ESJ Social Sciences

Using Classical Test and Item Response Theories to Evaluate Psychometric Quality of Teacher-Made Test in Ghana

Paul Kwame Butakor, PhD
University of Ghana, Ghana

Doi:10.19044/esj.2022.v18n1p139

Submitted: 19 October 2021
Accepted: 26 January 2022
Published: 31 January 2022

Cite As:
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

Abstract

Teaching, learning, and assessment are key concepts in education and their relationship can be seen as participants of a three-legged race. In this regard, classroom assessment practices such as teacher-made tests are important and meaningful when they support students' learning. The purpose of this study was to establish the psychometric quality of a teacher-made mathematics test item used in one of the Senior High Schools in Ghana. This study employed quantitative descriptive design where 400 selected students' responses to a teacher-made Mathematics test were collected and analyzed through various psychometric techniques. The results showed that the Mathematics test had low but acceptable reliability coefficient of 0.61. Also out of the 40 multiple-choice items, 26 were of satisfactory difficulty levels with only one test item found to be too difficult and three test items being too easy. The findings of the discrimination indices suggest that 25 test items had bad or weak discrimination indices and four items showed negative discrimination indices. The study further indicated that 30.8 percent of the options were functioning distractors whereas the majority of the options (69.2%) were non-functioning distractors.It is therefore recommended that in-service training on effective ways of developing test items should be organized regularly for in-service teachers to help improve of the quality of teacher-made tests across Senior High Schools in Ghana.

Keywords: Item difficulty, discrimination indices, teacher-made test, classical test theory, item response theory

Introduction

Assessment is key when it comes to education, specifically teaching and learning. It is worth acknowledging that the aim for which any assessment is developed and also validated remains a critical component of assessment (Awoniyi, 2016). Whereas formative assessment seeks to discover areas which need to be improved during the process of teaching-learning, summative assessment takes place after the process has ended, that is, at the end of term, year, or program for purposive decision making regarding the ended task, or completed program (Mensah, 2014). According to Amoako (2018), utilization of formative assessments remains a common methodology advocated within educational literature for improving teachers' pedagogical practices, and for providing specialized instructional support for less performing students. Literature has demonstrated that using formative assessment helps to improve instructional practices, identifies the gaps within the curriculum, and also contributes to the enhanced performance of students (Asare, 2015; Dunn & Mulvenon, 2009).

The utilization of formative assessments has even become more necessary particularly in the Covid-19 era, whereby the situation necessitated that schools be closed down due to the pandemic. Nonetheless, because Ghana overly relies on summative assessments for assessing final year students, students had to return to school to write their final examinations for instance, the Basic Education Certificate Examination (BECE) for Junior High Schools and the West Africa Senior School Certificate Examination (WASSCE) for Senior High School students to enable them progress to the next educational level. However, more emphasis can be placed on formative assessments which can be used in place of summative assessment in case of emergency in assessing students. Thus, it would not have been necessary for the Junior High School (JHS) and the Senior High school (SHS) final year students to return to school to write their exit examinations during this COVID-19 era if Ghana had a robust and trusted formative assessment regimes in our schools. It is therefore imperative to pay more attention to formative assessment practices in our schools moving forward as a country.

According to Quaigrain and Arhin (2017), teachers use tests/ assessments to understand the prowess of students regarding learning outcomes, whether they are affective, cognitive, or psychomotor. All teachers in whichever educational level they are, are responsible for preparing as well as administering numerous formalized teacher-made tests. According to Quaigrain and Arhin (2017), tests remain essential for providing feedback to teachers about their own performance and that of the students they teach. This makes test quality a crucial concern. Adhering strictly to the standards of test construction, administration, analysis, in addition to reporting remains relevant, particularly when there is the development of norm-referenced tests for instructional purposes (Mozaffer & Farhan, 2012).

In Ghana, there exist two main test formats. They include objective and essay tests, although objective tests (usually multiple-choice) are used more frequently (Quaigrain & Arhin, 2017). Designing multiple-choice questions (MCQs) to evaluate comprehensively the knowledge of students after each semester/term is very complex as well as time-consuming (Mozaffer & Farhan, 2012). Upon administering a test, a teacher must be able to determine test items' quality, and whether the items reflected the performance of the students with regards to those particular learning objectives taught over a period. According to Odili (2010), the interest in psychometric analyses of tests is due to the fact that education remains a mechanism for attaining equality among individuals. The situation where some examinees fail while others pass the examinations due to the difficulty encountered in the test has distorted the probability of failed examinees gaining admissions or promotion. Thus increasing class differentiation and the rate of school dropouts. This means that psychometric analysis of teacher-made tests should be key in the teaching profession to ensure that students are assessed fairly and objectively.

However, little attention has been paid to conducting psychometric analysis of teacher-made tests in Ghana. The objective of this study is to evaluate the psychometric properties of the teacher-made test using both classical test theory and item response theory approaches. Specifically, this study was conducted to answer the research question:what are the psychometric properties of the teacher-made mathematics test used in the selected Senior High School?

The Concept of Educational Assessments

Educational assessment involves the procedure of detailing the acquisition as well as mastery of knowledge to help in informed decisions making regarding the steps to be followed within an educational process (Ravela et al., 2009). Within this process, there is the consideration of students' aptitudes, learning styles, attitudes, progressions as well as outcomes. Decisions to be taken after the outcome may differ; from the implementation of system-wide programmes to improvement in classroom teaching/learning to modifying classroom instruction, or assessment of students' admission to higher education level such as the university (Clarke, 2011).

According to Ravela et al. (2009), effective educational assessment systems aid in the acquisition of quality information to meet decision-making needs, and to support and improve the learning of students. Similarly, Hockings (2010) asserted that educational assessment systems can inform as well as improve instruction and learning, assess progress, and offer partial accountability information. These objectives are anticipated to lead ultimately to improved educational quality. Salvia et al. (2012) recommended the involvement of students as associates in designing assessment so that engagement, as well as deep learning, will be increased among students. Also, Davies and Dempsey (2011) suggested the shifting away from the use of traditional assessment that undermines the capacity of students in judging their work to the implementation of research assessments which effectively prepare students for employment in future.

According to (Clarke, 2011) educational assessment could be categorized using the following distinctions: Classroom assessments, examinations, as well as large-scale assessments

Classroom Assessments

Classroom assessments are assessments that are undertaken by teachers and students as a daily activity (Heritage, 2010). They are made up of various standardized/non-standardized instruments as well as techniques for the gathering and interpretation of oral, written, and some other types of evidence for student learning/ achievement. Examples are; homework assignments, oral questions and feedback, diagnostic tests, student presentations, and quizzes. These assessments are aimed at providing 'real-time' data to assist in teaching and learning. According to Beziat and Coleman (2015), classroom assessment plays a very significant role in terms of how students learn, students' motivation to learn, and how teachers teach. Again, classroom assessment enables teachers/instructors to gain knowledge about what students have learnt and their misconceptions that can be used to plan as well as guide instruction and offer useful formative feedback to students.

Examinations

Examinations, often qualified by expressions such as 'end of cycle', 'public', or 'external' help to gather information for decisions to be made on individual students. Whether they are school-based or administered externally, their standardized nature ensures that every student is provided with equal opportunity to prove their level of knowledge or ability according to the curriculum, or any other acknowledged body of knowledge (Black & Wiliam,2010).

For instance, the exit certification examinations taken after compulsory education at either the basic and secondary levels are seen as "high-stakes" due to their significant influence in terms of what is taught as well as learnt. They might also influence the knowledge profile of graduates (Downer et al., 2010). The tests used in the exit examinations may be discouraging for candidates who are unable to perform well, hence contribute to them being excluded from their desired secondary schools or tertiary institutions. Disadvantaged groups of students are most vulnerable to these risks. Practices which create or promote inequities during these examinations occur in the form of scoring, exam registration fees, ranking, private tutoring, using certain language unfamiliar to the students and so on.

Forms of Assessment used in Ghana

Beside assessment in the classroom, popularly called the School-Based Assessment (SBA), and end of term or semester examinations, the various levels of grades (except Grade 9 and 12) including university and other tertiary levels have no external assessment scheme which evaluates the performance and achievement of students. For students in Grades 9 and 12, national examinations are conducted, these include the Basic Education Certificate Examination (BECE) after JHS 3 (Grade 9), and the West Africa Senior School Certificate Examination (WASSCE) after SHS 3 (Grade 12) (Mills & Mereku, 2016). Other assessments in Ghana, according to a report by the World Bank are large scale assessments conducted nationally. These include Early Grade Mathematics Assessment (EGMA), Early Grade Reading Assessment (EGRA) as well as international assessments conducted on a large scale – such as TIMSS (World Bank, 2013).

Quality of Teacher Made Test

Assessing students' learning is very essential in education. According to Bichi (2016), assessing students' academic skills, intellectual development, as well as cognitive abilities include some systems utilized in sampling students' performance with regards to a specific learning outcome. One of such systems is the test. A test is anticipated to illustrate students' performance (Gareis & Grant, 2015). Teacher-made tests are usually prepared and administered to measure the achievements of students in the classroom, evaluate the teaching method adopted by the teacher, in addition to other curricular programmes offered by the school. Teacher-made test thus remain one valuable instrument of the teacher in serving their purpose (O'Malley, 2010).

There are some principles that teachers should follow when designing teacher-made test. These include: the test items should be arranged based on levels of difficulty; the test is prepared by teachers that can be utilized for prognosis as well as diagnosis purpose; the test consists of the entire content area and is made up of several items; and item's preparation is done based on the blueprint (Gareis & Grant, 2015). This means that teacher-made tests are mostly used as instrument for formative assessment; and the teacher develops the test to determine the achievement as well as proficiency of the student in a particular subject.

In terms of preparation of the test, various forms of constructed response and objective test such as, short-answer, multiple-choice, and matching type could be constructed. After the construction, there should be a review of the test items by experts to ensure that the items are screened in terms of language, items' modalities, statements provided, accurate answers and distractors, in addition to other unintended errors. The recommendations from the experts would assist a test constructor to revise the items to make them much more acceptable and usable. After the test has been constructed, items must be organized from simple to complex order. In organizing the items, several approaches can be employed by the teacher, in terms of group, unit, topic and so on. There should also be the preparation of scoring key forthwith to prevent delay involved in the scoring. Furthermore, clear directions or instructions should be provided to prevent misunderstanding among the students (Gareis & Grant, 2015).

After careful planning, development, and administering the test, there is the need to conduct psychometric analysis on students test data to provide evidence to help evaluate and judge the quality of the test items. In this regard, there has been the development of testing theories that are helpful in evaluating tests with some indices. These indices mainly depend on two (2) common statistical frameworks. They are Classical Test Theory and Item Response Theory. These theories have a relationship with the item development process within the psychological and educational field to facilitate the quality of measuring instruments (Bichi, 2016).

Classical Test Theory (CTT)

CTT has been applied extensively over several years for determining reliability as well as additional features of measurement instruments. Hambleton and Russell (1993) asserted that CTT entails test scores that are made up of three concepts; test score (commonly known as observed score), true score; as well as error score. The fundamental equation of the CTT is:

$$X = T + E \qquad (1)$$

This is a simple linear model that links the observable test score (X) to the sum of two unobservable variables, true score (T) and error score (E).

The main assumptions which underline CTT include true scores, as well as error scores, remain uncorrelated, examinees obtain a zero average error score, and there is no correlation between the error scores of parallel tests. Also, Magno (2009) asserted that CTT assumes that examinees obtain true scores (unobservable) supposing there are no measurement errors. Nevertheless, because of the imperfection of instruments deployed, the observed score for every individual may vary from the true ability of an individual. This variance between the two scores (observed and true scores) results from a measurement error.

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

The standard measurement error is calculated as:

$$SE_M = S_X \sqrt{1} - R_{xx} \qquad (2)$$

SE_M=standard error of measurement
S_x = standard deviation of test scores
R_{xx}= reliability coefficient
Small SE_M indicates high reliability (Kaplan & Saccuzo, 1997).

CTT Statistics and Item Analysis

a. Item Difficulty

Item difficulty indicates the overall proportion of examinees that accurately answered an item. An item is considered easy when a large proportion of examinees accurately answer an item. Calculating the difficulty index of an item is carried out by dividing the sum of examinees accurately answering the item by the sum of examinees responding to the item. An item accurately answered by 75% of examinees would possess a difficulty index of .75. Also, an item that was accurately answered by 40% of examinees would possess item difficulty of .40 (Matlock-Hetzel, 1997). The higher the p-value, the easier the item and the vice versa. The item difficulty is denoted as p and given as:

$$P = \frac{R}{N} \qquad (3)$$

P = difficulty of an item
R = examinees who accurately respond to an item
N = sum of examinees
Guideline for interpreting item difficulty index is demonstrated in Table 1

Table 1. *Item Difficulty Indices Interpretation*

Difficulty Index (p)	**Interpretation**
$P < 0.20$	Very difficult
$0.20 \leq P \leq 0.40$	Difficult
$0.40 \leq p \leq 0.60$	Average
$0.60 < P > 0.80$	Easy
$P > 0.80$	Very Easy

Source: Hotui (2006)

b. Item Discrimination

It refers to a test item capability to discriminate between high and low ability examinees (Adegoke, 2013). Essentially, item discrimination aims at eliminating or dropping or modifying items which fail to function properly within a tested group (Courville, 2004).

The index of discrimination to ascertain an item's discriminating power could be calculated by two (2) indices: the item discrimination index (D), as well as the Item discrimination coefficient.

c. Item Discrimination Index (D)

This technique remains applicable in calculating the simple amount of an item's discriminating power by making use of the groups that are extreme (Matlock-Hetzel, 1997). To calculate D index, first of all, there is a rank order of students using their test scores. Secondly, 27% of students at the top, as well as 27 percent of students at the bottom are separated for analysis. According to Zubairi and Kassim (2006), "27 percent is used since evidence suggests that this value would maximize the differences in normal distributions, at the same time providing sufficient cases for analysis". D is given as:

$$D = \frac{Pu - Pi}{n} \qquad (4)$$

Pu and Pi represent accurate responses in the upper and lower group respectively and "n" is the group with the highest number of students. Because the index is on a scale from -1 to +1, the results of a negative index imply that the greater part in the lower group correctly answered the item, whereas a positive index implies that a greater part in the upper group correctly answered the item (Courville,2004).

d. Discrimination coefficients

Item's discrimination effectiveness consists of two indicators; point biserial correlation, in addition to the biserial correlation coefficient. A correlation selection is dependent on the question type to be answered. One major limitation of D is that just 54 percent (which is 27% upper + 27% lower) is used for computing the item discrimination, hence ignoring 46 percent of examinees. Likewise, the benefit in using the discrimination coefficients is that each examinee is used in the computation of the coefficients.

The definition of a point-biserial correlation coefficient (rpbi) is:

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq} \qquad (5)$$

Mp = whole-test mean of examinees who answered an item correctly,
Mq = whole-test mean for examines who answered an item incorrectly,
St = standard deviation for the whole test,
p = proportion of examines answering correctly
q= proportion of examinees answering incorrectly (Zubairi & Kassim, 2006).

Table 2. *Interpretation of Discrimination Indices*

Discrimination index	Quality of an item
$D \geq 0.40$	Item is functioning fairly satisfactory
$0.30 \leq D \leq 0.39$	Good item; therefore, less or no revision necessary
$0.20 \leq D \leq 0.29$	Item is marginal, hence should be revised
$D \leq 0.19$	Poor item: may have to be excluded or entirely changed

Source: Ebel (1979)

Item Response Theory (IRT)

Over the past years, IRT has increasingly become popular. IRT application is mainly found in psychological as well as educational testing, and more recently, it has become relevant for assessing health outcomes (Cai et al., 2016).

In the context of education, IRT emerged to resolve the shortcomings involved in the classic measurement theory of sample and test dependencies. IRT gives a theoretical framework which enables modeling of the relationship between the probability of answering an item correctly for an examinee of a given ability or trait.

$$P(X_i=c|\theta n) = f(\theta n) \tag{6}$$

- X_i represents the random variable symbolizing item *i* answer, with discrete varieties of response
- c represents the response which is observed
- If X is dichotomous, c=0,1 normally, 0 denotes inaccurate answers and 1 accurate answer.
- If X is polytomous, c=0,1,m (m>1).
- θn=nth person's trait parameter.

This represents the *item response function (IRF)*. IRF indicates a function relating the latent trait as well as the probability of responding accurately to an item and represented graphically by item characteristic curve.

Some popular unidimensional IRT models for determining dichotomously scored responses are described below.

The One Parameter Logistic Model (1PLM)

This type of IRF is called one-parameter logistic model as it includes just one (1) item parameter (ie. difficulty δ), with the functional form derived from the logistic function (Tendeiro, 2017). A special case of the 1PL is the Rasch Model. According to the 1PLM, an item *i* IRF is denoted by:

$$P\ (Xi = 1|\theta) = \frac{exp(\theta-\delta i)}{1+expexp\ (\theta-\delta i),} \qquad (7)$$

- θ represents the individual’s latent ability parameter.
- δi represents the item’s *difficulty* or popularity.

Figure 1 presents three (3) IRFs based on the 1PLM for three (3) items with different levels of difficulty: δ=−1(blue), δ=0 (red), δ=1.5 (green).

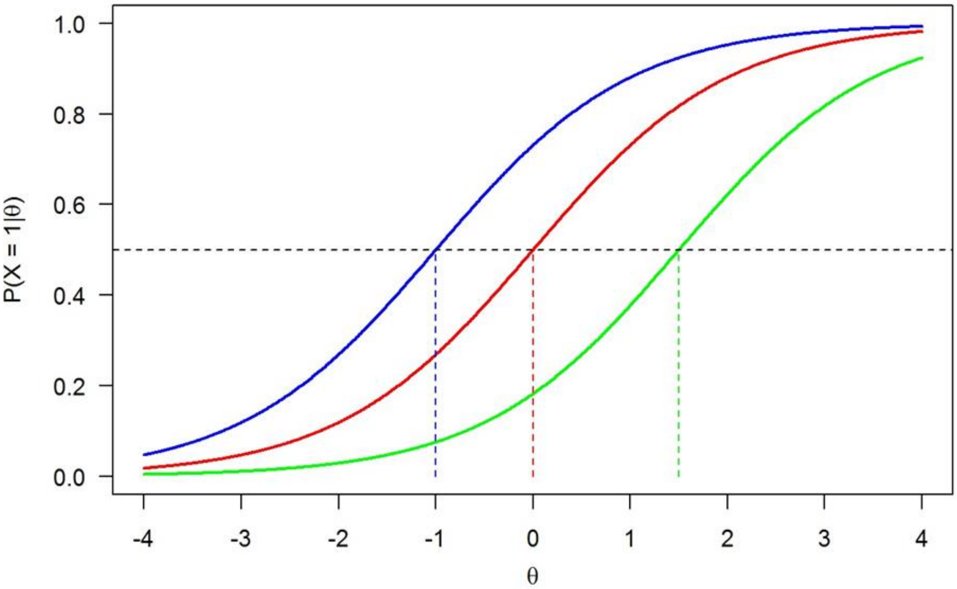


Figure 1. ICC for 1PLM

The 2PLM

The 2PLM extends the 1PLM by including another item parameter called the “*discrimination*” parameter. Based on the 2PLM, an item *i* IRF is:

$$P\ (Xi = 1|\theta) = \frac{exp\ [ai(\theta-\delta i)]}{1+(\theta-\delta i)],} \qquad (8)$$

αi represents discrimination parameter for item *I*, that’s denoted by the slopes of the curves, such that the steeper the slope, the more discriminatory the item is. This parameter carries positive values, even though those greater than, for instance, 4 or 5 remain uncommon (Tendeiro, 2017).

Figure 2 presents 3 IRFs for three items based on 2PLM. The parameters of discrimination: α=2 (blue), α=.5 (red), whereas α=1 (green).

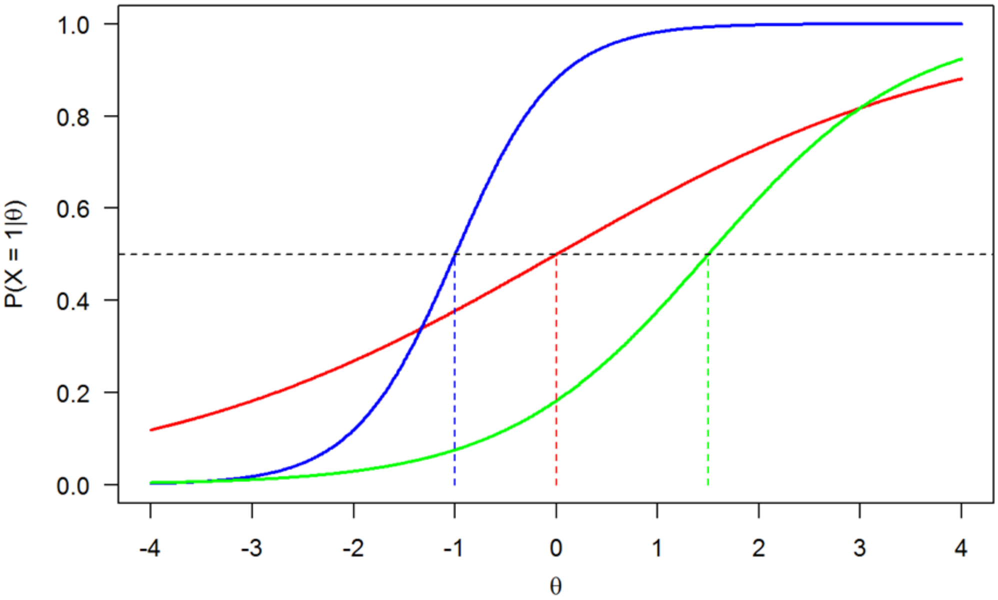


Figure 2. ICC for 2PLM

The 3PL model (3PLM)

The 3PLM extends the 2PLM by including another item parameter called the *(pseudo)guessing* parameter. Item *i* IRF is:

$$P\ (Xi = 1|\theta) = \frac{\gamma i+(1-\gamma i)exp\ [ai(\theta-\delta i)]}{1+(\theta-\delta i)]} \tag{12}$$

where γi represents the pseudo guessing parameter for item *i.* This carries 0 to 0.5 values. It assumes that even examinees with low ability possess the likelihood of correctly answering items merely using guessing randomly the right answer. For instance, γi is equivalent to .25 in a multiple-choice item having four-options, which suggests that an examinee who is only guessing an item’s right answer will select the right response with ¼ probability (Tendeiro,2017).

Figure 3 illustrates three (3) examples of IRFs based on the 3PLM. The pseudo guessing parameters include γ=0 (blue), γ=.25 (red), while γ=.20 (green).

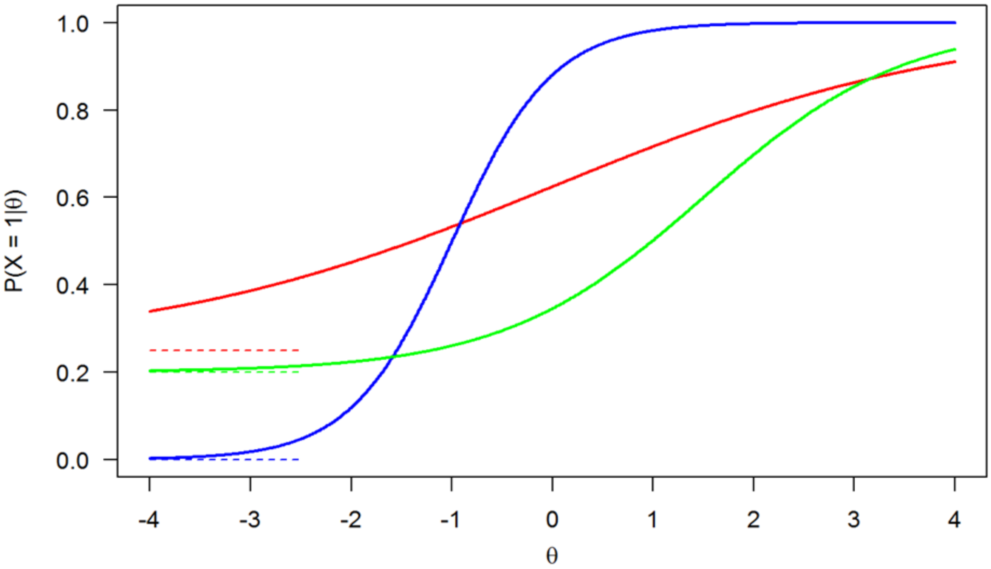


Figure 3. ICC for the 3PLM

Distractor Analysis

A traditional distractor analysis is usually performed with multiple-choice items. This type of analysis is based on CTT or IRT. Distractor analysis is aimed at eliminating non-functioning distractors and improving the discrimination power relating to multiple-choice items (MCIs) in differentiating between low as well as high ability examinees (Haladyna, 2016). Haladyna and Downing (1993) opined that if the number of examinees who select a distractor represents not more than 5 percent the distractor can be seen as low frequency.

According to Haladyna (2016), a trace plot is used in identifying non-discriminating as well as non-functioning distractors (see: Figure 4). For the horizontal axis, it implies the total score of examinees that is shared into 5 ordinal categories starting from the lowest group to the highest group. For the vertical axis, however, this signifies the overall percentage of examinees that selected the specific option. Thus, Option ‘A’ represents an accurate answer. Consequently, the overall percentage of examinees that choose Option ‘A’ rises when there is an increment in examinee’s ability. For Option B, this portrays well-functioning distractor. There is a decline in the overall percentage of examinees that choose Option B as examinee ability upsurges. In terms of Option C, this signifies a non-discriminating distractor, with a relatively unbroken rate of selection in various levels of ability. For Option D, this represents a non-functioning distractor, with overall percentage selection lesser than 5 percent in the entire ability groups. The trace line omitted indicates examinees that choose nothing in the item.

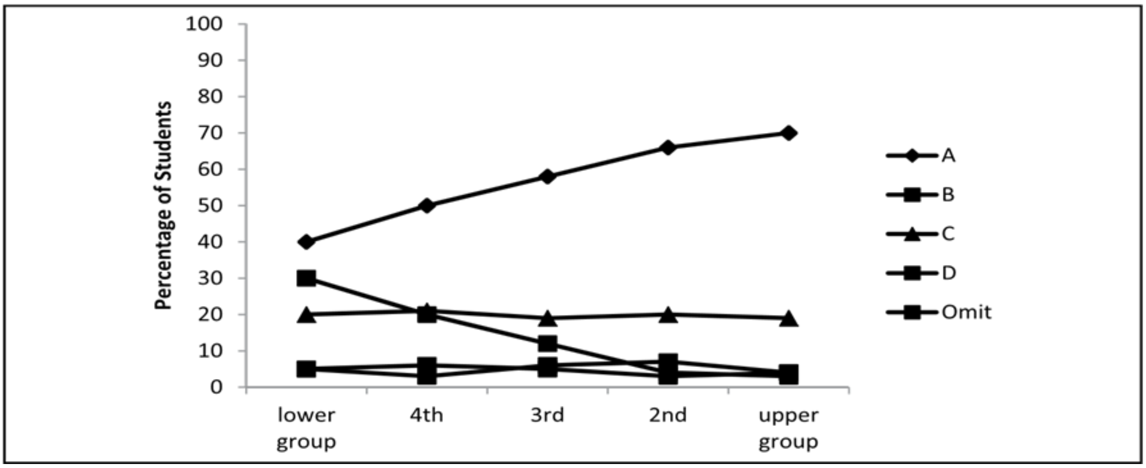


Figure 4. A trace plot for hypothetical Multiple-Choice Item (*Option A is the right answer)*

The choice mean of a distractor is another important tool which can be utilized to examine distractors (Haladyna & Rodriguez, 2013). In terms of an item that is well-discriminating, the choice mean in the accurate option is anticipated as higher than the choice mean of other distractors. When a distractor's choice mean remains higher than the accurate option's choice mean, there should be an assessment for content accuracy of that distractor. On the other hand, if there is a similarity in the choice means of the response options, the item then does not seem to be adequately discriminating.

The point-biserial correlation (correlation between continuous total test score and dichotomous item) or biserial correlation (correlation between continuous total test score and latent item score) is the most objective means of evaluating the choice mean of a distractor (Attali & Fraenkel, 2000). These scholars indicate that in calculating the point–biserial correlation to assess a distractor, there is the need for researchers to contrast examinees who select the distractor with examinees who select the right option, instead of the examinees who fail to select the distractor. The formula for the point– biserial is:

$$PBDC = \frac{MD - MDC}{SDC} \quad \frac{PD}{PC} \tag{6}$$

M_D represents distractor *D's* choice mean; M_{DC} represents total scores mean of examinees that select either a distractor or right option; S_{DC} represents the standard deviation relating to those examinee group that selects a distractor or right option; while P_D, as well as PC, indicate those examinees that chose a distractor as well as right option respectively.

Attali and Fraenkel (2000) indicate that an item having a PB_{DC} value to be higher than −0.05 is not adequately discriminating, whereas values less than −0.05 could be considered discriminately.

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

In comparing CTT distractor to IRT distractor analysis, the IRT distractor analysis does not just assess the proper functioning of a distractor, but it also enables the analyst to make use of distractors to estimate the abilities of students. There exist two (2) common IRT models in assessing distractor analysis. They include the nominal-response model (NRM) by Bock (1972), as well as the graded response model (GRM) by Samejima (1979).

Bock (1972) recommended the NRM for analysing distractors present in MCIs. Rather than the traditional IRT models which involve approximating the likelihood of answering items accurately, the NRM is associated with estimating the likelihood of selecting individual option devoid of assuming between the options. The NRM could be presented as:

$$P(Xj = k|\theta) = \frac{exp(ak(\theta - bk))}{\sum_{h=1}^{mj} \; exp(ah(\theta - bh))} \tag{7}$$

$P(x_j = \mathrm{K} \mid \theta)$ represents the probability that k in item j will be chosen assuming the examinee's ability is θ (usually between −4 and 4); a_k represents item discrimination for distractor k; b_k represents the difficulty of distractor k; whiles m_j represents the overall quantity of options in item j.

A major shortcoming of the NRM remains that, if a candidate's ability reduces, the likelihood of selecting one specific distractor increases. Nevertheless, this is not possible in reality as examinees having the low ability have the likelihood of guessing the right option randomly. In overcoming this shortcoming, Samejima (1979) suggested a variation of the NRM which considers the number of examinees who guess randomly the right option. The model presupposes the existence of a latent type of "*don't know*" (DK). Examinees belonging to the DK group would guess randomly, with the likelihood to guess seen by modelling the likelihood of choosing an individual option. The GRM is presented as:

$$P(Xj = k|\theta) = \frac{exp(ak(\theta - bk))}{\sum_{h=1}^{mj} \; exp(ah(\theta - bh))} + dk \frac{exp(ak(\theta - bo))}{\sum_{h=1}^{mj} \; exp(ah(\theta - bh))}, \tag{8}$$

where d_k is fixed to 1/ m_j to signify that examinee belonging to latent group *DK* will randomly guess.

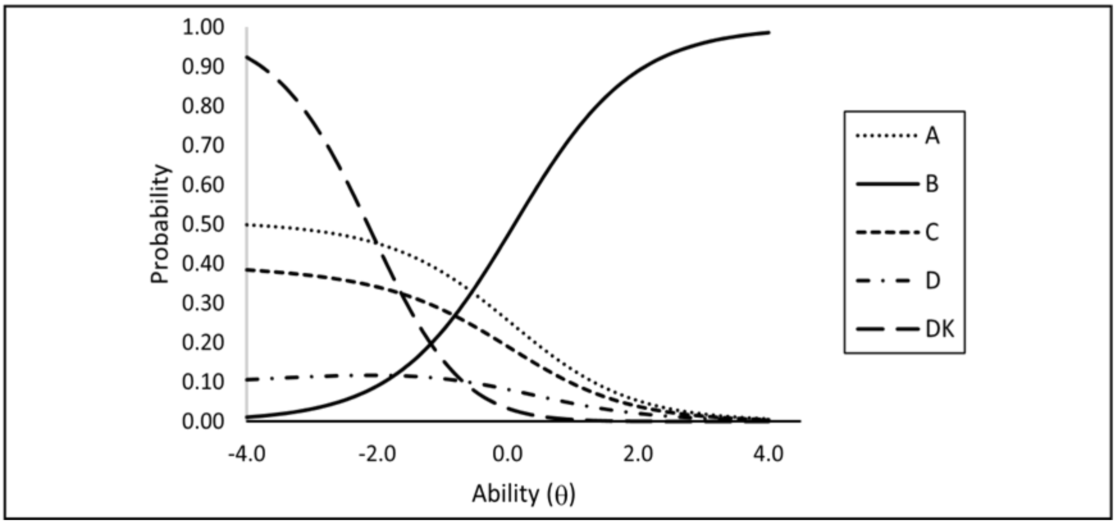


Figure 5. ICC for response options of MCQ *(Note:* The right response is Option B; DK: Don’t know).

$$\frac{exp\left(a_0\left(\theta-b_0\right)\right)}{\sum_{h=1}^{m_j} exp\left(a_h\left(\theta-b_h\right)\right)} \qquad (9)$$

This the probability that a student belongs to the latent group DK.

The two (2) IRT models provided are seen as utilizing the item characteristic curve (ICC) plot as presented in Figure 2. Option B represents an accurate response. Based on Figure 2 , as the student ability upsurges, the likelihood that Option B is selected upsurges. Option D represents the instance of a distractor attracting examines with low ability. As ability upsurges, the likelihood that Option D will be chosen reduces. Option C represents the instance of a distractor attracting examinees possessing limited knowledge. The likelihood that Option C will be chosen increases at θ=−1 but reduces for smaller or higher ability levels. Option A represents the instance of a distractor which is non-discriminating. The probability that Option A will be chosen remains constant across ability levels. For DK (latent group), as ability reduces, the likelihood of belonging to DK group approaches to 1.

Methodology

The current study adopted the quantitative descriptive research design to describe the psychometric properties of teacher-made mathematics test. The choice of this design was considered suitable because the study was concerned about the state of teachers‘ testing practices already existing in Senior High Schools without the manipulation of any variables.

The data for this study was mathematics test data of second year (grade 11) students from a popular Senior High School in Ghana. The test consisted of 40 MCQs with 4 options. The test covered topics from the first and second years of mathematics syllabus (grades 10 and 11). The test was administered in the paper-and-pencil form in the second semester of 2021 academic year for a duration of an hour. The total number of students who participated in the test were 600, however, a simple random sample of 400 students were selected. This sample size was arrived at because most psychometric analysis techniques require a minimum sample size of 200 for effective calibration (Martin & Larkin, 2018). The test booklets of the sampled students were digitized for the analysis. Prior to the collection and use of the student's data, appropriate ethical clearance was sought from the appropriate institutions.

To achieve the main objective of this study, both CTT and IRT analysis techniques were employed. For the CTT and dichotomous IRT analysis, the test data were dichotomously scored. However, for the distractor analysis, the multiple-choice data indicating the specific options selected for each of the items were used.

The distractor analysis was done using the nominal-response model (Bock, 1972). The response frequency in which a non-functioning distractor is defined as <5% and examination of option characteristics curves (trace lines) (Haladyna & Downing, 1993) were used. Specifically, the trace lines graphically display the response patterns of the item options but typically require a large sample of examinees above 200 (Osterlind, 1998) in evaluating distractor quality.

Results

The first part includes the frequency distribution of gender of students who were selected for this study. Table 3 presents the frequency distribution of the students. The sample of students for the mathematics test was approximately distributed with 172 representing 43 percent being male students and 228 representing 57 percent being female students.

Table 3. *Frequency distribution of gender of students*

Gender of student	Number	Percentage (%)
Male	172	43
Female	228	57
Total	400	100

Further, the mean test score and the standard deviation were 23.13 and 4.29 respectively. The median score was 23 which is almost the same as the mean test score and Cronbach's alpha was 0.61. Teacher made test needs to demonstrate reliability coefficient of approximately 0.50 or 0.60 (Quaigrain & Arhin, 2017).

Classical Test and Item Response Theories

To examine the psychometric properties of the test, the difficulty (p), the standard deviation and discrimination indices of each item (D) were estimated using the CTT first followed by IRT, and then the distractor analysis. Table 5 presents the item difficulty, the standard deviation of items and item discrimination of the test items.

Table 5. *Psychometric Result of Test Items*

Item Number	Item difficulty(P)	SD	Item Discrimination (D)
1	0.482	0.500	0.077
2	0.584	0.493	0.181
3	0.825	0.380	-0.011
4	0.186	0.389	0.076
5	0.866	0.341	0.210
6	0.526	0.499	0.401
7	0.890	0.312	0.104
8	0.942	0.233	0.135
9	0.784	0.412	0.186
10	0.247	0.431	0.270
11	0.501	0.500	0.207
12	0.934	0.248	0.176
13	0.622	0.485	0.079
14	0.504	0.500	0.066
15	0.200	0.400	0.145
16	0.499	0.500	0.225
17	0.937	0.243	0.187
18	0.792	0.406	0.318
19	0.340	0.474	-0.021
20	0.759	0.428	0.320
21	0.882	0.322	0.193
22	0.427	0.495	0.033
23	0.523	0.499	0.284
24	0.233	0.423	0.095
25	0.877	0.329	0.075
26	0.482	0.500	0.023
27	0.871	0.335	0.169
28	0.340	0.474	-0.020
29	0.438	0.496	0.143
30	0.307	0.461	0.047
31	0.340	0.474	0.220
32	0.529	0.499	0.124
33	0.530	0.499	0.166
34	0.354	0.478	0.085
35	0.397	0.489	-0.058
36	0.548	0498	0.199
37	0.542	0.498	0.088
38	0.800	0.400	0.339
39	0.682	0.466	0.232
40	0.605	0.489	0.222

Note. SD = Standard Deviation

The difficulty (p) indices ranges from 0.186 to 0.942. The p-value translates to a number when multiplied by 100, which is the percentage of students who got the item correct. The higher the p-value, the easier the items are, which means the higher the index of difficulty, the easier it is to respond correctly to the item. Based on Hotui (2006) guidelines, if $0.20 \leq p \leq 0.90$, then the item is considered as good and acceptable. The item is considered excellent if $0.40 \leq p \leq 0.60$. However, if $p<0.20$, then the item is too difficult, and if $p> 0.90$, the item is classified as too easy.

As can be seen from Table 5, majority of items 37 representing 92.5 percent were of acceptable difficulty level with p-value within the range of 20-90 percent, while 14 items representing 35 percent with p-value range of 40-60 percent were excellent among them. Three items were found to be too difficult (7.5 percent) (p value < 20 percent) and only one item was found to be too straightforward or easy (p > 90 percent).

The test item's discriminating indices (D) were classified based on Ebel's (1979) guidelines on classical test theory item review: If $D \geq 0.40$, the item is functioning satisfactorily, if $0.20 \leq D \leq 0.29$, then the item is marginal and needs revision and if $D \leq 0.19$, then the item should be eliminated or completely revised.

The item with highest D was item 6 (0.40) which means that the item is functioning satisfactorily and the item with the lowest D was item 35(-0.058) which means that the item should be eliminated or completely revised. Three items representing 7.5% of the items 18, 20, 38 were good and needed little or no revision. Ten items representing 25% of the test items, that is items 5, 10, 11, 16, 23, 26, 31, 35, 39 and 40 are marginal and need revision. 25 items representing 65% of the items should be completely revised or replaced as they have very low discriminating power as presented by the table above. Four items had negative discrimination indices. Items with negative discrimination index are useless and can also reduce the validity of the test (Quaigrain & Arhin, 2017).

Table 6 shows the mean difficulty level and mean discrimination of the test with their respective standard deviation. The mean difficulty of the test was 0.58 which shows the mathematics test is of acceptable difficulty level for examinees. Mean discrimination of the test was 0.15 which shows majority of the test items has low discriminating power.

Table 6. *Summary of test parameter*

Parameter	Mean	Standard deviation
Difficulty index (p)	0.58	0.23
Discrimination index (DI)	0.15	0.10

For the dichotomous IRT model, the 3PLM fitted the data well as compared to the 1-and 2-PLMs. The difficulty (threshold or "b" parameter),

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

and the discrimination (slope or “a”) estimated for all the items did not depart so much from the CTT results. However, the average guessing parameter was 0.15, meaning that an ill-prepared examinee can guess and score correctly 15% of the 40 test items.

Distractor results

Bock nominal model 2.0.0 was used to conduct the analysis. Distractor analysis was performed in order to identify both functioning and non-functioning options. The response frequency and the option characteristics curves were used to evaluate distractor performance. The distractors serve as an indicator of the functionality of each option. Distractors which are chosen by one or more examinees are called functioning distractors and those not chosen by anyone are called non-functioning distractors. A non-functioning distracter is an option with a response frequency of <5% and a functioning distracter has a response frequency of ≥ 5 percent (Haladyna & Downing, 1993).

Table 7 presents the frequency distribution for all the 40 test items, which included 160 options made up of 120 distractors and 40 correct answers. 37 options representing 30.8 percent of the 120 distractors had a choice frequency of ≥ 5 percent out of all 120 distractors or non-correct choices evaluated. On the other hand, of the 120 distractors, 83 options representing 69.2 percent had a choice frequency of < 5 percent which indicate that 83 distractors are non-functioning and are perhaps implausible and of little use as distractors in multiple choice items and need to be replaced in future use of such items (Quaigrain & Arhin, 2017).

Table 7. *Frequency Distribution of Test Options*

		Frequency			
Item	Options	A	B	C	D
1		0.146	-0.198	0.022	**0.030**
2		-1.061	**-1.458**	-0.811	3.330
3		**-0.230**	-0.111	-0.323	0.664
4		0.312	**0.317**	-0.330	-0.300
5		-0.396	**0.868**	-0.352	-0.120
6		-0.273	-0.480	-0.466	**-0.214**
7		**0.147**	-0.155	0.182	-0.214
8		-0.347	0.241	**0.391**	-0.044
9		-0.643	0.389	**0.563**	-0.309
10		**0.723**	-0.311	-0.110	-0.302
11		-0.462	**0.441**	0.035	-0.015
12		0.061	0.063	-1.262	**1.201**
13		-0.347	0.041	0.084	**0.221**
14		-0.467	**0.335**	-0.014	0.146
15		0.652	-0.600	-0.504	**0.453**
16		-0.581	**0.597**	-0.141	0.125
17		-0.167	**0.467**	0.011	-0.310

18	-0.336	-0.375	**0.871**	-0.160
19	**0.034**	-0.437	0.391	0.012
20	0.069	**0.776**	-0.515	-0.331
21	1.229	-6.105	2.077	**1.642**
22	0.166	0.336	-0.515	**0.013**
23	-0.006	-0.237	-0.417	**0.660**
24	-0.223	**0.330**	-0.028	-0.079
25	-0.349	**0.088**	-0.069	0.330
26	-1.312	0.066	0.101	**0.432**
27	0.148	-0.533	0.151	**0.234**
28	0.317	-0.073	-0.090	**-0.154**
29	0.107	0.397	0.077	**0.296**
30	0.445	-0.331	-0.219	**0.105**
31	-0.752	**0.005**	-0.476	-0.268
32	-0.259	-0.098	0.346	**0.010**
33	**0.064**	0.321	-0.642	0.024
34	**0.088**	-0.201	-0.073	0.186
35	-0.311	-0.354	0.810	**0.341**
36	0.212	-0.288	**0.458**	-0.382
37	**0.128**	0.070	-0.102	-0.095
38	-0.206	-0.293	-0.545	**1.044**
39	**0.604**	-0.030	0.034	-0.608
40	0.423	-0.584	**0.567**	-0.406

Option characteristics curves for sample items are provided in figures 8-11. Figure 8 presents the option characteristics curve for test item 5. As illustrated in Figure 8 , option B represents the right answer. The overall percentage of examinees that choose option B increases when there is an increase in examinee’s ability. Based on the figure the likelihood that option B is selected increases. Option A, C and D portrays well-functioning distractors. Option A, C and D represent the instance of a distractor attracting low ability students. As ability increases, the likelihood that option A, C and D will be chosen reduces. Option A, C and D represents the instance of a distractor attracting examinees possessing limited knowledge, but however reduces for smaller or higher ability levels. Option A, C and D represent the instance of a distractor which is a discriminating or a functioning distractor.

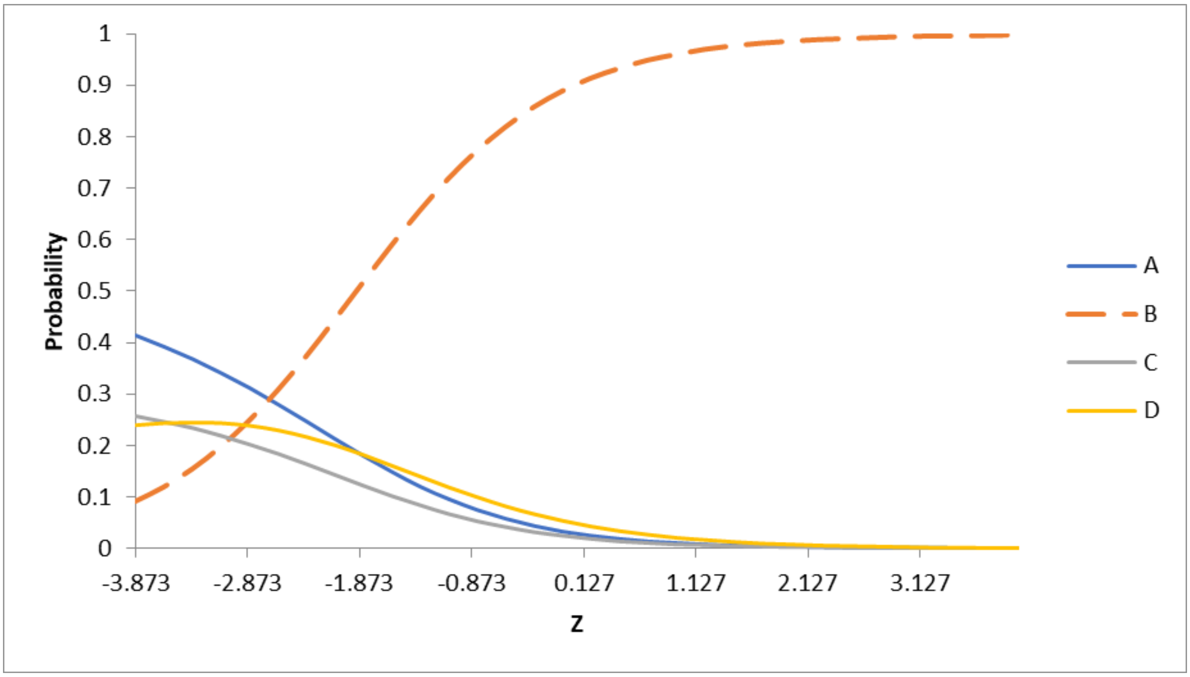


Figure 8. Item characteristics curve for every response option within multiple choice item 5 (*Note*. The right response is option B; Z. student's ability)

As illustrated in Figure 9 option, D represents the right answer for item 6. The overall percentage of examinees that choose option B increases when there is an increase in examinee's ability. Options A, B and C represent the instance of distractors attracting low ability students. As ability increases, the likelihood that option A, B and C will be chosen reduces. Option A, B and C represents the instance of a distractor attracting examinees possessing limited knowledge, but however reduces for smaller or higher ability levels. Option A, B and C represent the instance of a distractor which is a discriminating or a functioning distractor.

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

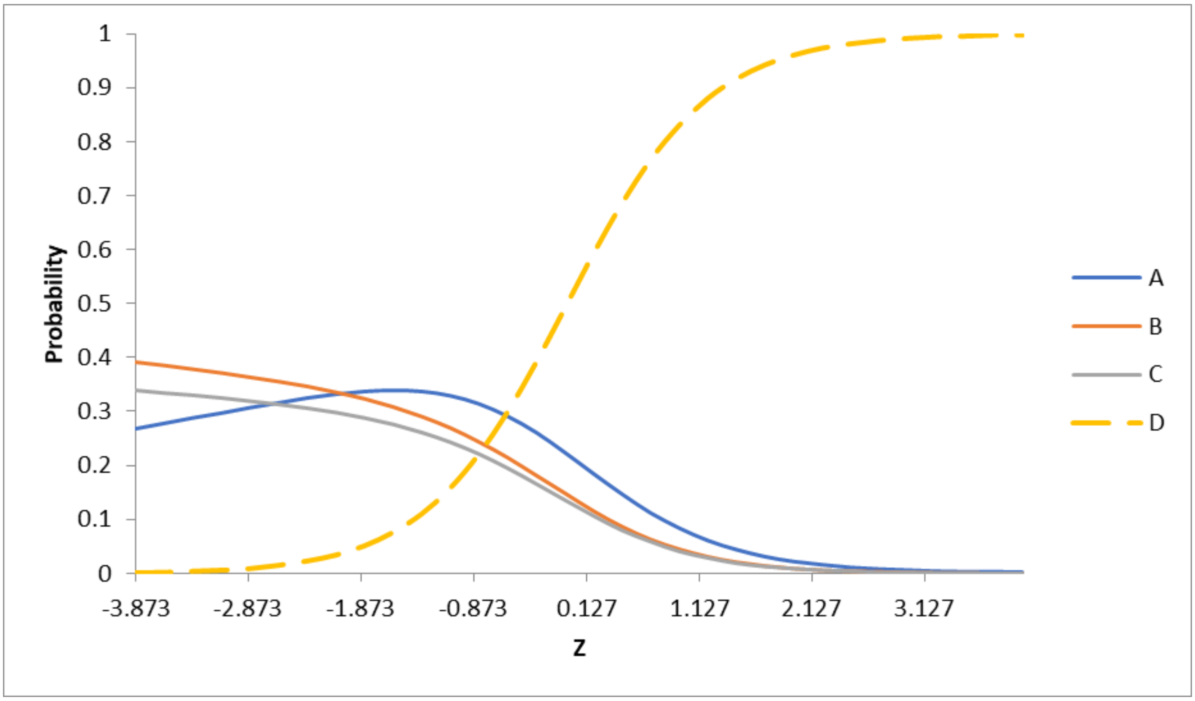


Figure 9. Item characteristics curve for every response option within multiple choice item 6 **(***Note*. The right response is option D; Z. student's ability)

As illustrated in figure 10, Option D is the correct option for item 28. but as ability increases the probability of selecting non-correct option C increases and the probability of selecting correct option D decreases. Option C and D signifies a non-discriminating distractor. For options B and C represents a non-functioning distractor. Distracters should be able to distinguish between low scoring student who have not grasped the subject content. Correct option D and non-correct option A, failed to discriminate between the informed student and the uninformed students thus need to be revised.

ISSN: 1857-7881 (Print) e - ISSN 1857-7431

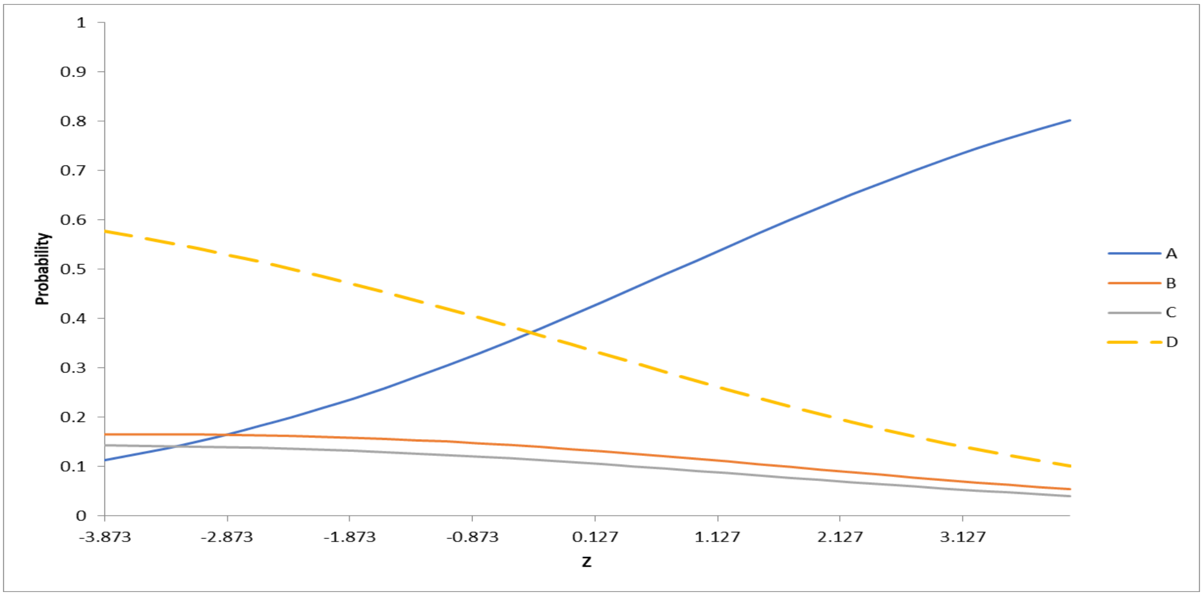
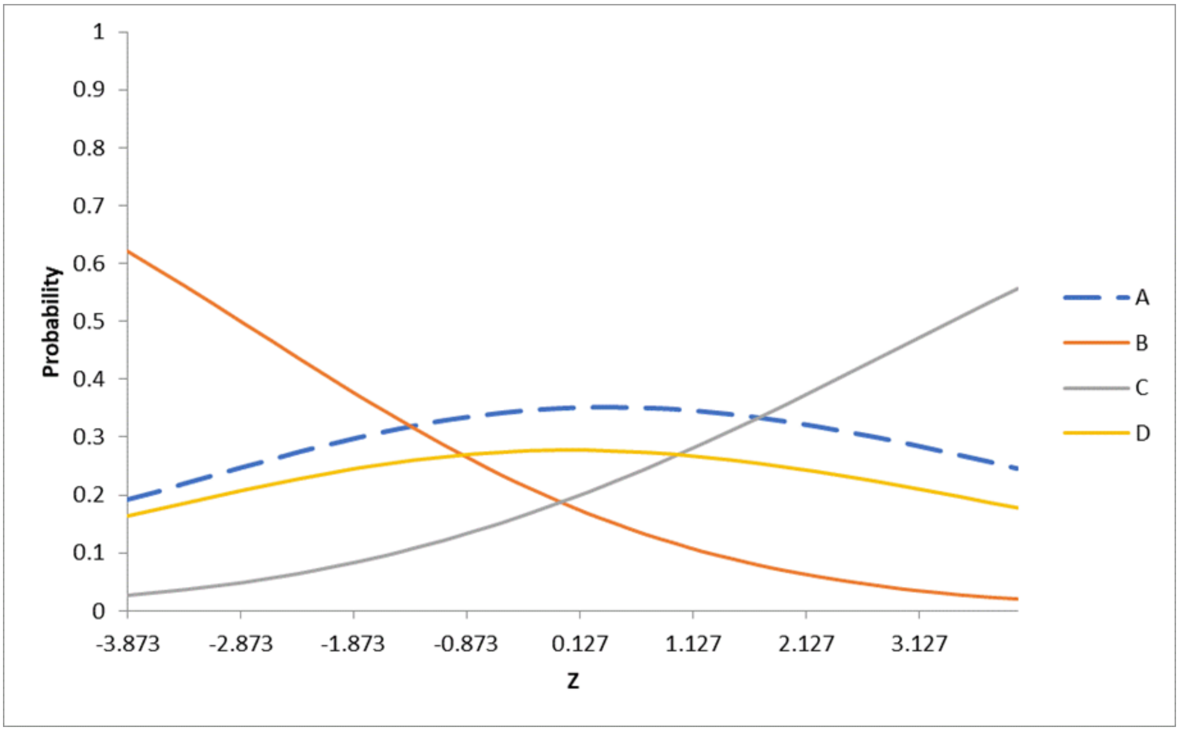


Figure 10. Item characteristics curve for every response option within multiple choice item 28 (*Note*. The right response is option D; Z. student's ability)

As illustrated in figure 11 Option A is the correct answer for item 19 and this signifies a non-discriminating option is the correct option. Option B portrays a well-functioning distractor. Option D and C represent non-functioning distractors. Distractors should be able to distinguish between low scoring student who have not grasped the subject content, and choose the distractors more often, whereas high scorers reject them more often while choosing the correct option.

Figure 11. Item characteristics curve for every response option within multiple choice item 19



Note. The right response is option A; Z. student’s ability

Discussion of Results

Classroom instruction must align with the test items to attain instructional validity and this includes setting valid questions. Mozaffer and Farhan (2012) stressed the significance of instructors' test construction steps and by means of statistical analysis to strengthen their instructional approaches and test building skills. Indictors that teachers will utilize to verify whether or not test items are well designed are the difficulty index, discrimination indices and distractor analysis. Most of the test items (26) are of satisfactory to excellent difficulty levels with only one test item found to be too difficult and three test items being too easy as indicated from the results of the analysis.

The findings of the discrimination indices suggest that 25 test items should be entirely replaced and four items showing negative discrimination indices. The negative discrimination indices could be as a result of entering an incorrect key or unclear nature of the test items (Quaigrain & Arhin 2017). Distractors were also analysed in this study for their functionality. Distractors are analyzed according to Quaigrain and Arhin (2017) to show how important each option is in each test item and if test takers repeatedly do not choose such options for multiple choices, these options may be implausible and hence of no use as distractors to students.

Constructing plausible and minimizing non-functioning distractors remains a significant feature of multiple-choice questions for framing consistency. The study indicates that 30.8percent of the options were functioning distractors whereas the majority of the options (69.2% of the options) were non-functioning distractors. The findings show that there are more non-functioning distractors as compared to functioning distractors. According to Tarrent et al. (2009), this small percentage of options with functioning options was not fully unexpected, reason being that when tests are created by teachers or examiners, some of whom have limited knowledge in item construction, a condition probably the same in most Senior High Schools in Ghana. This usually happens because studies indicate that there are rarely more than two functional distractors (Tarrant et al., 2009) in teacher-made test items and standardized exams. Research by Haladyna and Downing (1993) found that there were only one or two working distractors in items with four options they tested and items with five options had almost four non-functioning distractors. However, a test item with two probable distractors are suitable (Crehan et al., 1993) to a test question with more than two implausible distractors.

There is no psychometric explanation, on the other hand, that all test items must have the same number of choices as some test items will necessarily have more or less possible distractors than others (Tarrant et al., 2009). So, although three options would be appropriate in most circumstances, test developers must construct several better distractors given the subject field being evaluated (Haladyna & Downing, 1989). Nevertheless, many teacher-developed tests must comply with official procedures of the school as to how many test choices they must provide. These recommendations are rarely evidence-based (Haladyna & Downing,1989) and are probable to be founded on standard procedures and or proven procedures.

Conclusion

The findings of this study have demonstrated the relevance of evaluating the psychometric properties of test items after the administration of the test and using the results to improve or remove test items that have not functioned properly in order to enhance or improve the performance of the test items in future test administration. Item review and item analysis procedures should be used to measure the output of each question and their respective options. The method for item analysis includes the evaluation of test items relative to the distribution of responses (Tarrant et al., 2009).

According to Haladyna (2004), 50 percent or more of test items teachers and test developers write fail to perform as anticipated. So, it is imperative for item analysis to be performed in order to obtain useful data for question enhancement and can be integrated into the test creation and review process. In this regard, this study, that is rare within the Ghanaian context, examined the level of difficulty, discrimination indices, and distractor functioning of mathematics multiple-choice test designed by a teacher.

The conclusion drawn from this study is that teacher-made mathematics test had acceptable psychometrics properties. Specifically, the average difficulty and discrimination indices of the test were within acceptable ranges according to the criteria adopted for this study. However, the majority of the distractors were not functioning as they should. That is, 69.2 percent of the options were non-functioning distractors.

Limitation

This study investigated only one senior high school in Ghana, so these results cannot be generalized to cover all test items within the country. Further studies are needed on a larger scale in order to generalize the results.

Recommendation

Based on the findings of this study, the following recommendations were suggested:

- Teachers responsible for developing, validating and administering test in Senior High Schools need to perform psychometric analysis before (pilot testing) and after administering the test to ensure or improve the quality of test items.
- In-service training should be organized for teachers to enable them to acquire and develop the basic psychometric analytic skills.

Conflict of Interest

The author has no conflict of interest to declare.

References:

1. Adegoke, B. A. (2013). Comparison of item statistics of physics achievement test using classical test and item response theory frameworks. *Journal of Education and Practice*, 4(22), 87-96. https://www.iiste.org/Journals/index.php/JEP/article/view/8331
2. Amoako, I. (2018). Formative assessment practices among distanceeducation tutors in Ghana. *African Journal of Teacher Education*, *7*(3), 22-36. http://doi.org/10.21083/ajote.v7i3.4325
3. Asare, K. (2015). Exploring the kindergarten teachers' assessment practices in Ghana. *Developing Country Studies*, 5(8), 2225-0565.

4. Attali, Y., & Fraenkel, T. (2000). The point-biserial as a discrimination index for distractors in multiple-choice items: Deficiencies in usage and an alternative. *Journal of Educational Measurement,* 37(1), 77-86.
5. Awoniyi, F. C. (2016). The understanding of senior high school mathematics teachers of school-based assessment and its challenges in the Cape coast metropolis. *British Journal of Education,* 4(10), 22-38.
6. Beziat, T. L., & Coleman, B. K. (2015). Classroom assessment literacy: Evaluating pre-service teachers. *The Researcher*, *27*(1), 25-30. http://www.nrmera.org/wp-content/uploads/2016/02/Beziat.and_.Coleman.2015.Vol_.27.Issue_.1.pdf
7. Bichi, A. A. (2016). Classical test theory: An introduction to linear modelling approach to test and item analysis. *International Journal for Social Studies*, 2(9), 27-33. https://doi.org/10.26643/ijss.v2i9.6690
8. Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*(1), 81-90.
9. Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika,* 37, 29–51. https://doi.org/10.1007/BF02291411.
10. Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, 3, 297-321. https://doi/10.1146/annurev-statistics-041715-033702
11. Clarke, M. (2011). Framework for building an effective student assessment system: READ/SABER Working Paper. *World Bank.*
12. Courville, T. G. (2004). An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics. Unpublished Ph.D Dissertation, Texas A & M University.
13. Crehan K.D., Haladyna T. M., & Brewer B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. Educational and psychological measurement. *Sage Journals*, 53(1): 241-247. https://doi.org/10.1177/0013164493053001027
14. Davies, M., & Dempsey, I. (2011). Australian policies to support inclusive assessments. In *Handbook of accessible achievement tests for all students* (pp. 83-96). Springer, New York, NY.
15. Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The individualized classroom assessment scoring system (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interactions. *Early childhood research quarterly*, *25*(1), 1-16. https://doi.org/10.1016/j.ecresq.2009.08.004.

16. Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research, and Evaluation,* 14(1), 7. https://doi.org/10.7275/jg4h-rb87
17. Ebel, R. L. (1979). The Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.https://doi.org/10.1177/001316448104100337
18. Gareis, C. R., & Grant, L. W. (2015). Teacher-made assessments: How to connect curriculum, instruction, and student learning. Routledge.
19. Haladyna T. M., & Downing S. M. (1989). A taxanomy of multiple-choice item-writing rules. *Applied measurement in education journal*, 2(1):37-50. https://doi.org/10.1207/s15324818ame0201_3.
20. Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Routledge. https://doi.org/10.4324/9780203825945.
21. Haladyna, T. M. (2016). Item analysis for selected-response test items. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 392–409).
22. Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item?. *Educational and psychological measurement*, 53(4), 999-1010. https://doi.org/10.1177/0013164493053004013.
23. Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. Routledge. https://doi.org/10.4324/9780203850381
24. Hambleton, R. K., & Russell, W. J. (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.
25. Heritage, M. (2010). Formative assessment: Making it happen in the classroom. Corwin Press. http://dx.doi.org/10.4135/9781452219493.
26. Hockings, C. (2010). Inclusive learning and teaching in higher education: a synthesis of research. *York: Higher Education Academy*.
27. Hotiu, A. (2006). The relationship between item difficulty and discrimination indices in a physical science course (MSc thesis). Florida Atlantic university, BocoRaton, FL. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.599.5172 &rep=rep1&type=pdf
28. Kaplan, R. M., & Saccuzo, D. P. (1997). Psychological Testing: Principles, Applications and Issues. Pacific Grove, CA: Brooks/Cole Pub. Co.
29. Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The*

International Journal of Educational and Psychological Assessment,1(1), 1-11.

30. Martin, C. R., & Larkin, D. (2018). Minimum sample size requirements for a validation study of the Schizophrenia Quality of Life Scale-Revision 4 (SQLS-R4). *Journal of Basic and Clinical Health Sciences*, *2*(3), 76+. https://doi.org/10.30621/jbachs.2018.411
31. Matlock-Hetzel, S. (1997). Basic Concepts in Item and Test Analysis. Texas A & M University, USA. https://eric.ed.gov/?id=ED406441
32. Mensah, F. (2014). Evaluation of Social Studies Students' learning Using Formative Assessment in Selected Colleges of Education in Ghana. *British Journal of Education*, *2*(1), 39-48.
33. Mills, E. D., & Mereku, D. K. (2016). Students' performance on the Ghanaian junior high school mathematics national minimum standards in the Efutu Municipality. *African Journal of Educational Studies in Mathematics and Sciences*, *12*, 25-34.
34. Mozaffer, R. H., & Farhan, J. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *Journal of Pakistan Medical Association*, *62*,142–147. http://jpma.org.pk/PdfDownload/3255.pdf
35. Odili, J. N. (2010). Effect of language manipulation on differential item functioning of test items in Biology in a multicultural setting. *Journal of Educational assessment in Africa*, 4-268.
36. O'Malley, P. (2010, November). Students evaluation: Steps for creating teacher-made test. In Assessment Group Conference-School programme. Maryland: Kennedy Krieger Institute.
37. 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. https://doi.org/10.1080/2331186X.2017.1301013
38. Ravela, P., Arregui, P., Valverde, G., Wolfe, R., Ferrer, G., Rizo, F. M., Aylwin, M., & Wolff, L. (2009). The Educational Assessments that Latin America Needs. Washington, DC: PREAL.
39. Rodriguez, M., C. (2005). Three options are optimal for multiple choice items: a meta-analysis of 80 years of research. *Educational Measurement Issues and Practice*, 24(2),3-13. https://doi.org/10.1111/j.1745-3992.2005.00006.x
40. Salvia, J., Ysseldyke, J., & Witmer, S. (2012). Assessment: In special and inclusive education. Cengage Learning. https://doi.org/10.1177/0731948719826296
41. Samejima, F. (1979). A New Family of Models for the Multiple-Choice Item. Tennessee Univ Knoxville Dept. of Psychology. https://doi.org/10.21236/ada080350

42. Tarrant, M., Ware J., & Mohammed, A. M. (2009). An assessment of functioning and non- functioning distractors in multiple-choice questions: A descriptive analysis, *BMC Medical education*, 9(40), 2-20. https://doi.org/10.1186/1472-6920-9-40
43. Tendeiro, J. N. (2017). The lz (p)* person-fit statistic in an unfolding model context. *Applied psychological measurement*, 41(1), 44-59. https://doi.org/10.1177/0146621616669336.
44. World Bank (2013). System Assessment Benchmarking for Education Results: Ghana SABER Country Report. Available at http://wbgfiles.worldbank.org/documents/hdn/ed/saber/supporting_doc/CountryReport s/SAS/SABER_SA_Ghana_CR_Final_2013.pdf.
45. Zubairi, A. M., & Kassim, N. L. A. (2006). Classical and Rasch Analysis of Dichotomously Scored Reading Comprehension Test Items. *Malaysian Journal of ELT Research*, 2, 1-20. https://www.researchgate.net/publication/254504568_Classical_And_Rasch_Analyses_Of_Dichotomously_Scored_Reading_Comprehension_Test_Items