	**0.278**
Q7	0.189	1.947	1.837	0.211
Q8	0.338	-0.359	1.170	0.738
Q9	0.000	-1.628	0.794	0.785
Q10	0.216	2.732	1.072	0.256
Q11	0.145	2.598	1.072	0.195
Q12	0.145	-0.761	0.952	0.721
Q13	0.139	-0.552	1.748	0.762
Q14	0.000	-1.994	0.801	0.832
Q15	0.068	1.537	2.279	0.095
Q16	0.128	2.286	2.289	0.133
Q17	0.123	0.103	1.818	0.521
Q18	0.018	-1.592	1.799	0.947
Q19	**0.009**	**-0.965**	**0.165**	**0.544**
Q20	0.231	0.050	1.364	0.602
Q21	**0.003**	**0.963**	**0.355**	**0.417**
Q22	0.170	-0.622	1.704	0.786
Q23	0.000	-2.188	2.318	0.994
Q24	0.000	-2.276	1.200	0.939
Q25	0.000	-0.987	1.305	0.784
Q26	0.146	-0.631	1.366	0.746
Q27	0.236	-0.335	1.712	0.725
Q28	**0.204**	**-31.877**	**-0.078**	**0.265**
Q29	0.030	-19.540	-0.052	0.287
Q30	0.091	-0.153	1.351	0.592
Q31	0.000	-4.201	0.279	0.763
Q32	0.204	0.977	1.246	0.386
Q33	0.122	-3.096	-1.842	0.125
Q34	0.352	-0.542	2.132	0.845
Q35	0.051	-0.993	0.918	0.728
Q36	0.336	1.287	0.833	0.505
Q37	0.000	-0.676	1.068	0.673
Q38	0.236	0.845	1.435	0.411
Q39	0.249	1.069	1.013	0.439
Q40	0.173	0.561	1.070	0.466

2020 RURAL PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.186	0.189	2.806	0.488
Q2	0.220	0.670	1.139	0.468
Q3	0.160	-0.158	2.053	0.647
Q4	0.000	-2.875	0.627	0.859
Q5	0.168	-0.363	1.503	0.695
Q6	0.000	-1.821	1.291	0.913
Q7	0.207	0.953	1.911	0.318
Q8	0.129	-0.607	1.326	0.731
Q9	0.273	0.153	1.765	0.588
Q10	0.135	-0.600	2.082	0.807
Q11	0.002	-0.660	0.451	0.575
Q12	0.144	-0.215	1.502	0.640
Q13	0.445	0.133	1.613	0.693
Q14	0.218	0.440	0.516	0.565
Q15	0.000	-1.140	1.310	0.817
Q16	0.000	-2.225	0.468	0.739
Q17	0.157	-0.822	1.594	0.821
Q18	0.121	1.416	1.597	0.204
Q19	0.129	-3.314	-1.268	0.142
Q20	0.240	-0.802	1.370	0.810
Q21	0.164	1.837	1.264	0.238
Q22	0.177	1.115	1.288	0.335
Q23	0.000	-1.564	2.228	0.970
Q24	0.063	-0.055	1.038	0.545
Q25	0.167	0.305	1.867	0.468
Q26	**0.275**	**1.057**	**1.155**	**0.441**
Q27	0.000	-1.290	1.273	0.838
Q28	0.020	-1.288	0.264	0.592
Q29	0.204	0.143	1.653	0.555
Q30	0.299	0.383	1.064	0.579
Q31	**0.000**	**-1.698**	**1.365**	**0.910**
Q32	0.000	-3.051	-1.353	0.016
Q33	0.004	-0.758	0.620	0.617
Q34	**0.230**	**-1.163**	**2.756**	**0.970**
Q35	0.001	-1.461	0.426	0.651
Q36	0.126	-0.180	1.842	0.635
Q37	0.000	-1.742	-0.390	0.336
Q38	0.129	-0.498	2.176	0.780
Q39	0.140	-0.756	0.877	0.707
Q40	**0.266**	**-0.623**	**1.606**	**0.803**

2020 URBAN PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.164	0.089	2.584	0.534
Q2	0.259	0.631	1.294	0.486
Q3	0.178	-0.241	2.076	0.690
Q4	0.406	-1.658	0.668	0.853
Q5	0.189	-0.478	1.425	0.728
Q6	0.000	-2.141	1.150	0.921
Q7	0.202	1.147	1.891	0.284
Q8	0.221	-0.504	1.345	0.738
Q9	0.224	-0.001	1.538	0.612
Q10	0.148	-0.709	1.914	0.826
Q11	0.004	-0.763	0.462	0.589
Q12	0.190	-0.272	1.438	0.673
Q13	0.492	0.282	1.919	0.679
Q14	0.311	0.921	0.630	0.558
Q15	0.000	-1.349	1.380	0.865
Q16	0.000	-2.970	0.406	0.770
Q17	0.191	-0.903	1.569	0.842
Q18	0.116	1.515	1.291	0.226
Q19	0.140	-4.060	-1.178	0.147
Q20	0.196	-1.010	1.209	0.817
Q21	0.186	1.923	1.240	0.255
Q22	0.184	1.076	1.404	0.332
Q23	0.002	-1.707	2.089	0.973
Q24	0.001	-0.232	0.937	0.555
Q25	0.158	-0.015	1.936	0.585
Q26	**0.266**	**0.994**	**1.102**	**0.450**
Q27	0.000	-1.424	1.308	0.866
Q28	0.035	-1.950	0.175	0.599
Q29	0.198	0.112	1.581	0.563
Q30	0.320	0.620	0.946	0.563
Q31	**0.000**	**-2.031**	**1.327**	**0.937**
Q32	0.000	-3.016	-1.344	0.017
Q33	0.005	-0.522	0.578	0.577
Q34	**0.149**	**-1.306**	**2.654**	**0.974**
Q35	0.003	-1.772	0.450	0.691
Q36	0.139	-0.328	1.787	0.692
Q37	0.000	-1.539	-0.458	0.331
Q38	0.126	-0.641	2.035	0.814
Q39	0.081	-1.112	0.793	0.731
Q40	**0.219**	**-0.886**	**1.574**	**0.845**

APPENDIX I

SCHOOL TYPE 3-PARAMETER ESTIMATES

2018 PUBLIC PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.003	-0.233	0.762	0.545
Q2	0.027	-4.927	-0.095	0.402
Q3	**0.009**	**-0.634**	**0.630**	**0.602**
Q4	0.036	-0.726	1.290	0.729
Q5	0.001	-0.133	0.656	0.522
Q6	0.003	-0.116	0.967	0.529
Q7	0.078	17.098	0.058	0.326
Q8	0.128	2.424	0.972	0.203
Q9	0.208	1.684	2.285	0.225
Q10	0.133	-0.113	0.919	0.589
Q11	**0.000**	**-1.311**	**0.499**	**0.658**
Q12	0.210	1.020	1.794	0.319
Q13	0.000	-1.173	2.010	0.914
Q14	0.315	1.101	2.200	0.370
Q15	0.000	-0.589	1.013	0.645
Q16	0.285	2.476	1.680	0.296
Q17	0.111	0.983	1.483	0.279
Q18	**0.231**	**0.254**	**1.989**	**0.520**
Q19	**0.164**	**-0.406**	**1.636**	**0.716**
Q20	0.026	0.394	0.108	0.502
Q21	**0.021**	**1.546**	**0.470**	**0.340**
Q22	0.101	1.925	2.463	0.109
Q23	0.001	-1.410	1.132	0.832
Q24	**0.421**	**-0.382**	**1.686**	**0.801**
Q25	0.181	13.562	0.089	0.369
Q26	0.200	0.639	1.887	0.384
Q27	**0.000**	**-1.627**	**2.069**	**0.967**
Q28	0.215	2.762	2.028	0.217
Q29	0.135	-0.060	1.839	0.591
Q30	0.099	1.980	0.591	0.312
Q31	0.409	0.487	1.139	0.625
Q32	**0.001**	**-3.761**	**0.272**	**0.736**
Q33	0.008	-2.442	-0.137	0.422
Q34	0.077	-0.677	1.753	0.784
Q35	0.146	0.799	1.256	0.375
Q36	0.178	1.969	1.327	0.234
Q37	0.187	1.776	1.157	0.280
Q38	0.265	1.013	1.205	0.433
Q39	0.155	0.661	1.586	0.374
Q40	**0.002**	**-0.537**	**0.788**	**0.605**

2018 PRIVATE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.000	0.432	0.392	0.458
Q2	0.027	-6.627	-0.031	0.464
Q3	**0.521**	**2.529**	**0.992**	**0.557**
Q4	0.000	-0.937	0.690	0.656
Q5	0.178	1.253	0.546	0.453
Q6	0.000	0.743	0.435	0.420
Q7	0.031	-132.611	-0.006	0.344
Q8	0.101	2.774	1.321	0.124
Q9	0.025	11.774	0.127	0.204
Q10	0.187	0.956	0.641	0.472
Q11	**0.000**	**-1.969**	**0.333**	**0.658**
Q12	0.222	2.745	1.599	0.232
Q13	0.000	-2.018	0.919	0.865
Q14	0.296	2.398	6.105	0.296
Q15	0.008	0.327	0.336	0.477
Q16	0.074	81.404	0.018	0.251
Q17	0.128	2.246	1.574	0.153
Q18	**0.211**	**1.589**	**1.576**	**0.271**
Q19	**0.178**	**0.727**	**0.877**	**0.462**
Q20	0.000	2.432	-0.081	0.549
Q21	**0.000**	**2.033**	**0.422**	**0.298**
Q22	0.079	20.407	1.223	0.079
Q23	0.000	-1.366	0.657	0.710
Q24	**0.136**	**-0.431**	**0.905**	**0.651**
Q25	0.359	-3.277	-1.221	0.371
Q26	0.024	1.490	0.610	0.304
Q27	**0.000**	**-2.249**	**1.542**	**0.970**
Q28	0.000	-6.724	-0.209	0.197
Q29	0.146	0.562	0.886	0.469
Q30	0.120	3.250	0.347	0.335
Q31	0.508	1.655	3.006	0.511
Q32	**0.000**	**-7.401**	**0.152**	**0.755**
Q33	0.042	23.314	0.011	0.462
Q34	0.000	-0.528	0.924	0.620
Q35	0.003	2.868	0.320	0.287
Q36	0.175	2.809	1.338	0.194
Q37	0.196	2.265	1.264	0.239
Q38	0.271	2.777	0.448	0.434
Q39	0.000	1.360	0.480	0.342
Q40	**0.000**	**-2.223**	**0.153**	**0.584**

2019 PUBLIC PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.239	0.551	1.811	0.444
Q2	**0.103**	**0.527**	**0.886**	**0.449**
Q3	0.139	0.469	1.928	0.388
Q4	0.275	0.938	2.919	0.319
Q5	0.009	-2.251	1.538	0.970
Q6	**0.021**	**3.142**	**0.335**	**0.274**
Q7	0.201	1.982	2.128	0.213
Q8	0.252	-0.616	1.123	0.750
Q9	0.001	-1.467	0.829	0.772
Q10	0.226	2.530	1.291	0.255
Q11	0.138	2.477	1.104	0.191
Q12	0.001	-1.361	0.915	0.777
Q13	0.170	-0.432	1.856	0.743
Q14	0.001	-1.714	0.876	0.818
Q15	0.079	1.645	2.432	0.096
Q16	0.120	2.277	2.257	0.125
Q17	0.112	0.232	1.896	0.460
Q18	**0.099**	**-1.246**	**1.936**	**0.926**
Q19	**0.023**	**-0.730**	**0.163**	**0.541**
Q20	0.228	0.054	1.426	0.599
Q21	**0.020**	**1.190**	**0.360**	**0.406**
Q22	0.169	-0.525	1.810	0.768
Q23	**0.000**	**-1.961**	**2.438**	**0.992**
Q24	0.000	-2.126	1.290	0.940
Q25	0.018	-0.762	1.444	0.755
Q26	0.104	-0.736	1.331	0.756
Q27	0.257	-0.240	1.827	0.709
Q28	0.114	-22.227	-0.067	0.278
Q29	0.092	-51.263	-0.027	0.274
Q30	0.104	0.004	1.459	0.551
Q31	0.001	-3.111	0.383	0.767
Q32	0.184	0.974	1.143	0.386
Q33	0.123	-3.188	-1.748	0.126
Q34	0.350	-0.458	2.145	0.823
Q35	0.114	-0.654	1.105	0.710
Q36	0.374	1.464	0.987	0.493
Q37	0.051	-0.486	1.194	0.660
Q38	0.238	0.873	1.491	0.401
Q39	0.251	1.263	0.996	0.417
Q40	0.184	0.707	1.123	0.438

2019 PRIVATE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.435	2.086	6.848	0.435
Q2	**0.316**	**2.208**	**1.185**	**0.362**
Q3	0.145	1.360	1.304	0.269
Q4	0.213	1.893	2.180	0.226
Q5	0.006	-4.986	0.430	0.896
Q6	**0.001**	**4.342**	**0.274**	**0.234**
Q7	0.126	8.442	0.270	0.207
Q8	0.530	1.423	1.388	0.587
Q9	0.079	-2.160	0.399	0.726
Q10	0.190	2.812	1.481	0.202
Q11	0.137	2.693	1.919	0.141
Q12	0.367	0.365	1.019	0.625
Q13	0.273	0.376	0.896	0.576
Q14	0.005	-4.091	0.330	0.795
Q15	0.048	2.665	2.773	0.049
Q16	0.120	2.955	2.148	0.121
Q17	0.117	1.226	0.688	0.383
Q18	**0.000**	**-2.399**	**0.969**	**0.911**
Q19	**0.025**	**0.047**	**0.055**	**0.512**
Q20	0.000	-0.094	0.475	0.511
Q21	**0.000**	**2.084**	**0.285**	**0.356**
Q22	0.121	-0.485	0.984	0.663
Q23	**0.000**	**-2.783**	**1.871**	**0.995**
Q24	0.000	-3.769	0.609	0.908
Q25	0.000	-1.393	0.647	0.711
Q26	0.284	0.122	1.429	0.611
Q27	0.138	0.098	1.012	0.548
Q28	0.050	12.762	0.097	0.264
Q29	0.012	-7.127	-0.124	0.302
Q30	0.000	0.115	0.492	0.486
Q31	0.000	-5.250	0.259	0.796
Q32	0.311	2.054	1.451	0.345
Q33	0.125	5.649	0.853	0.132
Q34	0.000	-1.221	0.775	0.720
Q35	0.395	0.395	0.945	0.642
Q36	0.018	2.145	0.134	0.439
Q37	0.000	-0.531	0.751	0.598
Q38	0.295	2.218	1.358	0.328
Q39	0.409	2.479	1.331	0.430
Q40	0.368	2.312	1.216	0.403

2020 PUBLIC PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.165	0.182	2.779	0.479
Q2	0.255	0.731	1.280	0.465
Q3	0.169	-0.083	2.233	0.623
Q4	0.003	-2.868	0.612	0.853
Q5	0.181	-0.282	1.616	0.682
Q6	0.000	-1.756	1.325	0.911
Q7	0.197	1.065	1.759	0.304
Q8	0.141	-0.576	1.338	0.728
Q9	0.210	0.063	1.617	0.585
Q10	0.134	-0.596	2.076	0.805
Q11	0.004	-0.586	0.481	0.572
Q12	0.167	-0.137	1.469	0.625
Q13	0.470	0.263	1.671	0.678
Q14	**0.282**	**0.837**	**0.564**	**0.558**
Q15	0.000	-1.048	1.397	0.812
Q16	0.000	-2.285	0.512	0.763
Q17	0.138	-0.875	1.627	0.833
Q18	0.120	1.523	1.487	0.203
Q19	0.133	-3.785	-1.062	0.148
Q20	0.229	-0.840	1.301	0.806
Q21	0.178	1.928	1.226	0.248
Q22	0.178	1.093	1.300	0.338
Q23	0.050	-1.494	2.288	0.970
Q24	0.040	-0.053	1.008	0.533
Q25	0.155	0.221	1.909	0.489
Q26	**0.249**	**1.024**	**1.052**	**0.440**
Q27	0.049	-1.159	1.367	0.838
Q28	**0.008**	**-1.504**	**0.224**	**0.587**
Q29	0.198	0.188	1.652	0.537
Q30	**0.286**	**0.473**	**0.982**	**0.562**
Q31	0.000	-1.724	1.299	0.904
Q32	**0.000**	**-2.983**	**-1.292**	**0.021**
Q33	0.005	-0.615	0.622	0.597
Q34	**0.198**	**-1.186**	**2.740**	**0.970**
Q35	0.016	-1.322	0.448	0.650
Q36	0.134	-0.149	1.950	0.629
Q37	**0.001**	**-1.715**	**-0.386**	**0.341**
Q38	**0.126**	**-0.418**	**2.224**	**0.753**
Q39	0.082	-0.895	0.835	0.705
Q40	0.262	-0.591	1.655	0.798

2020 PRIVATE PARAMETERS 3PL

	Gussng	Dffclt	Dscrmn	P(x=1|z=0)
Q1	0.169	1.016	1.910	0.274
Q2	0.376	1.585	1.297	0.447
Q3	0.187	0.218	1.098	0.545
Q4	0.757	0.962	0.812	0.833
Q5	0.380	0.591	1.221	0.583
Q6	0.001	-3.572	0.608	0.898
Q7	0.229	1.849	1.983	0.248
Q8	0.221	-0.251	0.688	0.644
Q9	0.264	0.669	1.063	0.507
Q10	0.097	-0.795	0.868	0.699
Q11	0.002	-0.156	0.316	0.513
Q12	0.213	0.462	0.926	0.524
Q13	0.571	1.566	1.798	0.595
Q14	**0.316**	**2.206**	**0.444**	**0.503**
Q15	0.003	-1.794	0.676	0.772
Q16	0.000	-3.541	0.341	0.770
Q17	0.453	0.212	1.102	0.695
Q18	0.153	2.008	3.708	0.153
Q19	0.003	-11.712	-0.154	0.144
Q20	0.463	-0.117	1.253	0.751
Q21	0.211	2.447	4.719	0.211
Q22	0.195	1.728	2.952	0.200
Q23	0.001	-2.675	1.019	0.939
Q24	0.000	0.217	0.594	0.468
Q25	0.195	1.145	0.915	0.404
Q26	**0.418**	**2.172**	**12.780**	**0.418**
Q27	0.000	-1.631	0.935	0.821
Q28	**0.000**	**-3.790**	**0.226**	**0.702**
Q29	0.072	0.298	0.649	0.491
Q30	**0.472**	**1.744**	**3.181**	**0.474**
Q31	0.000	-2.895	0.775	0.904
Q32	**0.010**	**-3.396**	**-1.557**	**0.015**
Q33	0.000	-0.411	0.425	0.544
Q34	**0.000**	**-1.760**	**1.576**	**0.941**
Q35	0.000	-2.275	0.279	0.654
Q36	0.216	0.222	0.824	0.572
Q37	**0.008**	**-1.429**	**-0.151**	**0.450**
Q38	**0.000**	**-0.752**	**0.986**	**0.677**
Q39	0.001	-1.683	0.438	0.677
Q40	0.000	-1.838	0.610	0.754