ASSEREN
Journal of Educational Research and Development
VOLUME 11, JUNE 2024 PAGE 30-37
ISSN: 2536-6899 E – ISSN: 2814 – 3248 Website: www.theasseren.org.ng

ASSESSMENT OF ITEM PARAMETERS OF 2017 NATIONAL EXAMINATIONS COUNCIL'S ENGLISH LANGUAGE MULTIPLE CHOICE TEST USING ITEM RESPONSE THEORY

Bolaji, G. T., Adediwura, A. A. & Omidiora J. F

Submitted: 16 October, 2023 Revised: 25 April, 2024 Accepted: 29 May, 2024

Abstract

The study aimed to evaluate the adequacy of the NECO 2017 English Language examination items in measuring their intended objectives among secondary school students in Osun State. It assessed the dimensionality, local independence, item difficulty, and discrimination indices of the examination. A non-experimental design of descriptive research type was employed, involving 26,127 senior secondary students who sat for the exam in June/July 2017. Using a stratified random sampling technique, the students were divided based on characteristics such as gender, school type, and academic performance. Data from the NECO headquarters, including responses and scores, were analyzed using the two-parameter logistic (2PL) IRT model. Three research questions were addressed, revealing that the test was unidimensional with a maximum DETECT value of -0.1331, ASSI of -0.211, and RATIO of 0.142. However, 30 item pairs were locally dependent. Additionally, 19 out of 100 items were found to be of poor quality based on difficulty parameters outside the acceptable range (-3 to 3), and 29 items had poor discrimination indices (less than zero). The study concluded that while the exam's multiple-choice items were largely unidimensional and independent, suggesting reliable scores. It was recommended that there is a need to identify specific skills influencing student performance and develop targeted interventions to improve overall test performance.

Keywords: Unidimensionality, Item difficulty, Item discrimination, Validity, Reliability.

Introduction

The importance of English language is stressed in the National Policy on Education, making it a central topic for students in the Nigerian educational system. Apart from being a powerful learning tool, it is also a means through which people can access information from all over the world, as well as develop positive values and attitudes, build and maintain meaningful relationships with others, increase their cultural awareness, and broaden their knowledge and worldviews. The significance of the English Language outspreads beyond enhancing academic and social growth, educational achievement, career progression, personal fulfilment, and cultural understanding. Adesulu (2012) highlighted the significance of the 2002 NECO/SSCE examination, noting it as the inaugural session. This year marked a positive turning point for many secondary school graduates, with a notable increase in candidates passing their registered subjects, including English Language, in contrast to previous years under WAEC/SSCE. The success trend continued to rise until after 2007. The preceding five years continuously saw a poor decline of test takers performance. According to (Faleye & Olajide 2012), standardized achievement test are tests whose items have been validated and commercially produced mostly by public examination bodies. Examples of standardized tests are tests conducted by examination bodies in Nigeria such as WAEC, NABTEB and NECO among others in Nigeria. Adediwura (2012), highlighted the features of a standardized test which are as follows:

1. The test items should be of high technical quality, established by educational and testing experts, tested experimentally (pre – test), and chosen based on difficulty (facility), discriminating ability, and relationship to clearly specified and rigid collection of speculations.
2. The instruction for conducting and scoring the tests should be clear, and the procedures be consistent across users. The developer of a standardized test strives for uniform directions for all the examinees and favourable environment conditions.
3. As a tool for interpreting test results, using norms based on representative groups of people. The standard allows an individual's test score to be compared to that of identified groups.
4. The test manual and other required materials are included as a guide for conducting, scoring, assessing the technical qualities of the exam, likewise interpreting and using the findings.

Early in the 20th century is when bias analysis first emerged (McNamara & Roever, 2006), Researchers at that time were concerned with developing tests that measured 'raw intelligence'. However, a number of studies carried out at the time demonstrated that the test takers' socioeconomic status was a confounding factor. Flores (2000) and Lam (1995) found that a test can be biased, affecting examinees' scores disproportionately due to the existence of non-target constructs such as gender, ethnicity, race, socioeconomic status. IRT tests which measure item parameters (i.e., item complexity and item discrimination) that are independent of the sample of examinees, are used in parametric DIF method .It's the same as statistical bias, which occurs when one or more statistical model parameters are underestimated or overestimated (Camilli, 2006).

Item response theory (IRT) is a set of latent variable techniques for modelling the interaction between a subject's "ability" and item level stimuli (difficulty, guessing, etc) (Chalmers, 2012). Item parameter pertains to evaluating the adequacy of every item comprising the test instrument. It is categorized into two basic components which include: the parameter relating to individual examinee and the parameter relating to each of the items of a given test. In the parameter relating to individual examinee, each examinee responding to a test item is presumed to have some level of underlying capacity. That is, each examinee is considered to have a score, a numerical value which makes him/her to be placed somewhere on the ability scale. While for the parameter relating to each of the items of a given test the analysis of an item includes a number of statistics that can help improve the consistency and accuracy of multiple choice items, therefore items in a given task are characterized by, at most, three parameters depending on the type of logistic model being considered which include:

i. discrimination index of the test item represented by the letter a

ii. difficulty index of the test represented by letter b and

iii. vulnerability to guessing index represented by letter c.

The degree to which an examinee's response to an item in a cognitive task varies with or related to their characteristic or ability level is indicated by the a-parameter. The item or task's cognitive resistance is represented by the b-parameter and the c-parameter shows the likelihood that a person who fully lacks the trait under estimation will answer the question correctly (Dibu Ojerinde et al, 2012). It is possible to create an item that can effectively differentiate between individual, the item difficulty and item discrimination are independent of the sample. The achievement of an item of a known difficulty level may be used to characterize a testee's proficiency. Since the English language is widely known around the world, it is important for students to learn it. Despite being required as a pre-requisite for admission into universities and as a core subject in secondary schools, students still performs poorly in it. To address students' English language success, various initiatives have been implemented, including improvements in the standard of instructional methods used by teachers, the availability of trained English language teachers, and so on. However, despite the steps in place, students' performance in English language in external examinations has not been impressive. Thus, there is need to investigate if NECO English Language multiple-choice Examination questions provide invariant measurement despite using the result for comparing students' performance across the country. Hence the study. The specific objectives of the study were to:

a. assess the dimensionality of the NECO 2017English language multiple choice examination items among Osun state Secondary School students;

b. determine the item difficulty of the NECO 2017 English Language multiple choice items; and

c. determine the item discrimination of the NECO 2017 English Language multiple choice items.

The following research questions were derived from the research objectives

1. how many dimensions underlie the NECO 2017 English Language test among Osun state secondary schools students?
2. what is the level of item difficulty of the NECO 2017 multiple choice items in the English language?
3. what is the level of item discrimination of the NECO 2017 multiple choice items in the English language?

Methodology

The study adopted non-experimental design of descriptive research type. The population and sample of the study comprised 26,127 senior secondary school students in Osun State that sat for National Examination Council (NECO) English Language Multiple-choice (paper II) in June/July 2017. The Optical Marks Record sheets for the June/July 2017 English Language multiple choice questions from the National Examination Council (NECO) serve as the study's instruments. Examinees' answers to the NECO June/July 2017 English Language multiple choice questions were contained in the OMR sheets. There are five response alternatives for each of the 100 multiple-choice questions in the English Language exam, which is a dichotomously scored test. Examinees' responses were given a score of 1 or 0 for right and wrong answers. The responses and scores of candidates who wrote the NECO English Language SSCE June/July 2017 in Osun state as indicated on the OMR sheets were collected from NECO headquarters. The 2PL IRT model was used to analyze the collected data. Research question one was analysed using Stout's test of essential unidimensionality test (Zhang & Stout, 1999) implemented in Supplementary Item Response Theory Models, (SIRT) package (Robitzsch, 2019) of R Language and Environment for statistical computing (R Core, 2019). Research question two and three were subjected to test calibration based on 2PL IRT model that fitted the test data, the difficulty and discrimination estimates of the multiple choice test items were extracted.

Results

Research Question One: How many dimensions underlie the NECO 2017 English language multiple choice test items among Osun State secondary schools students?

Table 1: Unidimensionality assessment of 2017 NECO English language multiple choice test

	Unweighted	Weighted
DETECT	-0.1330967	-0.1330967
ASSI	-0.2109091	-0.2109091
RATIO	-0.1418236	-0.1418236

Table 1 displays the outcome of the NECO English Language multiple choice test's dimensionality evaluation. The maximum DETECT value = -0.1331 (< .20), ASSI = -0.211 (< 0.25) and RATIO = -0.142 (< 0.36)) indicate that the 2017 NECO English language multiple choice test was largely unidimensional. As a result, the unidimensionality assumption was not rejected. This result showed that one dominant dimension accounted for the variation observed in student's responses to the English multiple choice test items. Based on the result it can be concluded that the 2017 NECO English language multiple choice test was unidimensional, with variances in candidates' performance being explained by a single dominating skill.

Research Question two: What is the level of item difficulty of the NECO 2017 English Language Items?

Table 2: Difficulty parameter of the 2017 NECO English language test items

Item	Est	Remark	Item	Est	Remark	Item	Est	Remark	Item	Est	Remark
1	-1.98	Good	26	-1.34	Good	51	-0.99	Good	76	-1.61	Good
2	-0.69	Good	27	-0.01	Good	52	-2.07	Good	77	-1.90	Good
3	-3.37	Poor	28	-1.34	Good	53	-3.49	Poor	78	-2.05	Good
4	-1.13	Good	29	-2.22	Good	54	-0.75	Good	79	-0.91	Good
5	-1.90	Good	30	-2.97	Good	55	-3.90	Poor	80	-0.48	Good
6	-1.21	Good	31	-0.98	Good	56	-1.83	Good	81	-4.15	Poor
7	-3.23	Poor	32	-1.83	Good	57	-4.92	Poor	82	-1.03	Good
8	-2.27	Good	33	-1.88	Good	58	-1.70	Good	83	-0.71	Good
9	-2.66	Good	34	-1.43	Good	59	-0.71	Good	84	-0.48	Good
10	-0.83	Good	35	-0.64	Good	60	-5.76	Poor	85	-1.42	Good
11	-0.51	Good	36	-1.94	Good	61	-0.53	Good	86	-2.87	Good
12	-1.03	Good	37	-1.55	Good	62	-2.73	Good	87	-2.41	Good
13	-1.10	Good	38	-3.05	Poor	63	-0.20	Good	88	-2.23	Good
14	-4.64	Poor	39	-2.03	Good	64	-2.10	Good	89	-2.12	Good
15	-3.45	Poor	40	-2.13	Good	65	-3.12	Poor	90	-1.01	Good
16	-0.24	Good	41	-0.68	Good	66	-1.50	Good	91	-3.80	Poor
17	-4.82	Poor	42	-2.39	Good	67	-3.44	Poor	92	-0.90	Good
18	-2.57	Good	43	-1.20	Good	68	-0.29	Good	93	-6.63	Poor
19	-2.19	Good	44	-2.36	Good	69	-0.99	Good	94	-2.23	Good
20	-1.45	Good	45	-2.86	Good	70	-0.23	Good	95	-2.19	Good
21	-2.26	Good	46	-0.53	Good	71	-0.59	Good	96	-4.63	Poor
22	-1.07	Good	47	-2.94	Good	72	-7.98	Poor	97	-15.11	Poor
23	-1.22	Good	48	-2.11	Good	73	-1.17	Good	98	-2.23	Good
24	-3.19	Poor	49	-2.21	Good	74	-2.92	Good	99	-0.38	Good
25	-1.81	Good	50	-0.43	Good	75	-1.64	Good	100	-1.48	Good

est = estimates

Table 2 presented the items in the 2017 NECO English Language test in terms of difficulty. The test item difficulty indices are represented in the table's estimate column, and the remarks reflect the assessment of how appropriate the item difficulty level is in relation to established standards. These estimation show the students' level of difficulty or ease with the tasks. The difficulty indices of harder items are higher (positive) and those of easier items are lower (negative). The majority of the items had a decent degree of difficulty according to the table. The table showed that 19 items (item 3, 7, 14, 15, 17, 24, 38, 53, 55, 57, 60, 65, 67, 72, 81, 91, 93, 96, 97) representing 19.0% of the test's items were poor as their difficulty parameters (-3.37, -3.23, -4.64, -3.45, -4.82, -3.19, -3.05, -3.49, -3.90, -4.92, -5.76, -3.12, -3.44, -7.98, -4.15, -3.80, -6.63, -4.63 and -15.11 respectively) were outside the range (-3 to 3) for which items difficulty parameter estimates are considered good (Baker, 2001; Hambleton & Jones; De Mars, 2010). The remaining 81.0% of the test items were of good level of difficulty. The result showed that the 2017 NECO English language test items has acceptable difficulty level.

Research Question Three: What is the level of item discrimination of the NECO 2017 English Language Items?

Item	Est	Remark	Item	Est	Remark	Item	Est	Remark	Item	est	Remark
1	1.06	Good	26	-1.50	Poor	51	1.26	Good	76	-1.27	Poor
2	-1.66	Poor	27	2.72	Good	52	0.88	Good	77	1.00	Good
3	0.79	Good	28	1.25	Good	53	0.71	Good	78	0.73	Good
4	1.33	Good	29	0.91	Good	54	-1.61	Poor	79	1.25	Good
5	0.98	Good	30	0.73	Good	55	0.71	Good	80	-1.74	Poor
6	1.09	Good	31	1.39	Good	56	0.90	Good	81	-0.68	Poor
7	0.60	Good	32	1.09	Good	57	0.45	Good	82	-1.44	Poor
8	0.83	Good	33	1.09	Good	58	0.91	Good	83	-1.96	Poor
9	0.83	Good	34	1.31	Good	59	-1.90	Poor	84	1.60	Good
10	1.45	Good	35	1.82	Good	60	0.32	Good	85	0.81	Good
11	-1.78	Poor	36	0.98	Good	61	-1.85	Poor	86	0.54	Good
12	1.27	Good	37	1.22	Good	62	0.76	Good	87	0.69	Good
13	1.52	Good	38	0.65	Good	63	-2.24	Poor	88	-0.94	Poor
14	0.46	Good	39	0.83	Good	64	0.86	Good	89	-1.01	Poor
15	0.55	Good	40	0.80	Good	65	0.58	Good	90	1.01	Good
16	-1.83	Poor	41	-1.79	Poor	66	0.89	Good	91	-0.75	Poor
17	0.48	Good	42	0.97	Good	67	0.50	Good	92	1.52	Good
18	0.61	Good	43	-1.76	Poor	68	-2.23	Poor	93	-0.47	Poor
19	0.87	Good	44	0.84	Good	69	1.49	Good	94	-1.09	Poor
20	1.03	Good	45	0.77	Good	70	2.07	Good	95	-1.26	Poor
21	1.07	Good	46	-1.50	Poor	71	-2.01	Poor	96	-0.63	Poor
22	1.53	Good	47	0.68	Good	72	-0.54	Poor	97	-0.24	Poor
23	1.52	Good	48	0.87	Good	73	1.28	Good	98	-1.07	Poor
24	0.68	Good	49	0.94	Good	74	0.68	Good	99	1.62	Good
25	1.20	Good	50	-1.84	Poor	75	1.12	Good	100	0.97	Good

The discrimination of the 2017 NECO English language test items is displayed in Table 3. The estimate column in the table shows the test item discrimination indices, and the remark column shows the adequacy of how well the item level of discrimination meets specified standards. These estimations demonstrate the degree to which high an item can discriminate and how discriminating the items were between examinee with low ability and those with high ability in English language multiple choice test. Poorly discriminating item has lower (negative) discriminating indices and appropriate discriminating items have higher (positive) discrimination indices. The table shows that 29 items (item 2, 11, 16, 26, 41, 43, 46, 50, 54, 59, 61, 63, 68, 71, 72, 76, 80, 81, 82, 83, 88, 89, 91, 93, 94, 95, 96, 97 and 98) representing 29.0% of the test's items returned discrimination values that were less than zero (-1.66, -1.78, -1.83, -1.50, -1.79, -1.76, -1.50, -1.84, -1.61, -1.90, -1.85, -2.24, -2.23, -2.01, -0.54, -1.27, -1.74, -0.68, -1.44, -1.96, -0.94, -1.01, -0.75, -0.47, -1.09, -1.26, -0.63, -0.24 and -1.07 respectively). The result showed that the items were poor as their discrimination indices were less than 0.4 the minimum discrimination parameter an item must have to be considered good (De Mars, 2010). The remaining 71.0% of the test items has good level of discrimination. The result showed that most of the items of the 2017 NECO English language test items has good level of discrimination.

Discussion of Findings

The 2017 NECO English Language Multiple Choice test was unidimensional, based on the findings of the research question one on unidimensionality. Confirming unidimensionality supports the validity of the measurement instrument indicating that the items collectively measure a single construct which is crucial for ensuring that the instrument accurately reflects the intended latent traits. This corroborate those of Haberman and Sinharay (2010) which offer practical insights into the validity, reliability and interpretability of measurement instruments when unidimensionality is established. Jimoh (2021), who concluded that the 2016 NECO Mathematics Test was largely one-dimensional. Hambleton and Swaminathan (1985) also illustrate the importance of unidimensionality in measurement. In another study, Ubi, Joshua and Umoinyang (2012) took a random sample of candidates who sat for the joint Admission and UME in Cross River State, Nigeria, in the years 2002 and 2003 with the aim to determine the dimensionality of mathematics products concluded that examinations designed to select candidates may not be solely unidimensional, particularly when items are fielded from a large syllabus, based on the findings that the JAMB – UME test revealed five significant dimensions. This study implies that in order to increase students' overall test scores, it is critical to identify and develop this skill. It also emphasises the necessity of individualised teaching strategies to accommodate students' differing ability levels.

The research question two further revealed that 81.0% of the test items were of good difficulty while 19.0% of the test items were poor as their difficulty parameters were outside the range for which items difficulty parameter estimates are considered good i.e. -3 to 3 (Baker, 2001; Hambleton & Jones, De Mars, 2010). 71.0% of the test items were of good level of discrimination while 29.0% had discrimination values that were less than zero. Embretson and Reise (2000) emphasise that item difficulty can be visualised using item characteristic curves (ICCs), which show the probability of a correct response across different levels of ability. In our study, we utilised these ICCs to illustrate the difficulty of each item and to ensure that the items covered a wide range of ability levels.

Research question three revealed that 29.0% of multiple choice test items were poor as their discrimination indices were less than 0.4 the minimum discrimination parameter an item must have to be considered good (De Mars, 2010). The remaining 71.0% of the test items were of good level of discrimination. Based on these findings, the identification of the 19.0% of the multiple choice test items as poor in terms of difficulty parameters and 29.0% as poor in terms of discrimination indices raises concerns about the quality of these multiple choice items which suggests that a significant portion of the test may not effectively assess students' abilities or differentiate between high and low performers.

Conclusion

Based on the findings, several conclusions can be drawn regarding the 2017 NECO English Language multiple choice test. The test was found to be unidimensional, indicating that a specific skill or talent significantly has effect on students' performance which submits that focusing on developing the skill could lead to improved test outcomes. A substantial portion of the test items (19.0%) were identified as poor in terms of difficulty parameters, and 29.0% were poor in terms of discrimination indices. This raises concern about the quality of these items and propose a need for revision or replacement. There was variability in the specific skill among students, indicating that not all students possess the same level of proficiency. This highlights the importance of suitable teaching approaches to address varying skill levels.

Recommendations

The following suggestions are offered in response to study questions one, two and three, respectively, in light of these findings.

1. Skill identification and Development: Specific skill or talent that significantly influences students' performance on the test should be identify while targeted interventions and instructional strategies should be developed to enhance skill among students, aiming to improve the overall test performance. Teaching approaches that cater to the varying levels of the identified skill among students can also be implemented, additional support and resources for students who may need help in developing this skill should be provided.
2. Item writing guidelines and item selection process: To improve the quality of test items by ensuring they align with desired difficulty parameters, it is recommended to enhance the item selection process, emphasize adherence to guidelines to maintain item quality and consider using item banks that contain pre-tested items with known difficulty parameters which can help ensure that test items meet required standards for difficulty and reduce the risk of including poor-quality items in the test.
3. Item Analysis: Item analysis should be regularly conducted to monitor the discrimination indices of test items. Item with consistently low discrimination indices should be identified and removed or revised to maintain the overall quality of the test.

References

Adediwura, A. A. (2012). Teacher's perception of school-based assessment in Nigerian secondary schools. *Mediterranean Journal of Social Sciences 3*(1).

Afolayan, A. (1977). Acceptability of English as a second language in Nigeria. Acceptability in Language.

Baker, F. (2001). The basics of item response theory (2ed.): ERIC Clearing on Assessment and Evaluation.

Camili, G. (2006). Test fairness. I R.L Brenan (ed), educational measurement (4th ed.), Westport, CT: American Council on Education & Praeger. 4, pp 221-256.

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software.*

Demars, C. (2010). *Item response theory. Understanding statistics and measurement*. Oxford University Press.

Embretson, S. E., & Reise, S. P. (2000). Using item response theory to improve the psychometric properties of an instrument.

Faleye, B. A., & Olajide, A. A (2012). A revalidation of students' evaluation of teaching effectiveness rating scale. *IFE Psychology, 20.*

Flores, G. S. (2000). What is cultural validity in assessment? Retrieved from http://www.edgateway.net/cs/cvap/print/dcos/cvap/news.htm.

Hambleton, R. K. & Jones, R. V. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practices, 12*(3); 38 – 47.

Jimoh, K. (2021). Gender and culture-related differential item functioning in 2016 National Examinations Council Mathematics multiple choice questions in Nigeria. Unpublished Ph.D. Thesis, Obafemi Awolowo University, Ile-Ife.

Lam, T. C. M. (1995). Fairness in performance assessment, Eric Clearing house on counselling and student service Greensboro NC. Retrieved from http://www.ericfacility.net/ericdigest/ed391982.htm

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*, Malden, MA & Oxford: Blackwell.

Ojerinde, O., Popoola, K., Ojo, F. & Onyeneho, P. (2012). *Introduction to item response theory, parameter models, estimation and application*. Marvelouse Mike Press LTD

R Core Team (2019). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Robitzsch, A. (2019). Sirt: Supplementary item response theory models. R package version 3.7-40. https://CRAN.R-project.org/package=sirt

Wilberg, M. (2007). Differential item functioning analysis of high stake test in terms of gender. *Malaysian Online Journal of Educational Sciences*.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64*(2), 213-249. Doi:10.1007/BF02294536.