Sumerianz Journal of Education, Linguistics and Literature, 2020, Vol. 3, No. 11, pp. 249-253
ISSN(e): 2617-1201, ISSN(p): 2617-1732
Website: https://www.sumerianz.com
DOI: https://doi.org/10.47752/sjell.311.249.253

Original Article Open Access

Differential Test let Functioning (DTLF) in Senior School Certificate Mathematics Examination Using Multilevel Measurement Modelling

Emily Oluseyi Adeyemo*
Department of Educational Foundations and Counselling, Faculty of Education. Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria
Email: seyiadeyemo2007@yahoo.com

Oluwaseyi Aina Opesemowo
Department of Educational Foundations and Counselling, Faculty of Education. Obafemi Awolowo University, Ile-Ife, Osun State, Nigeria

Article History
Received: September 11, 2020
Revised: November 10, 2020
Accepted: November 12, 2020
Published: November 14, 2020

Abstract

The study determined the parameter estimate of the Senior School Certificate Mathematics items of June/July 2017 NECO examinations and testlet effect under Multilevel Measurement Modeling with the aim of providing information on the psychometric properties and quality of the items. The research design was an ex-post facto, The examinees response were the multiple – choice items of the National Examinations Council Mathematics paper two for June/July 2017 which consisted the data for the study group. The targeted population consisted of 26,086 senior secondary three examinees who registered for Mathematics Senior School Certificate (NECO) in June/July 2017 in Osun State. A total of 318 private schools and 179 public schools registered for the paper. The results revealed the following items to be good which implied that such items functioned well, these were: items 1, 4, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 20, 21, 22, 23, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 40, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 54, 57, 58, 59, 60 whereas items that were considered to be bad included 2, 3, 5, 6, 8, 13, 16, 19, 24, 25, 30, 38, 39, 41, 49, 53, 55 and 56. Furthermore, investigating an average bundle of item statistics under the measurement framework indicated that the Item Discrimination Means value and Standard Deviation under IRT approach were 1.26 and 0.60 respectively while the Mean value difference was 1.26. Although, item difficulty Mean value and the Standard Deviation were 0.26 and 4.26, respectively, whereas the Mean value difference was 0.26. Similarly, the guessing Mean value and the Standard Deviation were 0.15 and 0.19, respectively, whereas the Mean value difference stood at 0.15. The study concluded that any standardized examination, especially from an examining body in charge of certificate examination, if issues like differential testlet effect is not taking into consideration, it could harm the validity of the items and also alter the ability estimates of the examinees The validity of the test would be strengthened when issues like differential testlet effect is adequately taken care of.

Keywords: Differential testlet functioning; Multilevel measurement modeling; Testlet effect; Validity.

1. Introduction

Test fairness is a major concern in the credibility of items since this serves as part of what determines the validity of tests irrespective of how reliable the test could be. Studies have revealed that a set of items could be reliable but may not be valid since reliability is a function of consistency. As long as there is consistency even in what is seemingly wrong, it could have a high value of reliability. Adeyemo (2018) Fair assessment requires invariance of measure across different samples within population.

Everyone (male or female) of comparable abilities are expected to have equal probabilities of providing a correct response to any given item. In order to determine if a test is fair across groups within the targeted population, analysis of differential item functioning is commonly conducted. The effect of person grouping factor could be determined through the impact of Differential Item Functioning (DIF) analysis while the item grouping factor could be captured by analyzing the testlet effect.

One criteria of a good test is the absence of DIF. The items must function in the same way in all important subpopulations of the examinee. Ethnic background, item grouping factor such as common inputs, common response format and item chaining could affect item difficulty and hence threatens the validity of the items. There must be independent of observations as opined by Hox (2010), violation of this may lead to underestimation of standard error which in turns may result to spurious rejection of the null hypothesis.

Item measure may be affected by person grouping such as gender, Local item and ethnic background among other factors such as common input or stimulus, common response format and item chaining. In either case, item difficulty is affected by a factor irrelevant to the main construct, hence construct validity of the test consisted of the items is threatened. The effect of person grouping factors can be studied through impact and differential item functioning (DIF) analysis while the effect of the item grouping can be captured by studying testlet effect. Differential models have been proposed to study DIF; for examples Logistic regression, Mantel Hanzel, Item

*Corresponding Author

response model and multiple indicator multiple causes (MIMC) models, however when Local item Independence is violated, DIF detection may be affected according to Bolt (2002).

Testlet is defined as a set of items that are constructed and implemented together as a unit of measurement (Wainer and Kiely, 1987). It is a fungible unit of a test, an interrelated and integrated group of items always presented as a single unit. For example, in a reading comprehension test, a series of question may be based upon a common reading passage. Items responses on questions within a testltet may not be totally independent of each other because the level of understanding of the reading passage may be affected by the students' knowledge or interest in the contents of reading passage. For example, when item responses on a question is not totally independent on each other, the level of understanding of a student in one item may affect the performance on the other dependent item. If an examinee need to derive an answer to a question based on the answer to a previous question, the students who missed the preceding question will also be at disadvantage to answer succeeding question correctly, this renders the test unfair since test fairness is a major concern in standardized testing when establishing the validity of a test score. If a test were built to be administered as a unit, it is important that the items be analysed that way, if not, one is likely to get a wrong answer.

Standardized tests are often composed of testlets especially the reading comprehension tests where sets of items are associated with the passages. In situation like this, according to Wainer and Kiely (1987), Item Response Theory may not be independent of each other and hence parameter estimates in IRT models may be biased especially for item discrimination parameters. Also DIF magnitude and item parameters may not be estimated accurately under the violation of the item local independence assumption according to Fukuhara (2009), ignoring local item dependence (LID) would result in overestimation of precision of the ability estimates.

Determination of DIF at the testlet level accordindg to Wainer *et al.* (1991) has three disadvantages over confining the investigation of the items. It allows the analysis model to match the test construction, DIF cancellation through balancing and the uncovering of DIF because of its size evades detection at the item level but can become visible with some aggregation. Roznowski and Reith (1999) in his own pointed out that because decisions are made at the scale or test level, DIF at the item level may have only limited importance. It is sensible therefore to consider an aggregate measure of DIF. The present study aimed at analyzing testlet effect and impact of differential testlet functioning on Senior Secondary School Certificate Mathematics paper two of National Examination Council (NECO) of 2017 in order to identify the status of the psychometric properties of the items and also determine testlet effect under Multilevel Measurement Modeling in order to make recommendations on the quality of the test items. The objectives of the study were to determine the parameter estimate of the Senior School Certificate Mathematics items of June/July 2017 NECO examinations and to determine testlet effect under Multilevel Measurement Modelling with the aim of providing information on the psychometric properties and quality of NECO items.

2. Methods

The research design was an ex-post facto, a causal comparative research design since the research was conducted after variation in the independent variable had been determined in the natural course of events. The examinees response were the multiple – choice items of the National Examinations Council Mathematics paper two for June/July 2017 which consisted the data for the study group. In as much as the design was an ex-post facto, there was no room for manipulation of data obtained. The targeted population consisted of 26,086 senior secondary three examinees who registered for Mathematics Senior School Certificate (NECO) in June/July 2017 in Osun State. As part of this population was 13,120 male and 12,966 female. A total of 318 private schools and 179 public schools registered for the paper. The location included urban and rural, while the urban area accounted for 15,048 (57.7%) the rural was 11038 (42.3%).

Table-1. One showing Examinee's Characteristics

Variables		Frequency	Percentages	Total
	Male	13,120	50.5	26.086
	Female	12,966	49.52	
School Ownership	Public	179	36.02	497
	Private	318	63.98	
Location	Urban	15,048	57.7	26.086
	Rural	11.038	42.3	

Source; NECO 2017

A sample of 14,936 Senior Secondary School three (SSS 3) examinee was selected purposively on the basis of those who completed all the 60 items of the Mathematics paper two. This sample size consisted of 7,272 (48.7%) male while the female consisted of 7664 (51.3%) from the three senatorial districts of the state. The instrument for the study was the June/July 2017 NECO Mathematics paper two examination which was a dichotomous multiple choice examination consisting 60 items with 5-option key of four distracters. These were to be attempted for one hour and forty-five minutes. The Optimal Mark Recorder (OMR) sheets containing the response of these candidates were used as the data which were collected from the NECO head office. The demographic data of each examinee such as centre number, candidate number and sex were printed on the OMR sheets to ensure proper coding for computer analysis. Data were analyzed under Multi-level Measurement Modeling (MMMT-2 and MMMT-3).

The notable difference in the conceptualization of testlet effect in MMMT-3 and MMMT-2 as pointed by Walker (2012) is that the MMMT-3, testlet effect is person specific i.e. it contributes to person ability but in MMMT-2, it is item-specific i.e. it contributes to the difficulty of the items in the respective testlets. One of the

benefits of MMMT-2 is that it can simultaneously test for impact, DIF and DTLF. This study used the MMMT-3 testlet response model to assess the impact and the differential testlet functioning (DTLF) in the NECO Senior Secondary Mathematics paper two items.

3. Results

Research Question One: What are parameters estimate (such as difficulty, discrimination, and guessing) of the National Examination Council (NECO) Mathematics in 2017.

To answer this question, the examinees' responses in the 2017 NECO Mathematics items were calibrated using a 3-parameter logistic IRT model which align with the preliminary analysis indicating that the data was multidimensional. The results were presented in Table two, which showed the dificulty, discrimination and guessing parameters of each item of the 2017 NECO Mathematics items under Multi-level Measurements Modelling.

Table-2. NECO Mathematics Items Discrimination, Difficulty and Guessing Indices using Item Response Theory

Items	***A***	***b***	***C***	**Remark**
1	0.91	-1.94	0.07	**
2	0.84	0.96	0.4	*
3	0.72	-1.08	0.47	*
4	1.6	-0.84	0.13	**
5	0.21	-3.21	0.2	*
6	1.01	-0.19	0.69	*
7	1.1	-1.2	0.02	**
8	1.88	-0.14	0.58	*
9	1.33	-0.98	0.01	**
10	1.49	-0.12	0.3	**
11	1.78	-0.63	0.01	**
12	1.63	-0.16	0.04	**
13	0.82	21.61	0.08	*
14	0.98	-0.71	0.01	**
15	1.67	-0.24	0.06	**
16	1.61	-0.27	0.38	**
17	1.57	-0.56	0.04	**
18	0.24	1.59	0.04	**
19	0.79	-2.57	0.5	*
20	1.16	-1.54	0.02	**
21	0.74	0.27	0.02	**
22	1.35	-1.32	0.01	**
23	0.4	-0.26	0.03	**
24	1.12	-0.58	0.35	*
25	0.82	26.94	0.07	*
26	0.82	-1.74	0.01	**
27	0.61	-0.73	0.02	**
28	1.15	0.9	0.07	**
29	0.36	2.2	0.03	**
30	0.96	-0.53	0.45	*
31	1.97	-0.35	0.07	**
32	1.53	-1.01	0.01	**
33	1.35	-1.03	0.01	**
34	1.84	-0.77	0.04	**
35	1.39	-1.26	0.01	**
36	1.29	-0.57	0.03	**
37	2.72	0.13	0.12	**
38	2.95	0.24	0.45	*
39	1.89	-0.24	0.42	*
40	1.56	-0.34	0.02	**
41	1.51	0.23	0.44	*
42	2.37	0.06	0.14	**
43	1.89	-0.53	0.09	**
44	0.53	-1.39	0.03	**
45	1.3	-1.27	0.03	**
46	1.96	-0.79	0.08	**
47	0.52	-0.19	0.01	**
48	2	-0.2	0.32	**
49	1.06	-0.79	0.46	*

50	1.51	-0.12	0.17	**
51	0.97	-2.3	0.04	**
52	1.45	-1.92	0.02	**
53	2.22	-0.14	0.44	*
54	1.33	-1.17	0.04	**
55	0.16	1.56	0.06	*
56	1.46	-0.1	0.43	*
57	0.99	-1.53	0.03	**
58	1.25	-1.23	0.03	**
59	0.93	-0.76	0.01	**
60	0.21	0.36	0.08	**

Note: (*) represents a bad item and (**) represents a good item

It was observed from Table two how well the items function in terms of difficulty, discrimination and guessing. It ccould be seen that items with a double asterisk (**) under the remark column functioned well among examinees while items with a single asterisk (*) functioned poorly or badly. Based on these criteria, 42 items representing 70% out of the 60 NECO Mathematics items of 2017 functioned well while the remaining 18 items representing 30% function badly. The following items were said to be the good items which implied that such item functioned well, these were: items 1, 4, 7, 8, 9, 10, 11, 12, 14, 15, 17, 18, 20, 21, 22, 23, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 40, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52, 54, 57, 58, 59, 60 whereas items that were considered to be bad included 2, 3, 5, 6, 8, 13, 16, 19, 24, 25, 30, 38, 39, 41, 49, 53, 55 and 56. Furthermore, investigating an average bundle of item statistics under the measurement framework indicated that the Item Discrimination Means value and Standard Deviation under IRT approach were 1.26 and 0.60 respectively while the Mean value difference was 1.26. Although, item difficulty Mean value and the Standard Deviation were 0.26 and 4.26, respectively,whereas the Mean value difference was 0.26. Similarly, the guessing Mean value and the Standard Deviation were 0.15 and 0.19, respectively, whereas the Mean value difference stood at 0.15.

Research Objective Two:- Determine testlet effect under Multi-level Measurement Modelling of NECO Senior School Certificate Mathematics items of 2017.

Table-3. Testlet Statistics for NECO SSCE Mathematics Items of 2017

Testlet	Number of Items	Testlet Variance	Standard Deviation
MMMT-3		0.25	0.02
Testlet 1	7	0.25	0.01
Testlet 2	6	0.12	0.01
Testlet 3	22	0.84	0.02
Testlet4	16	0.59	0.02
Testlet 5	5	0.14	0.01
Testlet 6	4	0.11	0.01

To achieve this objective, items were subjected to MMMT-2 and MMMT-3. But under the MMMT-3, the variance (testlet effect) were considered to be fixed across all the testlet effects using Statistical Analysis System (SAS) and a different testlet effect for each testlet using the MMMT-2. From Table 3, the testlet effect variance showed the extent of Local Item Dependence (LID) among the items associated with a given testlet. The testlet effect variance will be zero when there is presence of Local Item Independence (LII). When there is an increase in the variance, the more the testlet items are locally interrelated or dependent. There is no generally acceptable decision or criteria for justifying the variance of the testlet effect but there had been some researchers who had come up with some certain criteria for the testlet effect variance. Glass, Wainer *et al.* (2000), Wang and Wilson (2005), submitted that a variance lower than 0.25 should be regarded as negligible small. Likewise, Wang and Wilson (2005), agreed that a variance is substantial when testlet effect variance ranges from 0s.5 to 2. Based on the aforementioned criteria, it was not awry to say that Testlets 1, 2, 5 and 6 showed a negligible testlet effect variance and the variances for Testlets 3 and 4 were considered substantial.

4. Discussion

Following the analysis, there had been an in-depth insight into differential testlet functioning of the NECO Senior School Certificate Mathematics Examination. Based on the preliminary analysis of of the study, it has been revealed that the data were multidimensional, hence in the submission of Carolyn *et al.* (2009), the application of unidimensionality IRT model with multidimensional data contradicts and violates the assumption of unidimensionality which infer threats on item parameters. The results revealed that the 2017 NECO Mathematics items were multidimensional an evidence of the fact that more than one construct were being measured, this therefore violated the IRT assumption of unidimensionality as opined by Thissen *et al.* (1993), and Wainer *et al.* (1991). This supported the findings of Abiri (2006) that difficulty indices of multiple- choice items test with a fewer number of options like four is better than anyone with more significant number of options like (the NECO five options). From the results of the ability estimate, the standard error (SE) were in line with those of previous studies like Tuerlinckx and De Boeck (2001), Sireci *et al.* (1991) which shown that ignoring local iten differential(LID) or

testlet variance could result into biased estimation of ability parameters at both low and high ends of ability distribution which could lead to inflated item difficulty estimates.

5. Conclusion

In any standardized examination, especially from an examining body in charge of certificate examination like NECO, if issues like differential testlet effect is not taking into consideration, it could harm the validity of the items and also alter the ability estimates of the examinees The validity of the test would be strengthened when issues like differential testlet effect is adequately taken care of.

References

Abiri, J. O. (2006). Comparative analysis of psychometric properties of Mathematics items constructed by WAEC and NECO in Nigeria. *Academic Journal of Evaluation, Measurement and Statistics, Techniques in Education, Ilorin.*

Adeyemo, E. O. (2018). Influence of situationally-induced response-faking on validity and reliability of a personality inventory. *ASSEREN Journal of Education,* 3(1): 125-35.

Bolt, D. M. (2002). A Monte Carlo Comparison of Parametric and Nonparametric polytomous DIF detection Methods. *Applied Measurement in Education,* 15(2): 113-41.

Carolyn, F., Furlow, C. F., Ross, T. R. and Gane, P. (2009). The impact of multidimemensionality on the detection of differential bundle functioning using simultaneous item bias test. *Applied Psychological Measurement,* 33(60441464): Available: http:/doi/10.1177/01466211609331959

Fukuhara, H. (2009). *A differential item functioning model for testlet-based items response model: A bayesian approach. Doctoral theses.* Florida State University.

Hox, J. J. (2010). *Multilevel Analysis :Techniques and applications.*Routledge: New York.

Roznowski, M. and Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items. Do biased item result in poor measurement? *Educational and Psychological Measurement,* 59(20): 248-70.

Sireci, S. G., Thissen, D. and Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement,* 28(3): 237-47.

Thissen, D., Steinberg, L. and Wainer, H. (1993). *Detection of differential item functioning using the parameters of item response models. In P. W. Holland and H. Wainer (Eds.), Differential item functioning.* Erlbaum: Hillsdale, NJ. 67-113.

Tuerlinckx, F. and De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods,* 6(2): 181-95.

Wainer, H. and Kiely, G. L. (1987). Item clusters and computerized adaptive testing. A case for testlets. *Journal of Educational Measurement,* 24(3): 185-201.

Wainer, H., Sireci, S. G. and Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement,* 28(3): 197-219.

Wainer, H., Bradlow, E. T. and Du, Z. (2000). *Testlet response theory: An analog for the 3PL useful in adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), Computerized adaptive testing: Theory and practice*. MA: Kluwer: Boston. 245-70.

Walker, C. M. (2012). Establishing effect size guidelines for interpreting the results of differential bundle functioning analyses using SIBTEST. *Educational and Psychological Measurement,* 72(3): 415-34.

Wang, W. C. and Wilson, M. (2005). Assessment of differential item functioning in testlet base items using the rasch testlet model. *Educationaland Psychological Measurement,* 65: 549-76. Available: http://Doi:10.1177/0013164404268677