UNIAFRICA JOURNAL OF EDUCATION (APRIL, 2026), VOLUME 5, ISSUE 1.
ISSN: 2971-6004
https://www.uniafricajournalofeducation.com

POLYTOMOUS IRT-BASED DETERMINATION OF DISCRIMINATORY CHARACTERISTICS OF 2022–2023 NECO SSCE MATHEMATICS ESSAY QUESTIONS

ONOJAIFE JUDE ONOVUGHAKPOR
Department of Mathematics, School of Sciences, Delta State College of Education, Mosogar, Delta State.
onojaifejude4@gmail.com
09065362030

ISAAC EFE AKPOMEMIYERE
Department of Mathematics, School of Sciences, Delta State College of Education, Mosogar, Delta State.
isaac.efe@descoem.edu.ng; isaacefe2@gmail.com
08034131247, 09156105371

Abstract
This study examined the discriminating parameters of essay test items in the National Examinations Council (NECO) Senior School Certificate Examination (SSCE) Mathematics papers for the 2022 and 2023 June/July examinations using an Item Response Theory (IRT) polytomous model. The study also determined the number of Mathematics essay items that satisfied the required IRT statistical conditions. A descriptive survey research design was adopted. The population comprised all Senior Secondary School III (SSS III) students who registered for the NECO SSCE in Ethiope West Local Government Area, Delta State, Nigeria. A sample of 1,000 SSS III Mathematics students was selected from ten public and ten private secondary schools using simple and stratified random sampling techniques. The instrument for data collection consisted of the 2022 and 2023 NECO SSCE Mathematics essay past question papers. Item parameters were estimated using Xcalibre 4.1 IRT parameter estimation software developed by Assessment Systems Corporation (ASC). The generated parameters were analyzed using frequency counts, percentages, mean, and standard deviation to answer the research questions, while the formulated hypothesis was tested using an independent samples t-test at the 0.05 level of significance. The findings revealed that some of the Mathematics essay test items for both 2022 and 2023 did not satisfy the IRT statistical conditions. The results also indicated that there was no significant difference in the discriminating parameters of the NECO SSCE Mathematics essay test items between the two years when analyzed using the IRT polytomous model. Based on these findings, it was recommended that examination bodies responsible for the development and validation of achievement tests should ensure rigorous psychometric validation of test items to align them with the ability levels of examinees.

Keywords: Application, IRT Polytomous Model, NECO SSCE, Mathematics, Essay Test Items, Discriminating Parameters

Introduction

Assessment is an essential component of the teaching–learning process because it provides evidence on whether educational goals and instructional objectives are being achieved. Through assessment, educators and policymakers obtain information that guides important decisions regarding grading, promotion, curriculum implementation, instructional improvement, and allocation of educational resources. In essence, assessment enables educators to address key questions such as: Are teachers delivering the intended curriculum? Are students acquiring the expected knowledge and skills? How can teaching strategies be improved to enhance learning outcomes? Providing answers to these questions requires a systematic process of collecting, analyzing, and utilizing information about

students' learning and educational programmes in order to improve instructional effectiveness and student achievement. Consequently, the outcomes of assessment are expected to provide reliable evidence for sound educational decision-making. Decisions regarding what to evaluate, the strategies used for evaluation, and the format of reporting results depend largely on the curriculum design, the relative importance of its components, and the needs of stakeholders who utilize the information generated through the evaluation process (Akporuno, 2015). For educational decisions to be meaningful and credible, they must be based on valid and reliable measures that accurately reflect students' knowledge and skills.

In Nigeria, the National Examinations Council (NECO) is one of the major examination bodies responsible for conducting the Senior School Certificate Examination (SSCE)**.** The examination employs both multiple-choice (objective) and essay-type test items to evaluate candidates across different subject areas, including Mathematics. However, the performance of students in Mathematics in the SSCE has consistently been a source of concern for educators, parents, and policymakers. Similar concerns have also been reported by the West African Examinations Council (WAEC) regarding the persistent poor performance of secondary school students in Mathematics. In response to this challenge, several studies have attempted to identify factors responsible for students' poor achievement in Mathematics in external examinations. Most of these studies have focused on variables such as teacher effectiveness, learning environment, instructional methods, and students' attitudes and study habits (Hassan, 2018; Crosswell, 2021; Gross, 2023). Comparatively little attention has been paid to the psychometric quality of the test items themselves.

Reports by the National Examinations Council indicate fluctuations in the overall performance of candidates in the SSCE over recent years. For instance, the proportion of candidates who obtained at least five credit passes including English Language and Mathematics in the examinations conducted in 2019, 2020, 2021, 2022, and 2023 were 35.99%, 39.82%, 48.61%, 76.36%, and 79.81% respectively (Global Financial Digest, February 2022). Although the recent improvement appears encouraging, concerns remain regarding the reliability and validity of the assessment instruments used in evaluating students' performance.

The educational system involves multiple stakeholders including students, teachers, parents, government agencies, peer groups, and institutional infrastructures. In many instances, parents and students attribute high failure rates in external examinations to examination bodies, sometimes alleging that such bodies deliberately fail candidates. On the other hand, examination bodies often attribute poor results to students' lack of preparation and commitment to academic work. These conflicting perspectives have prompted psychometricians and educational researchers to examine the quality and characteristics of test items used in large-scale examinations.

Most previous psychometric investigations of external examinations have focused primarily on multiple-choice items using the Classical Test Theory (CTT)**.** Although CTT has been widely applied in educational measurement, it has certain limitations, particularly in its ability to provide detailed information about item characteristics and examinee ability levels, especially for essay or constructed-response items. To address these limitations, modern measurement models such as Item Response Theory (IRT) were developed. Unlike CTT, IRT provides more robust statistical procedures for estimating item parameters and can be applied to both dichotomous (right or wrong) and polytomous (multiple-score category) items. The polytomous IRT model is particularly useful for analyzing essay-type items where responses are graded across multiple score levels.

Given the importance of ensuring the psychometric quality of assessment instruments, it becomes necessary to examine the statistical properties of essay test items used in high-stakes examinations such as the NECO SSCE. Therefore, this study focuses on analyzing the discriminating parameters of past NECO SSCE Mathematics essay test items using the Item Response Theory polytomous model. By doing so, the study aims to determine whether these items meet the required IRT statistical conditions

and to provide empirical evidence that can guide improvements in the design and validation of essay-type test items in Mathematics examinations.

Statement of the Problem

A key issue in educational measurement is whether the test items used in examinations possess adequate discriminating power and meet established statistical standards. Test items with poor discrimination are unable to differentiate effectively between high-ability and low-ability students, thereby undermining the validity of the assessment results (Embretson & Reise, 2000). In many large-scale examinations, essay-type items are included because they assess higher-order cognitive skills such as reasoning, problem solving, and analytical thinking. However, these items are often more difficult to evaluate psychometrically compared to objective items, and as a result, they are rarely subjected to rigorous statistical validation.

Traditionally, analyses of examination items have been conducted using the Classical Test Theory (CTT) framework. Although CTT has been widely applied in educational measurement, it has several limitations, including its dependence on sample characteristics and its limited capacity to provide detailed item-level information. To overcome these limitations, modern measurement models such as Item Response Theory (IRT) have been developed. IRT provides more robust procedures for estimating item parameters such as difficulty and discrimination and is particularly useful in analyzing both dichotomous and polytomous test items (Baker & Kim, 2017). The polytomous IRT model is especially appropriate for analyzing essay-type questions, which are scored across multiple categories rather than simply right or wrong.

Despite the advantages of IRT, many studies examining examination items in Nigeria have concentrated mainly on multiple-choice questions, leaving essay-type items relatively underexplored. This gap is problematic because essay questions constitute an important component of Mathematics examinations in the NECO SSCE and are expected to measure students' deeper understanding of mathematical concepts. If these essay items do not possess adequate discriminating parameters or fail to meet IRT statistical conditions, the validity of the conclusions drawn from students' scores may be compromised.

Therefore, the problem of this study is the lack of empirical evidence regarding the psychometric quality particularly the discriminating parameters of NECO SSCE Mathematics essay test items when analyzed using the IRT polytomous model. Without such evidence, it remains uncertain whether these items effectively differentiate between students of varying ability levels. This study therefore seeks to address this gap by analyzing the discriminating parameters of the 2022 and 2023 NECO SSCE Mathematics essay test items using the IRT polytomous model in order to determine the extent to which they satisfy established IRT statistical conditions.

Objective of the Study

The objective of this study is to determine the discriminating parameters of the National Examination Council, Senior Secondary School Certificate Examination essay test items in Mathematics for the year 2022 and 2023. Specifically, the study is aimed at determining:

1. the discriminating parameters of the June/July NECO SSCE Mathematics essay test items for year 2022 and 2023
2. if the discriminating parameter satisfy the Item Response Theory statistical conditions for year 2022 and 2023
3. if there is difference between 2022 and 2023 discriminating parameters for the June/July NECO SSCE Mathematics Essay Test Items that satisfy the Item Response Theory statistical conditions.

Research Questions

The following research questions guided the study:

1. What are the discriminating parameters of the June/July NECO SSCE Mathematics essay test items for year 2022 and 2023?

2. Does the discriminating parameter satisfy the item response theory statistical conditions for year 2022 and 2023?
3. Are there differences between 2019 and 2020 discriminating parameters for the June/July NECO SSCE Mathematics Essay Test Items that satisfy the Item Response Theory statistical conditions?

Hypothesis

The following hypothesis was tested at 0.05 significant level:

1. There is no significant difference between 2022 and 2023 discriminating parameters for the June/July NECO SSCE Mathematics Essay Test Items that satisfy the Item Response Theory statistical conditions.

Literature Review

Theoretical Framework

Item Response Theory (IRT)

Item Response Theory (IRT) has recently emerged as a significant development in psychological and educational assessment. Often described as a "new psychometric term," IRT has gained considerable attention in psychological testing, test manuals, and academic journals. Familiarity with IRT is increasingly essential for psychologists, as those unaware of its methods may be disadvantaged in their research and practice (Hambleton, 2020). In countries like Spain, numerous publications on general testing practices now incorporate IRT principles.

Han and Hambleton (2017) highlighted that IRT provides a robust and widely applied framework for modeling educational and psychological test data. Its popularity stems from properties such as the invariance of item and examinee parameters, which are reported on a common scale. The theory is grounded in three key assumptions. First, unidimensionality assumes that all items measure a single latent trait, which can range from negative to positive infinity. In this context, the trait is treated as a unidimensional random variable, measurable on a standardized scale with a mean of 0.0 and a standard deviation of 1.0. Second, local independence stipulates that responses to items are statistically unrelated once the latent trait is accounted for. Third, the item response function (IRF) models the probability that a person with a specific ability level will answer an item correctly (Omorogiuwa, 2019).

Interest in IRT is expanding due to the increasing use of educational and psychological assessments and the demand for valid and reliable test items. Also known as latent trait theory, strong score theory, or modern mental test theory, IRT provides a framework for designing, analyzing, and scoring tests, questionnaires, and other instruments that measure abilities, attitudes, or related constructs. Previously, IRT was referred to as item characteristic curve theory (Umobong, 2019). Its foundation lies in the application of mathematical models to testing data, which offers advantages over Classical Test Theory (CTT).

High-stakes testing programs, including the Graduate Record Examination (GRE) and Graduate Management Admission Test (GMAT), often prefer IRT because it emphasizes individual item characteristics, unlike CTT, which focuses on overall test performance. The central aim of IRT is to overcome CTT limitations by establishing a scale on which examinee ability is independent of the specific test items administered (Hambleton, 2020). Historically, IRT research began in the 1940s and 1950s, becoming fully established approximately 30 years later (Lord, 1980; Hambleton, 2020). Pioneering work in the 1950s and 1960s was independently conducted by Frederic M. Lord, Danish mathematician Georg Rasch, and Austrian sociologist Paul Lazarsfeld. Although considered a modern psychometric theory, IRT builds upon concepts developed over more than seventy-five years, tracing back to L.L. Thurstone's 1925 work, *A Method of Scaling Psychological and Educational Tests*, which laid its conceptual foundation (Reeve, 2020).

In the 21st century, IRT finds diverse applications, including item analysis, item selection, test publishing, large-scale testing, test development, score equating, item bias studies, and computerized

adaptive testing. This study, in particular, focuses on the application of IRT in item analysis, examining its core assumptions, models, item parameters, and practical uses.

Core Assumptions of Item Response Theory

IRT is based on three fundamental assumptions: unidimensionality of the latent trait, local independence of items, and the item response function. While Han and Hambleton (2017) emphasize unidimensionality and local independence as the two primary assumptions, the item response function is critical for modeling examinee-item interactions.

1. **Unidimensionality**: Unidimensionality implies that each test item measures a single trait (θ). According to Umobong (2019), this assumption ensures that items assess only one domain of knowledge, for example, a Mathematics test should exclusively measure mathematical ability. Test scores are most meaningful when all items measure the same trait. In practice, many assessments are multidimensional, measuring multiple latent traits simultaneously, which violates this assumption.
2. **Local Independence**: Local independence indicates that, once the primary trait is accounted for, examinees' responses to different items are statistically independent (Han & Hambleton, 2017). In other words, the probability of answering one item correctly does not influence the probability of answering another correctly (Umobong, 2019). This assumption is closely linked to unidimensionality, although multidimensionality can emerge from other sources, and factor analysis is often used to investigate dimensionality.
3. **Item Response Function (IRF)**: The IRF, also known as the item characteristic curve (ICC), models the relationship between an examinee's latent ability and the probability of a correct response (Baker, 2021). Examinees with lower ability levels have a smaller probability of answering correctly, while those with higher ability levels are more likely to respond correctly. The IRF forms the basis for all IRT constructs and analyses, enabling precise evaluation of item properties and examinee performance. In the three-parameter logistic (3PL) model, the probability of a correct response to an item is mathematically expressed as:

$$P_i(\theta) = C_i + \frac{1-C_i}{1 + e^{-ai\,(\theta-bi)}}$$

Theta (θ) is the examinee ability and a_i, b_i and C_i are the item parameters which determine the shape of the IRF.

Empirical Studies on Psychometric Theory

Classical Test Theory (CTT) and Item Response Theory (IRT) are the two psychometric theories currently available for psychometricians for item selection and validation of measurement instruments. Several studies have been conducted by researchers, either to compare the two theories or to apply one of them for practical investigations based on their assumptions. Omorogiuwa (2019) conducted a study comparing CTT and IRT in the selection of physics achievement test items and concluded that the CTT item difficulty and discrimination indices are comparable to those of IRT. In other words, there is comparability in the items selected using the CTT and IRT procedures. Daniel (2016) compared and analyzed the examination questions and achievement levels of WASSCE and NECO SSCE general mathematics and concluded that the WASSCE and NECO SSCE question papers differ with respect to topic coverage and cognitive objective levels measured by the items for each Senior School Certificate Examination (SSCE). Osunde and Ethe (2003) in their study "Analysis of the Difficulty Index of SSCE Mathematics Multiple Choice Test Items between 1998-2002" reported that 64% of the items were accepted for 1998, and 60%, 60%, 74%, and 64% of the items were accepted for 1999, 2000, 2001, and 2002, respectively.

Baker and Kim (2017) confirmed that polytomous IRT models are particularly useful in analyzing performance-based and essay assessments because they capture multiple levels of achievement within

a single item. Their findings indicate that discrimination parameters estimated through models such as the GRM and GPCM provide deeper insights into item quality compared to classical test theory (CTT). In Mathematics essay tests, items with higher discrimination parameters are better able to differentiate students who demonstrate deeper conceptual understanding from those who exhibit limited procedural knowledge. De Ayala (2009) explained that the estimation of discrimination parameters in polytomous models helps test developers determine whether scoring categories function effectively. For Mathematics essay items, this analysis ensures that each score level (e.g., 0–4 marks) reflects meaningful differences in student ability and contributes to accurate measurement of mathematical competence.

Hassan (2018) investigated teacher effectiveness and its impact on secondary school students' performance in Mathematics in Nigeria. The study found that while teacher competence and instructional strategies significantly influenced student achievement, inadequacies in test items and assessment procedures also contributed to inconsistent student outcomes. This suggests that the quality of examination items can directly affect the validity of performance results. Crosswell (2021) conducted a study on factors affecting students' Mathematics achievement in West African secondary schools. The findings emphasized that beyond environmental and teacher-related factors, poorly constructed test items often fail to discriminate between high- and low-ability students, leading to unreliable assessment outcomes. The study recommended the adoption of psychometric evaluation techniques to enhance the quality of test items. Gross (2023) examined the influence of assessment instruments on students' Mathematics achievement. The study highlighted that multiple-choice items analyzed using Classical Test Theory (CTT) often overlook the complex nature of essay-type items, which assess higher-order cognitive skills. As a result, essay test items may not adequately differentiate students based on ability levels if they are not psychometrically validated.

In the context of psychometric models, Embretson and Reise (2000) argued that Item Response Theory (IRT) provides a more robust framework than CTT for evaluating test items, particularly those with multiple scoring categories (polytomous items). IRT allows for precise estimation of item parameters such as difficulty, discrimination, and guessing, thereby ensuring that the items reliably measure student ability. Baker and Kim (2017) further confirmed that the IRT polytomous model is suitable for essay-type questions, as it can capture the nuanced scoring of responses that range across multiple categories, unlike dichotomous scoring systems. In Nigeria, the performance of students in the NECO SSCE Mathematics examination has raised concerns over the past decade. The NECO reported that the percentage of candidates obtaining a minimum of five credit passes including English Language and Mathematics increased from 35.99% in 2019 to 79.81% in 2023 (Global Financial Digest, 2022). However, these aggregate statistics do not provide insights into the quality of test items or their capacity to discriminate among students of varying abilities. Several studies, therefore, underscore the need for empirical analysis of the test items themselves to ensure that assessments are valid and reliable (Akporuno, 2015; Hassan, 2018).

Methodology

This study adopted a descriptive survey research design, which is aimed at systematically collecting, analyzing, documenting, and describing the characteristics, features, and facts about a population under study (Omorogiuwa, 2016). The main purpose of using this design was to collect sample data on the NECO Senior School Certificate Examination (SSCE) Mathematics essay test items for the years 2022 and 2023, analyze the data, and describe the current status of item discriminating parameters as determined by the Item Response Theory (IRT) polytomous model. The population of the study comprised all Senior Secondary School III (SSS III) Mathematics students who registered to write the 2024 May/June NECO SSCE in Sapele Local Government Area of Delta State, Nigeria. A total of 1,000 SSS III students were selected for the study. These students were drawn from ten (10) public and ten (10) private secondary schools within the study area. The selection process combined simple random and stratified random sampling techniques to ensure proportional representation of different subgroups in the population.

The locality was divided into two zones: rural and urban schools. From each zone, five public and five private schools were selected using simple random sampling. The 1,000 students were then proportionally drawn from these schools, ensuring equal representation of public and private schools in both rural and urban areas. For public schools, 500 students (50 per school) were selected, and similarly, 500 students were selected from private schools. This approach ensured that each homogeneous subgroup in the population was represented proportionally in the sample, thereby satisfying the requirements for proportional stratified random sampling. The instrument for the study was the NECO SSCE Mathematics Essay Question Papers for the years 2022 and 2023. The essay test consisted of two sections: Part One: Five compulsory questions; Part Two: Seven optional questions (students were originally required to answer any five) For the purpose of this study, all twelve questions were treated as compulsory to facilitate uniform scoring and parameter estimation.

The study aimed to determine the acceptability of these test items according to IRT statistical conditions, and to estimate their item parameters, including discrimination (a). The instrument was not subjected to further validation, as it had already been validated and standardized by the National Examination Council (NECO). Reliability was established using the Cronbach alpha technique, yielding coefficients of 0.87 and 0.89 for the 2022 and 2023 examinations, respectively, indicating high internal consistency. The essay test papers were reproduced and administered to the sampled students with the permission of the school principals. Administration was conducted with the assistance of school teachers and the researcher. Completed scripts were collected and scored using the NECO scoring guide: Items 1–5: scored 1–8 marks each, based on steps completed; Items 6–12: scored 1–12 marks each

The students' scores were first subjected to item calibration using the Xcalibre 4.1 software, a Windows-based program designed for IRT parameter estimation (Guyer & Thompson, 2021). The software automatically selected items that satisfied IRT statistical assumptions and generated discrimination (a) parameters. Items were screened using the following criterion: Discrimination parameter (a) ≥ 0.30 – items meeting this threshold were considered acceptable (Lord, 1980; Omorogiuwa, 2019; Guyer & Thompson, 2021; Ethe, 2012). The number of acceptable and non-acceptable items for the years 2022 and 2023, as well as their parameters, were subjected to descriptive statistics, including mean, standard deviation, frequency, and percentage, to answer the research questions. For hypothesis testing, the discrimination parameters for 2022 and 2023 were compared using the independent samples t-test at the 0.05 level of significance to determine whether there was a statistically significant difference between the two years.

Results

Research Question One: What are the discriminating parameters of the June/July NECO SSCE Mathematics essay test items for year 2022 and 2023?

Table 1: Discriminating parameters of the June/July NECO SSCE Mathematics essay Test Items for Year 2022 and 2023

Year(s) Parameters/status Items	2022 a	Status	2023 a	status
01	-	**	-	**
02	-	**	2.928	*
03	1.727	*	1.465	*
04	1.533	*	3.980	**
05	2.498	*	-	**
06	3.518	*	-	**
07	2.951	*	3.212	*
08	2.872	*	0.625	*
09	-	**	-	**
10	3.204	*	-	**
11	2.468	**	3.229	*
12	2.420	*	1.150	*
Total no. Of items (16)	9		7	
Percentage of items	56.25%		43.75%	
Mean	2.577		2.370	
SD	0.648		1.272	
MIN	1.533		0.625	
Max	3.518		3.980	

* = Included Item(s), ** = Removed Item(s), a = Discriminating, SD = Standard deviation, Min = Minimum Value of the parameter, Max = Maximum Value of the parameter.

Results presented in Table 4.2 revealed the discriminating parameters of discriminating parameters of the June/July NECO SSCE mathematics essay Test Items for year 2022 and 2023 for all calibrated items from using the Xcalibre 4.1 IRT item parameter calibration report. From the Table, it was observed that 9 and 7 where the calibrated items for 2022 and 2023, representing 56.25% and 43.757% respectively. The minimum value is 1.533 and 3.518 is the maximum value for year 2022. For 2023 the minimum value is 0.625 and 3.980 is the maximum value respectively. Results also show mean values 2.577, 2.370, and standard deviations of 0.648, 1.272 for 2022 and 2023 accordingly.

Research Question Two: Does the June/July NECO SSCE mathematics essay Test Items discriminating parameter satisfy the item response theory statistical conditions for year 2022 and 2023?

*see results in Table 1

Results presented in Table 1 above revealed the number of the June/July NECO SSCE Mathematics essay test items that satisfied the Item Response Theory statistical conditions for year 2022 and 2023 discriminating parameter. The discriminating parameter ranges from 1.533 to 3.518 for 2022 and 0.625 to 3.980 for 2023. These discriminating parameters fall within the item response theory statistical conditions of 0.30 for an item to be acceptable.

Hypothesis One: There is no significant difference between 2022 and 2023 discriminating parameters for the June/July NECO SSCE Mathematics Essay Test Items that satisfy the Item Response Theory statistical conditions.

UNIAFRICA JOURNAL OF EDUCATION (APRIL, 2026), VOLUME 5, ISSUE 1.
ISSN: 2971-6004
https://www.uniafricajournalofeducation.com

Table 2: Independent Sampled t- Test of June/July NECO SSCE Mathematics Essay Test Items Discriminating Parameters for Year 2022 and 2023

	Years	N	Mean	Std. Deviation	$t_{critical}$	Df	P-Value
Discriminating parameters for year	2022	9	1.73	0.59			
	2023	7	1.09	0.18	5.354	14	0.000

Alpha (α) = 0.05 Level of Significance

Results in Table 2 revealed the Independent Sampled t- test of June/July NECO SSCE Mathematics essay test items discriminating parameters for year 2022 and 2023. The results showed a computed t-test value of 5.354, testing at 0.05 alpha (α) level of significance with a degree of freedom (df) of 14, P-value of 0.057. Since the p-value (0.000) is less than the alpha level of significance (0.05), the hypothesis which state that "there is no significant difference between year 2022 and 2023 June/July NECO SSCE Mathematics essay test items discriminating parameters using the Item Response Theory" was rejected at $p<0.05$ alpha (α) level of significance. Conclusion is therefore reached that, there is no significant difference between year 2022 and 2023 June/July NECO SSCE Mathematics Essay test items discriminating parameters using the Item

Discussion of Findings

Results presented in Table 1 show the discriminating parameters of all calibrated items as generated from the Xcalibre 4.1 Item Response Theory (IRT) parameter calibration software used in this study. The findings indicate that the range of discriminating parameters for the 2022 NECO Mathematics essay test items was 1.533 to 3.518, while for the 2023 items it ranged from 0.625 to 3.980. According to Guyer and Thompson (2021), items with discriminating parameter values of 0.30 or above are considered to discriminate effectively in the desired direction. Similarly, Lord (1980) emphasized that the higher the value of the discriminating parameter, the better the item is at differentiating examinees based on ability.
Response Theory.

From the results of this study, the lowest discriminating parameter observed was 0.625, suggesting that all calibrated items for both years were effectively discriminating. This conclusion was further supported by the mean and standard deviation values presented in Table 2. In particular, the fact that the mean values exceed the corresponding standard deviations indicates that the items achieved their intended measurement objectives (Osunde & Ethe, 2008). Akporuno (2015) further noted that the interpretation of the mean relative to the standard deviation depends on the nature of the data. Specifically, when the standard deviation is of greater magnitude than the mean, it may indicate a wider spread or variability in the data. For example, when measuring distances above and below sea level, a mean of zero (sea level) with a standard deviation of 20 feet would indicate that the measurements range within 20 feet above and 20 feet below sea level. By analogy, in this study, the relationship between the means and standard deviations reinforces the conclusion that the items are functioning appropriately and discriminating among examinees as expected.

Another important finding of the study showed that there was no significant difference between the discriminating parameters of the 2022 and 2023 NECO SSCE Mathematics essay test items when the IRT polytomous model was applied. This result indicates that the overall ability of the essay items to discriminate among students with different levels of mathematical proficiency remained relatively stable across the two examination years. The implication of this finding is that the quality of item construction in the NECO Mathematics essay examination may have been relatively consistent between 2022 and 2023. Stability in item discrimination parameters across examination years is often considered an indication of consistency in test development practices and examination standards.

This finding is in line with the principles of IRT which suggest that item parameters are expected to remain relatively invariant across different samples of examinees, provided that the items measure the same underlying construct. According to Samejima (1969), the invariance property of IRT allows item characteristics such as discrimination to remain stable across different administrations of a test. Similarly, Baker and Kim (2017) explained that when item parameters are estimated using appropriate IRT models, they tend to remain consistent across different examinee groups and test administrations. The absence of a statistically significant difference between the discrimination parameters of the 2022 and 2023 NECO Mathematics essay test items therefore suggests that the items in both examinations had relatively comparable capacity to distinguish between high- and low-ability students. This result may also indicate that the examination body responsible for the development of the Mathematics essay questions maintains similar levels of item quality and cognitive demand across examination years.

However, despite this overall similarity in discrimination parameters, the presence of items that failed to satisfy IRT statistical conditions highlights the need for continuous item analysis and validation during the test development process. Large-scale examinations such as those conducted by national examination bodies require rigorous psychometric evaluation to ensure that test items function appropriately and measure students' abilities accurately. Regular application of IRT models in the evaluation of examination items can therefore assist test developers in identifying weak items, improving scoring rubrics, and strengthening the overall quality of assessment instruments.

Conclusion

The findings of this study indicate that while the discrimination parameters of the NECO SSCE Mathematics essay test items remained relatively stable between 2022 and 2023, some items did not meet the required IRT statistical conditions. This underscores the importance of continuous psychometric evaluation and validation of examination items to ensure that they effectively measure students' mathematical abilities and maintain the integrity of large-scale educational assessments.

Recommendations

From the findings and conclusion of this study, the researcher therefore make the following recommendations that:

1. The National Examination Council (NECO) and similar examination bodies should ensure that all test items both psychological and achievement are rigorously validated to align with the abilities of examinees.
2. Evidence indicates that many Senior Secondary Certificate Mathematics essay items do not meet established statistical criteria under Item Response Theory (IRT). To enhance the quality and fairness of assessments, NECO should adopt modern psychometric approaches, including IRT, for item selection and validation.
3. Examination bodies should provide regular training for their staff on contemporary methods and technological tools for item analysis, validation, and selection.
4. All testing instruments must undergo thorough validation prior to their use in large-scale assessments to accurately measure the abilities of all examinees.

References

Akporuno, P. A. (2015). *Trend analysis of past NECO SSCE Mathematics essay test item difficulty and discriminating parameters using Item Response Theory Polytomous Model*. Ph.D. Thesis submitted to School of Postgraduate Studies, University of Benin, Nigeria.

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

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Crosswell, M. J. (2021). Research studies in public examination. Guilford, Associated Examining Board.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Gross, J. (2023). Exposing the cheat sheet with the students' aid. Retrieved from *http://www.Nytimes.com*

Guyer, R., & Thompson, N. A. (2021).*User's manual for xcalibre Item response theory calibration software, Version 4.1*. St. Paul MN: Assessment Systems Corporation.

Hambleton, R. K. (2020). *Emergence of item response modeling in instrument development and data analysis.* Medical care.

Han, K. T., & Hambleton, R. K. (2017). *User's manual for Wingen: Window software that generate IRT model parameters and Item Responses*. University of Massachusetts, Centre for Educational Assessment.

Hassan, T. (2018). *Understanding research in education*. Merrifeld publications united.

Lord, F. M (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, publishers. N.J.

Omorogiuwa, K. O. (2019). *An empirical comparison of the Classical test theory and Item Response Theory in the selection of Physics achievement test items. Ph.D*. thesis submitted to School of Postgraduate Studies, University of Benin, Benin City Nigeria.

Omorogiuwa, O. K (2016). *Research and applied statistics for Behavioural Sciences: An Introduction*. Mindex Publishing.

Osunde, A. U., & Ethe, N. (2003). Analysis of the difficulty index of SSCE Mathematics multiple choice test items (1998-2002). *Journal of Research in Counselling, 3*(3), 12-23.

Reeve, B. B. (2000). An introduction to modern measurement theory. Applied research program, Division of Cancer Control and Population Sciences national Cancer Institute.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*(4), 1–97.

Umobong, M. E. (2019). Item Response Theory: Introducing objectivity into educational measurement: In Afemikhe, O. A. and Adewale, J. G. (eds.) *Issues in Educational: Measurement and Evaluation in Nigeria, Nigeria*, Educational Research and Study Group, Institute of Education, University of Ibadan.