Journal of Integrated Elementary Education
ISSN 2776-1657 (online) 2828-223X (printed)
Volume 2, Number 1, March 2022, Page 43-54
Website: https://journal.walisongo.ac.id/index.php/jieed

Estimation of Item Parameter Indices of NECO Mathematics Multiple Choice Test Items among Nigerian Students

Jimoh Kasali, Adediwura Alaba Adeyemi*
Obafemi Awolowo University
*Correspondence author: yemtoyemtoy2000@gmail.com

DOI: https://doi.org/10.21580/jieed.v2i1.10187
Received: 24 December 2021, Revised: 1 January 2022,
Accepted: 21 March 2022, Published: 30 March 2022

Abstract

The study estimated the difficulty, discrimination, and vulnerability to guessing 2016 National Examination Council (NECO) Mathematics multiple-choice test items. The study employed an ex-post facto design with 276,338 samples. The research instruments used for the study were Optical Marks Record Sheets for the NECO June/July 2016 Mathematics objectives items. The responses of the testees were scored dichotomously. Data collected were calibrated using four parameters logistic model. The results showed that most items in the 2016 NECO Mathematics test were good as their difficulty parameters were within (-2 to 2). For those items, difficulty parameter estimates are considered good. Also, the results indicated that only 21.7% of 2016 NECO Mathematics test items had a very good discriminating power, and the majority of the items had poor discrimination power. The result implies that most of the items were not effective in discriminating between examinees with the required ability and those that lack the required ability. Finally, the result revealed that the 2016 NECO test items were not vulnerable to guessing (i.e., 86.7% of items were good in terms of guessing). The study concluded that the 2016 NECO Mathematics test items were good for difficulty and guessing parameter indices.

Keywords: item response parameter, item response theory, item discriminating, item difficulty, guessing parameter

INTRODUCTION

Item response parameters are not dependent on the sample used to generate the parameters. They are assumed to be invariant (within a linear transformation) across divergent groups within a research population and populations (Reeve, 2002). Item response theory models are described by the number of parameters they use. Estimation, model fit and equating. The one-parameter logistic (IPL) model

has generated substantial research during the review period. As is characteristic of research on IRT models, much of the basic research has been focused on problems of item parameter estimation. Since the IPL model parameter estimation procedure involves estimating only the difficulty parameters for items and the ability parameters for individuals, these two parameters are usually estimated simultaneously.

The study's research question is: to know how difficult the 2016 NECO Mathematics test items are, how discriminating the 2016 NECO Mathematics test items are, and how vulnerable to guessing are the 2016 NECO Mathematics test items?

The one-parameter logistic (1PL) model assumes that data have no discrimination and guessing. A single parameter only describes items in terms of location or difficulty (bi). The results in one-parameter models have specific objectivity properties; the rank of the item difficulty is the same for all respondents and independent of ability (Van Schuur, 2003). The level of the person's ability is the same for items independently of difficulty. The equation for the one-parameter model is given by the following:

$$\mathrm{P}(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-1(\theta-b)}}$$

Where: b is the difficulty parameter, and θ is the ability level. The above theories show that the item response theory is the modern theory that describes the students' ability to use the item by item performance rather than the classical test theory.

The 2-parameter (2P) and 3-parameter (3P) IRT models are simply generalizations of the IPL model, including additional parameters that describe aspects of the IRF. The 2P model permits items to vary in the discrimination parameter, and the 3P model adds the lower asymptote (pseudo-chance value) to the IRF. Being generalizations of the 1P model, the applications and utility of these models are essentially the same. They can provide sample-free measures of individuals, resulting in the same degree of "objectivity" as does the 1PL model. These IRT models also permit the measurement of individuals with any subset of items. However, the number-correct score for these models does not convey the same information as it does for the IPL model.

Consequently, new scoring methods have been developed to implement these models, as have additional methods for estimating item parameters. The two-parameter logistic (2PL) model assumes that the data have no guessing but that items can vary in terms of location (bi), i.e., difficulty and discrimination (ai). The equation for the two-parameter model is given below:

$$\mathrm{P}(\theta) = \frac{1}{1+e^{-L}} = \frac{1}{1+e^{-a\,(\theta-b)}}$$

Where: e is the constant, b is the difficulty parameter, a is the discrimination parameter, L = a (θ - b) is the logistic deviate (logit), and θ is an ability level. The difficulty parameter, denoted by b, is defined as the point on the ability scale is the probability of a correct response to the item.

The three-parameter logistic (3PL) model is named so because it employs three item parameters. Such as item difficulty, discrimination, and guessing parameter. The equation for the three-parameter model is: $P(\theta) = C + (1 + C) = \frac{1}{1+e^{-a(\theta-b)}}$ Where: b is the difficulty parameter, a is the discrimination parameter, c is the guessing parameter, and θ is the ability level. The parameter c is the probability of getting the item correct by guessing alone. It is important to note that by definition, the value of c does not vary as a function of the ability level. Thus, the lowest and highest ability examinees have the same probability of getting the item correct by guessing. One important characteristic of these models is that, like the lPL model's dichotomous case, integer scoring using equally distant weights preserves the IPL model attributes (Jimoh et al., 2020). Consequently, complex scoring procedures, such as are characteristic of the other IRT models, are not required.

Psychometrics is a field of study concerned with the theory and technique of psychological measurement (Ariyo, 2015). One part of the field is concerned with the objective measurement of skills and knowledge, abilities, attitudes, personality traits, and educational achievement. For example, some psychometric researchers have thus far concerned themselves with constructing and validating assessment instruments such as questionnaires, tests, raters judgments, and personality tests. Another field is concerned with statistical research on measurement theory (e.g., item response theory; interclass correlation).

As a result of these areas of focus, psychometric research involves two major tasks: (i) the construction of instruments and (ii) the development of procedures for measurement. "Psychometrics, or quantitative psychology, is the disciplinary home of statistical models and methods developed primarily to summarize, describe, and draw inferences from empirical data collected in psychological research" (Jones & Thissen, 2007). Psychometric requirement demands that such items be trial-tested while the responses and scores generated are subjected to statistical item analyses. Ary, Jacobs, & Razavieh (2002) opined that item analysis involves using statistics to provide relevant information for improving the quality and accuracy of multiple-choice questions. There are three popular forms of item analyses: item difficulty index, distractive index, and discriminatory index.

Item difficulty index indicates the degree of difficulty of the MCQ items concerning the cognitive ability of the testees (Boopathiraj & Chellamani 2013). It is calculated by finding the proportion of the testees that got the item correctly. An item is adjudged too difficult when the index is below 0.3. An item is adjudged too easily when the index is above 0.7. Depending on the test's purpose, the cut-off

points for easy or difficult things can be adjusted upward or downward. Generally, the rule is that life-sensitive or competitive activities require more technical/difficult items in screening.

In contrast, less sensitive activities or activities requiring motivation of testees often use less difficult things. For most summative assessments, such as those handled by the West African Examinations Council, a moderate difficulty index ranging around 0.5 is often preferred (Odukoya. et al. 2018). Item discrimination compares the number of high scorers and low scorers who answer an item correctly. It is how items discriminate among trainees in the high and low groups. The total test and each article should measure the same thing. High performers should be more likely to answer a good item correctly, and low performers should be more likely to answer incorrectly. Scores range from –1.00 to +1.00, with an ideal score of +1.00. Positive coefficients indicate that high-scoring examinees tended to have higher scores on the item, while a negative coefficient indicates that low-scoring students tended to have lower scores. More high scorers than low scorers will answer those items correctly on entities that discriminate well. The higher the discrimination index, the better the thing because high values indicate that the item discriminates in favor of the upper group, which should answer more items correctly. If more low scorers answer an item correctly, it will have a negative value and is probably flawed. A negative discrimination index occurs for too hard or poorly written items, making it difficult to select the correct answer. On these items, poor students may guess correctly, while good students, suspecting that a question is too easy, may answer incorrectly by reading too much into the question.

METHODOLOGY

The study employed an ex-post facto design. The population for the study comprised all candidates who enrolled and sat for June/July SSCE 2016 NECO Mathematics Examination in Nigeria. The sample for the survey comprised 276,338 candidates who sat for the examination in three purposively Geo-political zones in Nigeria (i.e., South-West, South-East, and North-West). The research instruments used for the study were Optical Marks Record Sheets for the National Examination Council (NECO) June/July 2016 SSCE Mathematics objectives items. The responses of the testees were scored dichotomously. Items were calibrated using the 4PL model.

RESULTS

1. How difficult are the 2016 NECO Mathematics test items?

The responses of the sampled examinees to the 2016 NECO Mathematics test items were calibrated using the 4PL model, and the result difficulty parameter is presented in table 1.

Table 1

Item difficulty parameters of 2016 NECO Mathematics test items

Item	b	Item	B
it1	54.272	it31	0.048
it2	-0.516	it32	0.007
it3	-1.022	it33	-0.479
it4	-0.043	it34	0.92
it5	0.417	it35	-0.084
it6	0.417	it36	0.466
it7	-0.985	it37	0.001
it8	-0.502	it38	-0.412
it9	-1.545	it39	0.106
it10	0.883	it40	0.417
it11	-0.994	it41	-0.069
it12	-0.189	it42	-1.54
it13	0.116	it43	-0.25
it14	-0.267	it44	88.422
it15	0.31	it45	98.73
it16	-0.396	it46	0.173
it17	-0.465	it47	0.119
it18	0.02	it48	-0.525
it19	-1.293	it49	-0.093
it20	-0.154	it50	-0.653
it21	-1.103	it51	-0.152
it22	-0.329	it52	-0.109
it23	228.23	it53	-0.046
it24	0.14	it54	-0.289
it25	-0.073	it55	-1.59
it26	0.324	it56	0.984

it27	-0.504	it57	-0.341
it28	-0.514	it58	-0.08
it29	-0.185	it59	-0.202
it30	-0.031	it60	-0.399

Table 1 shows the difficulty level of the 2016 NECO Mathematics multiple-choice test. In Table 1, the estimate column (b) represents the difficulty indices of the test items. These estimates indicate how easy or hard the items were for the students. Easier items have lower (negative) difficulty indices, with very easy items having values less than -2, and harder items have higher (positive) indices, with very hard items having values greater than 2.

The table showed that most of the items are good. A further look at the table showed that item 2 was the easiest and item 39 was the hardest. The item difficulty, or b-parameter, is on the same metric as ability (θ). When b = θ, the probability of a correct response is 50% for the 1PL and 2PL models (somewhat higher for the 3PL model). For example, items 1, 23, 44, and 45 had difficulty index of 54.272, 228.23, 88.422, and 98.73, respectively, so more than 50% of examinees with θ =54.272, 228.23, 88.422, and 98.73 would find it very difficult to answer items 1, 23, 44 and 45 correctly. More importantly, the table showed that apart from items 1, 23, 44, and 45. Other items in the 2016 NECO Mathematics test items were good items. Their difficulty parameters were within the range (-2 to 2) for which items' difficulty parameter estimates are considered good (Baker, 2001; Hambleton & Jones; De Mars, 2010).

2. How discriminating are the 2016 NECO Mathematics test items?

Table 2 shows the item discriminating parameters for the 2016 NECO Mathematics test items. These parameters show how an item can discriminate examinees with low ability from those with high ability in the 2016 NECO Mathematics test items. For discrimination indices, items with values ranging from 0–2 are considered good items, while items outside 0–2 are considered poor items. Here, only 13 items (items 1, 2, 3, 4, 6, 7, 15, 19, 21, 23, 44, 45, and 47) are found within 0–2. Therefore, only 21.7% of the 2016 NECO Mathematics test items had a very good discriminating power, and most of the items had poor discriminating power. The result implies that most of the items were not effective in discriminating between examinees with the required ability and those that lack the required ability.

Table 2

Item Discriminating Parameters of 2016 NECO Mathematics Test Items

Item	A	Decision	Item	a	Decision
it1	0.1	**Good**	it31	6.833	Poor
it2	5.9	**Good**	it32	2.49	Poor
it3	1.2	**Good**	it33	4.428	Poor
it4	1.8	**Good**	it34	3.798	Poor
it5	2.3	Poor	it35	2.471	Poor
it6	1.7	**Good**	it36	4.314	Poor
it7	1.1	**Good**	it37	2.811	Poor
it8	25.8	Poor	it38	3.909	Poor
it9	19.5	Poor	it39	2.31	Poor
it10	3.0	Poor	it40	5.158	Poor
it11	1.3	Poor	it41	2.675	Poor
it12	1.8	Poor	it42	22.561	Poor
it13	1.7	Poor	it43	3.034	Poor
it14	3.5	Poor	it44	0.133	**Good**
it15	1.9	**Good**	it45	0.119	**Good**
it16	3.2	Poor	it46	21.32	Poor
it17	4.8	Poor	it47	1.671	**Good**
it18	2.7	Poor	it48	6.133	Poor
it19	1.8	**Good**	it49	5.215	Poor
it20	2.2	Poor	it50	9.365	Poor
it21	1.4	**Good**	it51	4.743	Poor
it22	6.4	Poor	it52	2.465	Poor
it23	0.0	**Good**	it53	3.147	Poor
it24	3.6	Poor	it54	3.219	Poor
it25	6.0	Poor	it55	-16.481	Poor
it26	6.6	Poor	it56	-3.869	Poor
it27	7.5	Poor	it57	6.068	Poor

it28	6.8	Poor	it58	4.765	Poor
it29	3.5	Poor	it59	3.297	Poor
it30	2.5	Poor	it60	6.014	Poor

3. How vulnerable to guessing are the 2016 NECO Mathematics test items?

Table 3 shows the vulnerability to guessing the test items. According to Baker (2001) and Hambleton and Jones (1993), items greater than 0.35 of the benchmark are vulnerable to guessing. The table shows that based on the parameter estimate, eight items (items 5, 8, 26, 40, 44, 45, 49, and 51) are vulnerable to guessing, while 52 (86.7%) items were good in terms of guessing. The result showed that the 2016 NECO test items were not vulnerable to guessing.

Table 3

Item Guessing Parameters of 2016 NECO Mathematics Test Items

Item	c		Item	c	
it1	0.00	Good	it31	0.31	Good
it2	0.04	Good	it32	0.01	Good
it3	0.00	Good	it33	0.03	Good
it4	0.00	Good	it34	0.35	Good
it5	0.39	Vulnerable	it35	0.00	Good
it6	0.00	Good	it36	0.26	Good
it7	0.00	Good	it37	0.04	Good
it8	0.37	Vulnerable	it38	0.01	Good
it9	0.00	Good	it39	0.04	Good
it10	0.33	Good	it40	0.39	Vulnerable
it11	0.00	Good	it41	0.03	Good
it12	0.00	Good	it42	0.00	Good
it13	0.00	Good	it43	0.01	Good
it14	0.02	Good	it44	0.39	Vulnerable
it15	0.01	Good	it45	0.37	Vulnerable
it16	0.01	Good	it46	0.27	Good
it17	0.01	Good	it47	0.00	Good

it18	0.00	Good	it48	0.04	Good
it19	0.00	Good	it49	0.39	Vulnerable
it20	0.00	Good	it50	0.04	Good
it21	0.00	Good	it51	0.38	Vulnerable
it22	0.35	Good	it52	0.02	Good
it23	0.22	Good	it53	0.01	Good
it24	0.04	Good	it54	0.02	Good
it25	0.32	Good	it55	0.15	Good
it26	0.62	Vulnerable	it56	0.00	Good
it27	0.02	Good	it57	0.05	Good
it28	0.02	Good	it58	0.33	Good
it29	0.02	Good	it59	0.01	Good
it30	0.05	Good	it60	0.02	Good

DISCUSSION

The analysis results also revealed that most of the items in the 2016 NECO Mathematics test items were suitable because their difficulty parameters were within the range (- 2 to 2) for which item trouble boundary gauges are considered good. This research supports Olutola's (2016) findings that the WAEC SSCE multiple-choice Biology test is more difficult than the NECO SSCE multiple-choice Biology test. The mean difficulty of the WAEC SSCE multiple-choice Biology test is 0.42, while the NECO SSCE multiple-choice Biology test is 0.40.

Abiri's (2006) findings were also verified, stating that trouble files of multiple-choice tests with fewer options are superior to those with a larger number of options. Similarly, Nevid and McClelland (2013) found that students had difficulties reacting to evaluation and clarification inquiries at high psychological levels in Bloom's scientific categorization for a brain science course. These inquiries were the most important distinguishable for high-performing and low-performing students. Kim et al. (2012) discovered that the difficulty records of multiple-choice inquiries in drug store learning at the recollecting, understanding, and applying levels is higher than the inquiries at the review and union/assessment levels in another investigation.

Furthermore, the discrimination parameters in the 2016 NECO Mathematics test items indicate how an item can separate examinees with low capacity from those with high capacity. Objects with 0 to 2 are considered excellent segregating

items for prejudice lists, whereas those with values beyond 0 – 2 are considered powerless items. Only 13 objects (numbers 1, 2, 3, 4, 6, 7, 15, 19, 21, 23, 44, 45, and 47) are included within the 0–2 range. Following this discovery, it was discovered that only 21.7 percent of 2016 NECO Mathematics test items had a generally excellent separating force, with the majority of the items having helpless discrimination ability.

The findings imply that most items were ineffective in distinguishing between examinees who possessed the requisite ability and those who lacked it. Wiersma and Jurs (1990) stated that when students perform abysmally well or poorly, educators should investigate whether the low or high performance is due to a flaw in the test items guidelines or the students' abilities before making appropriate adjustments.

Furthermore, it has been observed that students' performance in Mathematics varies from year to year. On a serious note, the degree of good or bad showings is not constant but varies year to year, with lackluster showing having the greater prevalence. The variation in student presentation has been attributed to several variables; some attribute it to test features, while others attribute it to individual factors, such as ability level. As a result of these investigations into students' educational exhibitions, it was discovered that the psychometric properties could often trigger students' exhibitions in a test.

For example, Onunkwo (2002) stated that students' dissatisfaction is often due to issues inherent in the test's psychometric properties rather than their ineptitude. This error in a test's psychometric properties includes, among other things, the consideration of difficult objects, items with low to no discrimination ability, and insufficiently large numbers of choices (Abiri, 2006). The number of options in a test object thus influences the test's psychometric properties. Especially, items with fewer options progress to higher difficulty levels than those with a greater number of options. Olatunji in Olutola (2015) explained that objects with fewer options are segregated more effectively than those with a larger number of options.

CONCLUSION

The study concluded that most of the items in the 2016 NECO Mathematics test items were suitable because their difficulty parameters were within the range (- 2 to 2) for which item trouble boundary gauges are considered good.

REFERENCES

Ariyo, A. O. (2015). *Ensuring quality in the test development process through innovations in item calibration: a comparison of classical test theory and item response theory* [Conference presentation]. 33rd AEAA Conference, Accra, Ghana.

Ary, D., Jacobs, L.C., & Razavieh, A. (2002) Introduction to Research in Education, sixth eds. Wadsworth, Californi. https://books.google.co.id/books?hl=id&lr=&id=4RREDwAAQBAJ

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International journal of social science & interdisciplinary research*, *2*(2), 189-193.

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). United States of America: ERIC Clearing House on Assessment and Evaluation. https://eric.ed.gov/?id=ED458219

DeMars, C. (2010). *Item response theory: understanding statistics measurement.* City: Oxford University Press. https://books.google.co.id/books?hl=id&lr=&id=KOADeYBt7sIC

Hambleton, R. K., & Jones R. W. (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. https://eric.ed.gov/?id=EJ471935

Jimoh, M. I., Daramola, D. S., Oladele, J. I., & Sheu, A. L. (2020). Assessment of Items Prone to Guessing in SSCE Economics Multiple-Choice Tests among Students in Kwara State, Nigeria. *Anatolian Journal of Education*, *5*(1), 17-28. https://eric.ed.gov/?id=EJ1249146

Jones L.V, Thissen D (2007) A History and Overview of Psychometrics. Handbook of Statistics, Vol. 26, Sandip Sinharay, Hardbound. https://doi.org/10.1016/S0169-7161(06)26001-2

Kim, J., Frisbie, D. A., Kolen, M. J., & Kim, D. I. (2012). *A comparison of calibration methods and proficiency estimators for creating IRT vertical scales* [Paper presentation]—National Council on Measurement in Education Annual Meeting, Chicago, IL. https://www.proquest.com/openview/f28ebd6e63601cb5c61ecb89fde9e218/1?pq-origsite=gscholar&cbl=18750

Odukoya, J.A., Adekeye, O., Igbinoba, A.O., *et al.* Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university. *Qual Quant* **52,** 983–997 (2018). https://doi.org/10.1007/s11135-017-0499-2

Olutola, A. T. (2016). Assessing students' performance in senior school certificate multiple-choice test in biology. *Journal Issues and Ideas in Education, 4*(1), 11-20. https://doi.org/10.15415/iie.2016.41001

Onunkwo, G. I. N. (2002*). Fundamentals of education measurement and evaluation*. Owerri: Cape Publishers Int'l Ltd.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Massachusetts: Allyn and Bacon.

Revee, J. (2002). Self-determination theory applied to educational settings. In: E. L. Deci & R. M. Reyan (Eds.), *Handbook of self-determination research.* (pp. 183-203). Rochester, NY: University of Rochester Press. https://books.google.co.id/books?hl=id&lr=&id=DcAe2b7L-RgC

Van Schuur, W. (2003). Mokken Scale Analysis: Between the Guttman Scale and Parametric Item Response Theory. *Political Analysis, 11*(2), 139-163. https://doi.org/10.1093/pan/mpg002