ATATÜRK ÜNİVERSİTESİ YAYINLARI
ATATURK UNIVERSITY PUBLICATIONS

Araştırma Makalesi *Research Article* DOI: 10.5281/zenodo.11073124

Item Characteristics of National Examination Council's Economics Multiple-Choice Items: An Item Response Theory Exploration

Ulusal Sınav Konseyi Ekonomi Çoktan Seçmeli Maddelerinin Madde Özellikleri: Bir Madde Tepki Kuramı İncelenmesi

Yusuf Olayinka SHOGBESAN[1]
Osun State University, Faculty of Education, Department of Arts and Social Sciences Education, Osogbo, Nigeria

ABSTRACT

The study explored the IRT parameter estimates of Economics multiple-choice items using the 1, 2 and 3 parameter logistic models. The study adopted the explorative research design with a sample size of 1500 senior secondary school III Economics students' selected using multi-stage sampling procedure. The Economics Achievement Tests (EAT) and students' responses as contained in the optical mark reader (OMR) serve as instruments. Data collected was coded and analysed using Mirt package in R statistical software for item parameter calibrations. The results showed that the discrimination index estimated using the 2 Parameter Logistic (PL) and 3PL models indicated that 28 items and 25 items respectively are poor items while 32 items and 35 items are considered good items respectively. Also, the difficulty index estimated using the 1, 2 and 3 PL models shows that 23 items, 25 items and 35 items respectively are easy items, 35 items, 33 items and 23 items are moderately difficult items while 2 items are considered difficult items. Furthermore, the results of the 3PL model shows that only 9 items are considered to be vulnerable to guessing with 51 items not vulnerable to guessing. The study concluded that the IRT psychometric estimates of NECO Economics multiple-choice items possessed moderately difficult items with an average discrimination indices and items majorly found not vulnerable to guessing. It therefore recommended that test experts and examination bodies should regularly consider the use of IRT psychometric estimations to evaluate item parameters for quality check of test items.

Keywords: IRT, Item Characteristics, Item parameter estimates, NECO, Economics items

ÖZ

Çalışma, 1, 2 ve 3 parametreli lojistik modelleri kullanarak ekonomi çoktan seçmeli maddelerinin Madde Tepki Kuramı (MTK) parametre kestirimlerini araştırmayı amaçlamaktadır. Çalışmada keşfedici araştırma tasarımı benimsenerek, çok aşamalı örnekleme yöntemi ile 1500 'lise III ekonomi' son sınıf öğrencisi örneklem olarak alınmıştır. Ekonomi Başarı Testleri (EAT) ve öğrencilerin optik okuyucuda (OMR) yer alan yanıtları veri toplama aracı olarak kullanılmıştır. Toplanan veriler, madde parametre kestirimleri için R istatistik yazılımındaki Mirt paketi kullanılarak kodlanmış ve analiz edilmiştir. Sonuçlar, 2 ve 3 Parametreli Lojistik modeller kullanılarak kestirilen ayırt edicilik indekslerine göre, sırasıyla 28 ve 25 maddenin zayıf maddeler olduğunu, 32 ve 35 maddenin ise yeterli maddeler olduğunu göstermiştir. Ayrıca 1, 2 ve 3 Parametreli Lojistik modeller kullanılarak kestirilen madde güçlük indekslerine göre, sırasıyla 23, 25 ve 35 maddenin kolay olduğu; 35, 33 ve 23 maddenin orta derecede zor olduğu; 2 maddenin ise zor olduğu belirlenmiştir. Bunun yanında, 3 Parametreli Lojistik modelin sonuçları, yalnızca 9 maddenin şansla doğru yanıtlanabilir olduğunu, 51 maddenin ise şansla doğru yanıtlanabilir olmadığını göstermiştir. Çalışmada, NECO Economics'in çoktan seçmeli maddelerine ilişkin MTK'ya ilişkin kestirimlerine göre, ortalama ayırt edicilik indekslerine sahip, orta derecede zor maddeler içerdiği ve maddelerin büyük ölçüde şansla doğru yanıtlamaya açık olmadığı sonucuna varılmıştır. Bu nedenle, test uzmanlarının ve sınav kurumlarının, test maddelerinin kalite kontrolü için madde parametrelerini değerlendirmek üzere düzenli olarak MTK kestirimlerini yapmaları önerilir.

Anahtar Kelimeler: MTK, Madde özellikleri, Madde parameter kestirimleri, NECO, Ekonomi maddeleri

Geliş Tarihi/Received 04.03.2024
Kabul Tarihi/Accepted 03.04.2024
Yayın Tarihi/Publication Date 29.04.2024

Sorumlu Yazar/Corresponding author:
Yusuf Olayinka SHOGBESAN
E-mail:
yusuf.shogbesan@uniosun.edu.ng

Cite this article: Shogbesan, Y. O. (2024). Item characteristics of national examination council's economics multiple-choice items: An item response theory exploration. *Journal of Psychometric Research, 2*(1), 8-18.

Introduction

Item Response Theory (IRT) describes the application of mathematical models to analyze response data collected during testing/survey situations whose main objective is to measure individual persons' latent trait, ability, or skill levels as the probability of a response of an item is modeled via a mathematical function of the student's trait parameters and the item parameters (Embretson and Reise, 2000). It is a measurement framework used in the design and analysis of educational and psychological assessments (achievement tests, rating scales, inventories, or other instruments) that measure mental traits (Ogunsakin and Shogbesan, 2018). Item response theory (IRT) was originally developed to overcome the problems with Classical Test Theory (CTT) which is looking at the reliability of the test as a whole while IRT looks at each item that makes up the test (Linden, 2018).

Item response theory (IRT) is one of the statistical frameworks that generate a mathematical function to describe the relationship between student performance in a test and ability or trait level. Its procedure improves psychometric methodology and assessment instruments (Baker and Kim, 2004; De Ayala, 2009; Oguguo and Lotobi, 2019). IRT attempts to estimate the test parameters, explain the process and predict the outcome of a given measurement for validity purposes (Nenty, 2015). As a result, in theory, it focuses specifically on the items that make up the test, compares the items that make up a test, and then evaluates the extent at which the test measures the student's ability (Raykov and Marcoulides, 2018). IRT is a better framework when compared to the CTT framework as it can be exploited by researchers in analyzing cognitive data for assessment and evaluation research, and non-cognitive data in the areas of attitude, personality, cognitive, and developmental assessment as well as sociological, psychological and psychopathological assessments (Gierl et al., 2001; Ogunsakin and Shogbesan, 2018). The IRT models produce item statistics independent of examinee samples and person statistics independent of the particular set of items administered. This invariance property of item and person statistics of IRT has been illustrated theoretically (Hambleton and Swaminathan, 1985; Hambleton et al., 1991) has been widely accepted within the measurement community.

Before using IRT models in psychometric process, three basic assumptions must be met. These are unidimensionality, local independency and item model fit (DeMars, 2010). The assumption of unidimensionality means that only one trait or ability is measured by the items while in the local independency assumption, responses for different items are not related and an item does not provide any clue to answer another item correctly. Also, when considering the assumption of local independence, it should be noted that if the test common factor is partial led out from any two items, their residual covariances is zero. The assumption of local independence may not be precisely met particularly when the test format includes several items that are related by a common problem (Adebowale, 2007). However, the assumption of local independency can also be met without meeting the assumption of unidimensionality as long as all aspects that affect the test results are taken into account (McBride, 2001).

Using the Item Response Theory, Item parameters include difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptote). Three most commonly used IRT models are; one parameter logistic model (1PLM or Rasch model), two parameter logistic model (2PLM) and three parameter logistics model (3PLM). Based on the one-parameter model, with each multiple-choice item in the test, in addition to the parameter, Birnbaum (1968) proposed extending more one parameter, the discrimination parameter, to show the ability to examinee's classification (Doan et al., 2016). All three models have an item difficulty parameter (b), In addition, the 2PL and 3PLmodels possess a discrimination parameter (a), which allows the items to discriminate differently among the examinees. The 3PL model contains a third parameter, referred to as the pseudo-chance parameter (c) (Shamshad and Siddiqui, 2020). The pseudo-chance parameter (c) corresponds to the lower asymptote of the item characteristic curve (ICC) which represents the probability that low ability test takers will answer the item correctly and provide an estimate of the pseudo-chance parameter (Embretson and Reise, 2000). However, the 4-parameter logistic model incorporates response time and slowness parameter (Wang and Hanson, 2001) has been formally incorporated into the traditional IRT models.

The four-parameter logistic (4PL) model as an extension of the usual three parameter logistic (3PL) model with an upper asymptote with a value usually less than one (1). Magis (2013) indicated that 4PL model allows more robust estimation of ability due to weighting the log-likelihood function (the aberrant item responses are down-weighted and have less impact on the estimation of ability). The four-parameter logistic model (4PLM) assumes that even high ability examinees can make mistakes (e.g. due to carelessness) as reflected by the non-zero upper asymptote (d-parameter) of the IRT logistic curve (Ogunsakin and Shogbesan, 2018). Moreover, software for 4-parameter logistic model analysis software such as the "PP" package created in the R programming language are now available for it analysis and parameter estimations.

Within the general IRT framework, many model shave been formulated and applied to real test data. However, to be able to choose the right model, the number of item response categories must be considered. For dichotomous items, the 1, 2, and 3 parameter logistic models are most common (1PL, 2PL, 3PL), and models including an upper asymptote parameter (e.g., 4PL) are also possible. However, for polytomous items, variations of the Partial Credit Model, Rating Scale Model, Generalized Partial Credit Model, and Graded Response Model are available for ordered responses, and the Nominal Model is appropriate for items with a non-specified response order. All these models are the cornerstone of IRT; they are the pivots upon which the theory depends and they reveal information about the latent behavior of the items and the examinee which make it easy for measurement community to make right predictions (Ogunsakin and Shogbesan, 2018).

Furthermore, with respect to item the (b) parameter which refers to the difficulty of an item, it describes where an item functions along the ability scale. The easy item functions among low ability examinee and a hard item functions among the high ability examinees, thus difficulty is a location index. However, the (a) parameter which denotes item discrimination helps describes how well an item can differentiate between examinees having abilities below the item location and those having abilities above the item location. It is reflected in the steepness of the item characteristic curve in its middle section. The steeper the curve, the better the item can discriminate and the flatter the curve, the less the item is able to discriminate because the probability of a correct response at low ability level is nearly the same as it is at high levels (Shogbesan, 2021). The IRT one, two or three parameters logistic models employs one or more parameters whose numerical value define a particular item characteristic curve and provide a vehicle for communicating information about an item's technical properties as well as to help determine the quality of the test items and the test as a whole. While the 4PLM has been used in order to reduce the influence of examinees' early mistakes on estimation of their ability level in a more effective way than 3PLM (Liao et al., 2012; Loken and Rulison, 2010; Rulison and Loken, 2009).

In order to determine the psychometric quality of test items of any examination, item analysis of examines students' responses to individual test items are carried out to assess the quality of individual items and of the whole test (Adedoyin and Mokobi, 2013). According to Adetutu and Lawal (2023), the absence of item analysis in developing these multiple choice items undermines the integrity of assessments, selection, certification, and placement in our educational institutions. Adedoyin and Mokobi (2013) in a study explored the psychometric analysis of 2010 Botswana mathematics Junior Certificate paper 1. The mathematics paper 1 consisted of forty (40) multiple choice test items which was constructed using the three year Junior Certificate mathematics curriculum. The population for the study was all the 36,940 students who sat for the Junior Certificate mathematics examination in 2010, out of which a sample of 10,000 was selected randomly by the use of SPSS computer software. The students' responses were analysed using IRT (3PL) model to examine the psychometric parameter estimates of the forty test items which were: item difficulty, item discrimination, and the guessing value. The result showed that Twenty three (23) items fitted the 3PLM out of the forty (40) items, and were used in examining the psychometric qualities of the JC mathematics test paper 1.The findings from this study indicated that out of the twenty three (23) items that fitted the IRT model, twelve (12) items were classified as poor test items, ten (10) items were classified as fairly good test items which could be revised or improved and one (1) item was considered to be good test item. Similarly, Văn Cảnh (2021) analyze and evaluate the 50 multiple-choice items among 590 students who took the English 1 test organized at Dong Thap University in 2018.based on Item Response Theory (IRT) with two-parameter and three-parameter models through analysis results of data from R software (package ltm). By evaluating each multiple-choice item based on their difficulty, discrimination parameters and guessing parameter according to the models, the study has identified good items to put into item bank, and point out items that are not really optimal, thus should continue to be considered before being put into use.

Furthermore, Ajeigbe and Oderinde (2021) compared stability of item difficulty, discrimination and guessing tendencies across four different paper types of English Language multiple-choice tests of Distance Learning Centre in Obafemi Awolowo University. The study adopted a causal comparative design because students' responses were obtained from the database. The 2449 students who sat for the first contact examination during the 2015/2016 session were used as sample size for the study. The instruments used for the study were four different paper types of English Language for 2015/2016 first contact, consisting of 60 Multiple-choice items each. The items were calibrated to generate item difficulty, discrimination and guessing tendency using X-Calibre 4.2 software package. One-way analysis of variance was used to estimate statistical difference in terms of item difficulty, discrimination and guessing tendency across the four different paper types. Results obtained showed that each test paper type is unidimensional in nature. Also, out of the 60 items 25(41.7), 39(65.0) and 38(63.3); 32(53.3), 36(60.0) and 42(70.0); 30(50.0), 37(61.7) and 38(63.3); and 27(45.0), 36(60.0) and 40(66.7) fell under moderate difficulty, discrimination and acceptable guessing value of .00–.25 across the four different paper types respectively. It also concluded that the item calibrations, in terms of difficulty, discrimination and guessing parameter, are stable and comparable across paper types.

Moreover, Oguguo and Lotobi (2019) also determined the psychometric properties of the examination items in 2011 Basic Education Certificate Examination for Basic Science. The design adopted was survey research design. The instrument for data collection was the 2011 Delta State Basic Education Certificate Examination (BECE) in Basic Science Multiple Choice Test Items. The IRT model for item selection was used to determine the estimates of the item parameters. The findings of the study revealed that 45, 45 and 40 items satisfied the IRT difficulty, discrimination and guessing parameter respectively. While 38 items satisfied the combined three IRT parameter estimates. Also, Adetutu and Lawal (2023) in another recent study analzed a university semester examination where a total of 403 students took a compulsory general statistics course made up of 35 multiple choice items. Using the 1, 2 and 3 parameter logistic models for estimation, they found out that Items 15, 5, 3, 13, 28, 34, 23, and 11 were identified to be defectives in terms of item difficultly, while Item 29 and 34 were identified as the most discriminating among others with item 6, 7, 9 among others were found to discriminate poorly and needs to be remediated. Finally, items as 5, 23, and 3 are considered “poor” which were suggested to be defectives and must be revisited, moderated due to their high pseudo-guessing indices.

From all the above stated psychometric investigations carried out by various researchers and given the various findings obtained by researchers in a view to explore the psychometric properties of test items of various subjects, it should be known that the items making up a test may be defectives especially when the item properties such as difficulty, discrimination, and pseudo guessing indices (power) of each item lacks quality, and thereby unable to appropriately measure students’ ability or traits as intended. When this defects are observed, item analysis and moderation can be used for remediation (Adetutu and Lawal, 2023). Given that the National Examination Council (NECO) is a body charged with the conduct of Senior Secondary School Certificate Examinations (SSCE) in Nigeria, hence, this study intends to investigate the IRT psychometric estimates of NECO SSCE 2015 Economics Multiple-Choice items. Specifically, the objectives of the study are to;

1. estimate the difficulty index of NECO SSCE 2015 Economics multiple-choice items using the 1, 2 and 3 parameter logistic models;
2. estimate the discrimination index of NECO SSCE 2015 Economics multiple-choice items using the 1, 2 and 3 parameter logistic models; and
3. estimate the guessing index of NECO SSCE 2015 Economics multiple-choice items using the 3 parameter logistic model.

Methods

Research Model

The study adopts the explorative research design with the aim to determine the item parameters estimates of NECO Economics items among the tests-takers in Ogun State.

Universe and Sample

The population for the study comprised all secondary school students in Ogun State. A sample of 1500 senior secondary school III Economics students was used for the study using the multi-stage sampling procedure. From each of the three senatorial districts in the state, three Local Government Areas (LGAs) was selected using simple random sampling technique. From each of the selected LGAs, four secondary schools (2 public and 2 private) was selected using stratified random sampling technique with school type used as stratum. Furthermore, from each of the 36 secondary schools that was selected for the study, SSIII Economics students was selected through proportional sampling technique. However, as a result of the proportional selection, a total of 1500 SSIII Economics students was finally selected and used as the sample for the study.

Data Collection Instruments

With respect to instrumentation, two research instruments were used for data collection: Economics Achievement Tests (EAT) and an answer sheet. The EAT contained the demographic characteristics of the test takers such as candidate name, gender and school type in Section A and a set of 60 multiple-choice items with five options adopted from NECO SSCE 2015 Economics paper III while the answer sheet also contained two parts; the first part provided for a section where the candidate name (optional), gender and school type can be indicated while the second part contained a response option labelled a-e that was used to indicate students response to each items as test-takers will have to indicate the correct option by ticking appropriately. The Kuder-Richardson 20 approach was adopted in determining the reliability of the instrument and it provides a reliability coefficient of .921. Thus in line with the recommendation of Wiberg (2004) the instrument is valid enough for measuring what each section was designed to measure.

Furthermore, during data collection, in order to ensure proper preparation by the student, adequate notification was given to the students through their various Economics subject teachers on the specific date for the test administration. During the tests, the students was arranged properly to allow for independent attempt of test items and the test-takers was given the allotted testing time as indicated on the question paper to be able to respond to all items appropriately. Also, the test administration was conducted under strict examination condition to prevent or avoid any other forms of cheating during the test through proper invigilation and supervision by the researcher/research assistance(s) and the Economics teacher(s) during the tests administration to prevent any confounding factor from affecting the study outcome. At the end of the testing time, the EAT question and students' response as provided in the answer sheet was collected immediately.

Data Analysis

Data collected was analyzed based on the research questions raised in the study. Data collected was subjected to initial analysis to assess the unidimensionality and model-fit assumptions while the parameter estimation analysis was done using Multidimensional Item Response Theory (MIRT) package in R software (version 3.6.2) for item calibration to estimate the difficulty, discrimination and guessing parameters of the 1, 2 and 3 parameter logistic models.

Results

Preliminary Analysis: Assessing the Dimension of the EAT items and Model fit

To ascertain the dimension of the tests, the responses of the examinees to the EAT items were subjected to Stoust's Test of Essential Unidimensionality (STEU implemented in DIMTEST 2.0 package) (Stout, 2005) for undimensionality assessment. This is done by separating the test in to two subtests, the Assessment Subtest (AT) and the Partitioning test (PT). The AT are the items chosen as those that measure best along a dominant trait. They are chosen so that they measure best in the direction most opposite to that of the PT items. The Assessment Subtest (AT), was selected empirically, using the HCA/CCPROX cluster procedure and DETECT statistic in DIMTEST, and this item cluster was tested to see if it was dimensionally distinct from the remainder of the test. A random sample of 30% of the examinees responses was used to select the Assessment Subtest, and the remaining 70% of the examinees responses (PT) was used for the dimensionality test. Table 1 presents the result of Stout's test of essential unidimensionality used in testing the assumption of unidimensionality of the EAT items.

Table 1

Unidimensionality of EAT Items

TL	TGbar	T	p-value
12.8803	11.0264	1.8447	.0325

Table 1 that the AT was not dimensionally distinct from the remaining items of the test (T = 1.8447, p value = .0325 < .05); therefore, the assumption of unidimensionality was ascertained. This result shows that one dimension accounted for the variation observed in examinees responses to the test items. Hence, the EAT items are unidimensional. As such, the item parameters can be appropriately estimated using the unidimensional IRT models.

Furthermore, the IRT model-data fit investigation was conducted as the test data was calibrated using: the one-parameter logistic model, two-parameter logistic model and three-parameter logistic model respectively. The -2loglikelihood values obtained for each of the models were compared. The result of the model-data fit indicated that all the items fitted well for all the one-parameter logistic model, two-parameter logistic model and three-parameter logistic model. As such, the one-parameter, two-parameter and three-parameter logistic unidimensional IRT models are fit and used appropriately for this study.

Research Questions

What are the estimates of the item parameter estimates (difficulty, discrimination and guessing parameter indices) of NECO SSCE 2015 Economics multiple-choice items using the 1, 2 and 3 parameter logistic models?

In order to answer the research questions, the responses of the examinees to the EAT items were subjected to IRT test item calibration using the 1, 2 and 3 parameter logistic models for estimation. Specifically on the interpretation of item discrimination in line with Ebel and Frisbie (1991) and interpretation of difficulty indices in line with Henning (1987), Hambleton and Swaminathan (1985), De Ayala (2009), the following interpretations are providedforprovided for item discrimination and difficulty parameters respectively as presented in table 2 and 3 below.

Table 2

Interpretations of Items Discrimination Indices

Discrimination Indices (a)	Interpretations
$C\geq1.70$	Item is functioning quite satisfactorily
$1.35\leq C\leq1.69$	Good item.
$0.65\leq C\leq1.34$	Moderate, little or no revision is needed
$0.35\leq C\leq0.64$	Item is marginal and needed moderation
$C\leq0.34$	Poor item, should be eliminated or moderated

Table 3

Interpretations of Difficulty Values

Difficulty Value (b)	Interpretations
$-3 < b$	Poor (too easy)
$-3.00 \leq b \leq -2.00$	Very easy
$-2.00< b <-1.00$	Easy
$-1.00< b <1.00$	Moderately difficult
$1.00 < b < 2.00$	Difficult
$b > 2.00$	Very difficult

Moreover, after the response obtained from the EAT were subjected to IRT test item calibration using the 1, 2 and 3 parameter logistic models for estimation , the results of the item parameter estimates using the 1, 2 and 3 logistic models are presented in Table 4 below.

Table 4 shows the results of the item parameter estimates (difficulty, discrimination and guessing) of Economics Achievement Test (EAT) using the 1, 2 and 3 parameter logistic models. Form the table, the result of the discrimination index estimated using the 2PL and 3PL models shows that 28 items and 25 items respectively are poor items while 32 items and 35 items are considered good items respectively. However, it was observed that 3 items; Items 31, 43 and 47 considered to be poor items under the 2PL model are calibrated to have parameter estimates that are considered good under the 3PL model. The result of the difficulty index estimated using the 1PL, 2PL and 3PL models shows that 23 items, 25 items and 35 items respectively are easy items, 35 items, 33 items and 23 items are moderately difficult items while 2 items; Item 9 and 42 are considered difficult items using the 1PL, 2PL and 3PL models respectively. Finally, the result of the 3PL model shows that 9 items; Items 20, 23, 24, 26, 31, 34, 43, 44 and 47 are considered to be vulnerable to guessing with 51 items not vulnerable to guessing. Furthermore, it was observed that 3 items; Items 31, 43 and 47 considered to be poor items under the 2PL model are calibrated to have parameter estimates that are considered good under the 3PL model. Also, the 3PL model estimate have more relatively easy items (35 items) compared to the 23 and 25 easy items calibrated under the 1PL and 2PL models. While the 3PL model estimate have less relatively moderate items (23 items) compared to the 35 and 33 moderate items calibrated under the 1PL and 2PL models.

Discussion

The results of research question one of the current study which estimated the discrimination , difficulty and guessing parameters shows that discrimination index estimated using the 2PL and 3PL models indicated that 28 items and 25 items respectively are poor items while 32 items and 35 items are considered good items respectively. This implies that the discrimination index estimated using the 2PL and 3PL models reveals that majority of the items can really distinguish better between examines with high and low ability despite having a 5 option response format and as such are having moderate difficulty indices. However, since all the items are standardized items adopted from items developed by NECO, the finding was consistent with the findings of Olatunji (2007), Olutola (2015) and Thomas et. al. (2018) that item constructed by WAEC as a similar examination body have more discriminating items. This, however is contrary to the assertion of Olatunji (2007) which reported that test item with fewer options had the best discriminating index.

The result further indicated that the difficulty index estimated using the 1PL, 2PL and 3PL models shows that 23 items, 25 items and 35 items respectively are easy items, 35 items, 33 items and 23 items are moderately difficult items while 2 items are considered difficult items using the 1PL, 2PL and 3PL models respectively. The result of the 3PL model also indicated that only 9 items are considered to be vulnerable to guessing. The implication of this finding is that the range of complexity of task measured by each item is such that it tend to accommodate every student as tests have to include easier items as well as more difficult items. The easier items will allow students to show more of what they have learned and avoid the occurrence of 'floor effect' as well as prevent the most able students from providing evidence of their advanced achievements causing a 'ceiling' effect (Izard, 2005 as cited in Shogbesan, 2017). It should be noted that the spread in the range of complexity of tasks measured by the test items as suggested by Izard (2005) was at least as wide as the expected range of achievement for the students being assessed. Also, the majority of the items are not vulnerable to guessing among the examinee as this may be attributed to the fact that at least 50% of the items have been pre-known and the examinee may have interacted among peers to identify the correct answers to the compromised items and harvest them during testing, thereby reducing the possibility of any further guessing.

Conclusion and Recommendations

It can be concluded from the study that the IRT psychometric estimates of NECO Economics multiple-choice items are such that they possess moderately difficult items with an average discrimination indices and items majorly found not vulnerable to guessing.

Based on the findings of the study, the following recommendations are made:

1. Test experts and examination bodies should regularly consider the use of IRT psychometric estimation to evaluate item parameters as a statistical measure for ensuring stability and quality psychometric properties of test items.
2. Test experts and developers should also ensure that previously generated items that have been used recently should not be featured for re-use until after a long while and its psychometric properties should be re-evaluated before subsequent usage.
3. The 4-parameter logistic model which incorporates response time and slowness parameter should be used for more robust IRT parameter estimation as it assumes that even high ability examinees can make mistakes due to carelessness as reflected by the non-zero upper asymptote (d-parameter) of the IRT logistic curve.

Hakem Değerlendirmesi: Dış bağımsız.

Çıkar Çatışması: Yazar, çıkar çatışması olmadığını beyan etmiştir.

Finansal Destek: Yazar, bu çalışma için finansal destek almadığını beyan etmiştir.

Peer-review: Externally peer-reviewed.

Conflict of Interest: The author have no conflicts of interest to declare.

Financial Disclosure: The author declared that this study has received no financial support.

Table 4

Item Parameter Estimates of NECO SSCE 2015 Economics Multiple-Choice Items Using The 1, 2 and 3 Parameter Logistic Models

	a		a		b		b		b		c	
Model	2PL	Remark	3PL	Remark	1PL	Remark	2PL	Remark	3PL	Remark	3PL	Remark
1	0.425	P	0.488	P	-1.811	E	-1.697	E	-1.622	E	0.003	NV
2	-0.431	P	-0.485	P	-1.499	E	-1.389	E	-1.534	E	0.003	NV
3	0.952	G	1.935	G	-0.881	M	0.923	M	-1.815	E	0.151	NV
4	-0.083	P	-0.156	P	-2.155	E	-1.956	E	-2.279	E	0.030	NV
5	1.144	G	1.148	G	0.486	M	0.557	M	0.774	M	0.016	NV
6	-0.092	P	-0.187	P	-1.603	E	-1.440	E	-1.642	E	0.027	NV
7	0.898	G	0.871	G	0.529	M	0.557	M	0.735	M	0.010	NV
8	1.367	G	2.781	G	-0.947	M	-1.113	E	-2.111	E	0.123	NV
9	1.920	G	2.034	G	2.013	D	2.701	D	3.339	D	0.007	NV
10	0.431	P	0.418	P	-0.311	M	-0.284	M	-0.217	M	0.016	NV
11	1.808	G	2.192	G	-0.733	M	-0.951	M	-0.813	M	0.036	NV
12	0.905	G	1.525	G	-0.927	M	-0.961	M	-1.569	E	0.132	NV
13	2.535	G	3.064	G	-0.381	M	-0.516	M	-0.370	M	0.052	NV
14	1.545	G	1.530	G	0.105	M	0.172	M	0.473	M	0.004	NV
15	0.867	G	0.876	G	-0.189	M	-0.185	M	-0.008	M	0.007	NV
16	1.213	G	1.572	G	-0.524	M	-0.577	M	-0.673	M	0.092	NV
17	0.974	G	1.870	G	-1.559	E	-1.660	E	-2.468	E	0.080	NV
18	1.524	G	2.905	G	-1.099	E	-1.357	E	-2.194	E	0.090	NV
19	0.428	P	0.505	P	-4.344	E	-4.179	E	-4.226	E	0.002	NV
20	1.236	G	2.217	G	-0.224	M	-0.233	M	-0.964	M	0.216	V
21	-0.370	P	-0.398	P	-1.621	E	-1.491	E	-1.641	E	0.009	NV
22	1.034	G	1.099	G	-0.482	M	-0.507	M	-0.293	M	0.005	NV
23	1.180	G	9.363	G	-0.446	M	-0.484	M	-6.257	E	0.267	V
24	1.105	G	2.645	G	0.116	M	0.146	M	-1.107	E	0.331	V
25	0.291	P	0.362	P	-3.951	E	-3.738	E	-3.772	E	0.003	NV
26	0.938	G	2.749	G	-0.494	M	-0.509	M	-2.124	E	0.244	V
27	-0.198	P	-0.260	P	-1.466	E	-1.320	E	-1.443	E	0.011	NV
28	0.930	G	0.960	G	-0.125	M	-0.120	M	0.075	M	0.007	NV
29	0.955	G	1.009	G	0.105	M	0.123	M	0.272	M	0.033	NV
30	0.892	G	0.952	G	-1.594	E	-1.659	E	-1.460	E	0.001	NV
31	0.250	P	4.162	G	-1.267	E	-1.149	E	-6.772	E	0.212	V
32	0.619	P	0.588	P	-0.671	M	-0.644	M	-0.524	M	0.008	NV
33	0.392	P	0.370	P	-1.603	E	-1.490	E	-1.619	E	0.037	NV
34	1.321	G	4.011	G	-0.305	M	-0.331	M	-1.959	E	0.236	V
35	-0.089	P	-0.058	P	-1.667	E	-1.498	E	-1.964	E	0.067	NV
36	0.623	P	0.705	P	-0.968	M	-0.935	M	-0.801	M	0.003	NV
37	0.252	P	0.301	P	-0.790	M	-0.711	M	-0.670	M	0.008	NV
38	0.220	P	0.299	P	-3.951	E	-3.720	E	-3.835	E	0.005	NV
39	0.014	P	-0.037	P	-0.954	M	-0.845	M	-0.945	M	0.026	NV
40	0.775	G	2.016	G	-1.525	E	-1.538	E	-2.800	E	0.104	NV
41	-0.416	P	-0.532	P	-2.232	E	-2.086	E	-2.300	E	0.006	NV
42	1.681	G	1.512	G	1.622	D	2.071	D	2.326	D	0.015	NV
43	0.476	P	8.186	G	-0.375	M	-0.347	M	-9.616	E	0.369	V
44	1.368	G	9.456	G	-0.536	M	-0.613	M	-5.895	E	0.230	V
45	-0.323	P	-0.307	P	-0.746	M	-0.669	M	-0.749	M	0.002	NV
46	0.493	P	0.491	P	-1.742	E	-1.651	E	-1.591	E	0.010	NV
47	0.649	P	6.433	G	-0.907	M	-0.882	M	-7.624	E	0.247	V
48	0.493	P	0.455	P	0.024	M	0.028	M	0.103	M	0.012	NV
49	0.208	P	0.307	P	-1.781	E	-1.620	E	-1.593	E	0.005	NV
50	-0.237	P	-0.272	P	-2.001	E	-1.826	E	-1.970	E	0.009	NV
51	1.080	G	2.710	G	-1.015	E	-1.105	E	-2.476	E	0.143	NV

Table 4 (continue)

Item Parameter Estimates of NECO SSCE 2015 Economics Multiple-Choice Items Using The 1, 2 and 3 Parameter Logistic Models

Model	a	Remark	a	Remark	b	Remark	b	Remark	b	Remark	c	Remark
	2PL		3PL		1PL		2PL		3PL		3PL	
52	1.878	G	3.613	G	-0.572	M	-0.738	M	-1.410	E	0.119	NV
53	1.902	G	1.990	G	-0.010	M	0.053	M	0.431	M	0.001	NV
54	-0.024	P	-0.048	P	-1.283	E	-1.143	E	-1.604	E	0.087	NV
55	1.144	G	1.335	G	-0.881	M	-0.972	M	-0.866	M	0.032	NV
56	0.971	G	1.049	G	-0.131	M	-0.126	M	0.016	M	0.032	NV
57	0.624	P	0.649	P	0.367	M	0.361	M	0.396	M	0.060	NV
58	2.208	G	2.743	G	-0.803	M	-1.158	E	-1.116	E	0.041	NV
59	-0.064	P	-0.393	P	-2.106	E	-1.9091	E	-3.341	E	0.092	NV
60	1.500	G	1.479	G	0.291	M	0.395	M	0.628	M	0.029	NV

References

Adebowale, O. F. (2007). *A study of Differential item functioning (DIF) in physics examinations in selected secondary schools in Lagos State* [Master thesis]. Obafemi Awolowo University, Ile-Ife, Nigeria.

Adedoyin, O. O., & Mokobi, T. (2013). Using irt psychometric analysis in examining the quality of junior certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science, 3*(4), 992-1011.

Adetutu, O. M., & Lawal, H. B. (2023). Applications of item response theory models to assess item properties and students' abilities in dichotomous responses items. *Open Journal of Educational Development (OJED), 3*(1), 1-19.

Ajeigbe, T. O., & Oderinde O. I. (2021). Assessing unidimensionality and item parameter estimates of four different paper types of english language multiple-choice tests using three-parameter model. *African Journal of Theory and Practice of Educational Assessment (AJTPEA), 10*(1), 1-18.

Baker, F. B., & Kim, S. H. (2004). *Item response theory: parameter estimation techniques* (2nd ed.). Taylor and Francis.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-472). Addison-Wesley.

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

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

Doan, C. H., Le, V. A., & Pham, U. H. (2016). Applying three-parameter logistic model in validating the level of difficulty, discrimination and guessing of items in a multiple-choice test. *Ho Chi Minh City University of Education Journal of Science, 7*(8), 174-184.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall, Engelwood Cliffs.

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

Gierl, M. J., Bisanz, J. Bisanz, G. L., Boughton, K. A., & Khaliq, S. N. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice, 20*, 26-36.

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

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principals and applications*. Kluwer Academic Publishers.

Henning, G. (1987). *A guide to language testing: Development, evaluation research*. New Berry House Publisher.

Izard, J. (2005a). *Overview of test construction: quantitative research methods in educational planning*. International Institute for Educational Planning/UNESCO Paris, France http://www.sacmeq.org and http://www.unesco.org/iiep

Izard, J. (2005b). *Trial testing and item analysis in test construction: Quantitative research methods in educational planning*. International Institute for Educational Planning/UNESCO Paris, France http://www.sacmeq.org and http://www.unesco.org/iiep

Liao, W., Ho, R., & Yen, Y. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality, 40*(10), 1679-1694. https://doi.org/10.2224/sbp.2012.40.10.1679

Linden, A. (2018). Review of Tenko Raykov and George Marcoulides's: A course in item response theory and modeling with stata. *The Stata Journal: Promoting Communications on Statistics and Stata, 18*(2), 485-488. https://doi.org/10.1177/1536867X1801800213

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology, 63*(3), 509-25. https://doi.org/10.1348/000711009X474502

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304-315.

McBride, N. L. (2001). *An item response theory analysis of the scales from the international personality item pool and the neo personality inventory-revised* [Doctoral dissertation]. Virginia Tech.

Nenty, H. J. (2015, November). *Conjugal relationship between research and measurement*. A keynote address delivered at 1st EARNIA conference in Cameroon.

Oguguo, B. C. E., & Lotobi, R. A. (2019). Parameters of basic science test item's of 2011 basic education certificate examination using item response theory (irt) approach in Delta State, Nigeria. *European Journal of Educational Sciences, EJES. 6*(1), 22-36. http://dx.doi.org/10.19044/ejes.v6no1a2

Ogunsakin, I. B., & Shogbesan, Y. O. (2018). Item response theory (irt): A modern statistical theory for solving measurement problem in 21st century. *International Journal of Scientific Research in Education (IJSRE), 11*(3B), 627-635.

Olatunji, D. S. (2007). *Effects of number of options on psychometic properties of multiple choice tests in economics* [M.Ed thesis]. University of Ilorin, Ilorin.

Olutola, A. T. (2015). Item difficulty and discrimination indices of multiple choice biology tests. *Liceo Journal of Higher Education Research, 11*(1), 16-30. https://doi:http://dx.doi.org/10.7828/ljher.v11i1.890

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Raykov, T., & Marcoulides, G. A. (2018). *A Course in item response theory and modeling with stata*. Stata Press College Station.

Rulison, K. L., & Loken, E. (2009). I've fallen and i can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*(2), 83-101. https://doi.org/10.1177/0146621608324023

Shamshad, B., & Siddiqui, J. S. (2020). Testing procedure for item response probabilities of 2class latent model. *Mehran University Research Journal of Engineering and Technology, 39*(3), 657-667. https://doi.org/10.22581/muet1982.2003.20

Shogbesan, Y. O. (2017). *Effect of test facets on the construct validity of economics achievement tests in osun state secondary schools* [M.A.Ed. thesis].Obafemi Awolowo University, Ile-Ife, Nigeria.

Shogbesan, Y. O. (2021). *Sensitivity of economics multiple-choice item parameters to item compromise among secondary school students in Ogun State, Nigeria*. [PhD. Thesis]. Obafemi Awolowo University, Ile-Ife, Nigeria.

Stout, W. (2005). *DIMTEST* (Version 2.0) [Computer Software]. The William Stout Institute for Measurement.

Thomas, M. L., Brown, G. G., Gur, R. C., Moore, T. M., Patt, V. M., Risbrough, V. B., & Baker, D. G. (2018). A signal detection–item response theory model for evaluating neuropsychological measures. *Journal of clinical and experimental neuropsychology, 40*(8), 745-760.

Văn Cảnh, N. (2021). Applying the item response theory with two-parameter, three-parameter models in the evaluation of multiple choice tests. *Tạp chí Khoa học Đại học Đồng Tháp, 10*(4), 17-28.

Wang, T. , & Hanson, A. (2001, April). *Development and an item response model that incorporates response time* [Conference presentation]. Annual meeting of the American Education Research Association in Settle.

Wiberg, M. (2004).*Classical test theory vs. item response theory: An evaluation of the theory test in the swedish driving-license test* (No. 50). Kluwer Academic Publications.