CONSTRUCTION AND VALIDATION OF POLYTOMOUSLY-SCORED MULTIPLE-CHOICE ITEMS IN MATHEMATICS

R. O. IWINTOLU, E. R. I. AFOLABI

Obafemi Awolowo University, NIGERIA

Abstract. This study constructed and validated multiple-choice items using dichotomous and polytomous scoring in Mathematics. It examined the difficulty and discrimination indices of polytomous items using dichotomous scoring; the thresholds and discrimination indices of the items using polytomous scoring (GRM) and compared the discrimination indices of the items in both scoring formats. The population of the study comprised secondary schools' students in Lagos State, Nigeria. The sample consisted of 1015 students; these were selected from two Education Districts (EDs) from the six EDs in the state using simple random sampling technique. The results revealed that more than half of the items had high difficulty indices with corresponding high discrimination indices, using dichotomous scoring. It showed the items had moderate transition locations and discrimination indices, using polytomous scoring. It also indicated that there was a significant difference in the discrimination indices of both scoring methods.

Keywords: dichotomous scoring, polytomous scoring, difficulty index, discrimination index, transition locations

Introduction

The inevitability of tests in fulfilling important needs in the decision-making process permeating all facets of human endeavour seems to have made it captured the attention of the public, and at the same time inspires so much anxiety from classroom to work environment. Even the least interested citizen is aware of the growing use of testing in every facet of human endeavour; it is therefore expected that tests attained prominence and usage as far back as 3000 years ago. Educators, governments and establishments' continuous search for reliable and trusted means through which students and applicants could be held accountable of a self-acclaimed ability resulted in testing. Although, several schools of thought have argued that test or examination is not the best measure of a person's ability, while on the other hand, researchers (Lumsden, 1978; Kaplan & Saccuzzo, 2005, Afolabi, 2012) have maintained that it is everywhere evident that there is yet no viable alternative to tests in determining a person's academic ability.

A test may be defined as standard set of items which are specific stimuli to which a person overtly responds and which can be scored. There are different types of tests, with different peculiarities, strengths and weaknesses. The major and broad classification of test types is the essay and objective tests. From a practical point of view, the essay (free-response) test provides good measure of students' ability, in so far as it gives examinees the opportunity to respond freely to items in their own words, thereby building their reasoning and other skills in higher-thinking learning. The required response can be as simple as the writing of a single word or as complex as the design of a laboratory experiment to test a scientific hypothesis. It helps to recognise and reward different abilities of students through the assigning of partial credits to incomplete understanding of concepts. Despite the enormous credits accrued to the essay tests however, it has been critiqued on the basis of its difficulty to use; its scoring could be subject to raters' bias and inconsistency. On the other hand, the advocates of the objective tests type among which is the multiple-choice format have adduced reasons for its dominant use which includes; limited amount of testing time, its ability to sample a broad range of content and provide a good sample of test takers' knowledge. The responses can be scored by machine, making the scoring process faster and inexpensive, with no room for differences of opinion. In other words, objective tests allow the evaluation of a greater breadth of content in a fixed testing time under limited financial budgets.

Multiple-choice test (MCT), a variant of the objective test has gained credence from classroom assessment to professional licensure examinations (Scott, 2011). It is pertinent to note that the use to which MCT will be put, would determine the structure of the items in terms of construction of the stem, the response options (correct answer(s)) and the distracters which are predicated on its scoring. For the vast majority of multiple-choice tests, items are scored dichotomously (i.e., correct or incorrect). According to Osterlind & Everson (2009), items with two categories or values (possibilities to respond) are called dichotomous. In this stance, it involves presenting test question and a list of alternatives ranging from three, to four or five as the case may be. The testees are to make free choice of one correct answer from the alternatives given to the item. If an examinee selects the correct answer, in dichotomous scoring, he/she will be awarded a score of one, while if an incorrect answer a score of zero.

Generally, most users of the multiple-choice tests employ dichotomous scoring. However, educational reform efforts have led to an increased search of alternatives to the traditional dichotomously scored multiple-choice items as there has long been a need to assess objectives that require more than a single response (Albanese, 1993). Researchers in a bid to midmost the two item types (i.e., essay and objective test), explored the combination of the duo item responses and established that they produced a total score that is more reliable than scores separated by item type. Despite the alternative mixed-item format, they are still subject to both item types' strengths and weaknesses. Nonetheless, polytomous tests have been observed to provide more equitable approaches to testing than the dichotomously scored multiple-choice tests. They have also been adjudged to provide more information regarding the precision of trait-level estimation than dichotomous items (Jodoin, 2003; Penfield & Bergeron, 2005). The polytomous item type could therefore be perceived as the mediator between the dichotomously scored multiple-choice item and the essay test type; as it is contingent on the combined strengths of both item types in a bid to shrink their weaknesses.

In polytomous models, items in the test are not just scored right or wrong; instead, each of the categories of responses is evaluated and scored according to its degree of correctness or the amount of information provided toward the full answer. By implication, the items are constructed in such a way that it allows every examinee's efforts at items to be rewarded. Weights are assigned to options in ascending order as examinees knowledge on the item increases, in other words, polytomous items are believed to increase test validity. An advantage of the use of polytomous response items over dichotomous items appears to be the increased test information, on both the test-taker and each of the items in the test, which is one tenets of Item Response Theory (IRT).

It may not be an overstatement to say that the complexities involved in the construction of polytomous items could have account for one of the reasons teachers, test developers and examination bodies have not embraced its use, even where they are in use in advanced countries, most of them were in Mathematics. This may possibly be due to the role the subject plays and its distinctive contribution to the objectives of general education of man than any other subject (Odeyemi, 1991) and its flexibility in accommodating the peculiarity of the structure of the response options which may not be practicable in some other subjects.

Multiple choice test items using dichotomous scoring have gained considerable popularity among constructors of classroom and standard test. It is also generally known as the most widely applicable and useful type of objective test item. Unfortunately, Wilson & Masters (1993) referred to the practice of restricting each item to a single correct answer as a stifling limitation, noting that multiple-choice items with more than one correct option may have greater latitude for accuracy of ability estimates. In other words, the dichotomously scored items attempts to put a ceiling on low-ability students from being rewarded for their effort and will at items. This perhaps imply the over domineering of high-ability students and the underestimation of low-ability students. As a result, there is a need to come-up with items that could reduce errors, thereby catering for all ability groups.

The objectives of this study are to: examine the difficulty and discrimination indices of polytomous items using dichotomous scoring; determine the thresholds and discrimination indices of the items using polytomous scoring (GRM) and compare the discrimination indices of the items in both scoring formats. The research questions are: (1) what is the level of item difficulty and discrimination of the polytomous items using dichotomous scoring; (2) what are transition locations and the level of item difficulty of the items using the GRM and (3) are there differences in the discrimination indices of the items in both scoring methods.

Theoretical framework

Approaches for modeling responses to multiple-choice items fall into two broad categories: (a) those that group all distractor options into a single incorrect response category and model the probability of correct response using a dichotomous response model, and (b) those that retain the distinction between all response options and model the probability of each response option using a polytomous response model. However, the development and scoring in this study is predicated on the latter category, which deliberately conceive each response option as learning progressions in order to facilitate diagnostic assessment of student understanding. In this context, diagnostic assessment hinges upon the development of items (i.e., tasks, problems) to efficiently elicit student conceptions that can be related back to a hypothesized learning progression.

The advantage of using a polytomous model for multiple-choice items stems from the potential of extracting information from the distractors as well as the correct response. The incorporation of information pertaining to each of the distractors maximizes the information concerning the latent trait, and thus has the potential to lead to more precise estimation of the latent trait than the dichotomous response models, particularly at the lower end of the latent trait continuum (Bock, 1972; De Ayala, 1989; 1992). Consequently, Briggs et al. (2006) introduced Ordered Multiple-Choice (OMC) items as a means to this end. OMC items represent an attempt to combine the efficiency of traditional multiple-choice items with the qualitative richness of responses to open-ended questions.

The potential effectiveness comes because OMC items feature a constrained set of response options that can be scored objectively; the prospective qualitative richness comes because OMC response options are both designed to correspond to what students might answer in response to an open-ended question and explicitly linked to a discrete level of an underlying learning progression. The OMC item format belongs to a broader class of constrained assessment items in which the interest is not solely in whether a student has chosen the "scientifically correct" answer, but on diagnosing the reasons behind a student's choice of a less scientifically correct answer (c.f., Minstrell, 1992; 2000). An appealing aspect of such items is that they are consistent with the spirit behind learning progressions, which at root represent an attempt to classify the gray area of cognition that muddiest he notion that students either "get something" or they do not.

When items are scored polytomously, the categories of each multiple-choice item can represent different levels of difficulty which are referred to as thresholds. Thresholds refer only to a local relationship between a pair of adjacent categories, and are characterized as steps. The threshold between two adjacent categories is the ability level at which an examinee has equal probability to choose either one of two categories. The difficulty of reaching each threshold is identified not as the difficulty of reaching that threshold in relation to all other thresholds but only in relation to the previous threshold. Since the description of performance is referenced directly to the items making up the test, it is common practice to refer to the scale as an achievement scale. This reflects the fact that it is made up of observable behaviours - the students' achievements on the test. The underlying ability scale then takes its meaning from the descriptions and locations of these achievements. The process of establishing the achievement scale may involve several stages. Initially, at the item construction stage, there is an intention to construct items with a range of difficulties. Then the relative difficulty of each item is estimated from the data. Each item is then located on the scale. Following this, a description of the knowledge and skills addressed by each item is referenced to the item location, so that a profile of the performance demands of the test can be developed.

When polytomous models are applied to multiple-choice items, and partial credits are given to categories other than the best answers, the configuration of the thresholds of an item affects the information function of the item and thus the precision of ability estimation. Configuration of the thresholds includes two aspects, namely the order of the thresholds and the distances between them. Dodd & Koch (1985) found that item information functions for the partial credit model differs as a function of the thresholds. The distance between the first and last thresholds affected the shape of the information function of an item. Items with shorter distances between first and last thresholds had a more peaked information function for a narrower range of ability continuum.

For thresholds which are naturally ordered. If for example, the first threshold is located at -1.22 logits, the second at 0.23 logits. Persons with ability estimates less than -1.22 logits are most likely to fail the first threshold, and so score 0 on the item. Persons with ability estimates in the range -1.22 logits to 0.23 logits are most likely to pass the first threshold but fail the second, and so score 1. Persons with ability estimates greater than 0.23 logits are most likely to pass the first and the second thresholds, and so score 2. With increasing ability, the probability of exceeding the first threshold and then the second threshold also increases. Conversely, the greater the score, the greater the latent ability that is implied by the model. This is the underlying principle of the ordered categories and therefore the thresholds serve as boundaries between the categories and reflect an increasing amount of the attribute being measured (Andrich, 2002). In the case of ordered thresholds, for example: the candidate is required to simplify an equation.

Example 1: Simplify: $102/5 – 6\ \ 2/3 + 3$.

$52/5 – \ 20/3 + 3$ ……………………….................... 1 T

$\frac{156 – 100 + 45}{15} = \frac{101}{15} = 6\ 11/15$ ………............. 2 T

1T represent the first threshold; 2T represent the second threshold

To simplify the equation, the candidate must have the knowledge of converting a mixed fraction to an improper fraction; and then the simplification of the improper fraction. Here, the configuration of the threshold was ordered. The conversion of the equation from mixed fraction to an improper fraction is simpler than the second task of the simplification itself. This is because the final solution cannot be arrived at without first passing through the first task. Therefore, the first task could represent a less difficult threshold (first threshold) while the later task, which is the higher and require more knowledge to be able to solve the problem completely could amount to the second threshold.

Consequently, a candidate whose ability lies below the knowledge required in the first threshold and who ticks the incorrect response would be scored 0. It is logical to think that any candidate whose ability falls within the first threshold will be able to solve the first threshold; the candidate is awarded a score of 1. While a candidate whose ability falls within and above the first threshold will be able to solve both the first and the second threshold, thereby arriving at the full answer; the second threshold is awarded a score of 2.

Disordered thresholds are not taken to mean that there is a problem with the way an item is functioning, but only that the item is displayed in sequential order like the ordered items. The problem of using such items to construct the achievement scale is resolved by calculating an alternative set of thresholds, called Thurstone thresholds, based on the Thurstone Cumulative Probability Model. The thresholds derived from the Thurstone model are always in a natural order, irrespective of the order of the thresholds produced by the Rasch model. The first is that there is no evidence here to support the view that thresholds can be interpreted as steps, which refer only to a local relationship between categories. Interpreting thresholds as steps suggests that, as it is not unreasonable for later steps to be of lesser magnitude than earlier steps, it is also not unreasonable for later thresholds to have lower values than earlier thresholds.

Example 2: the coordinates of points *P* and *Q* are (4, 3) and (2, -1) respectively. Find the shortest distance between *P and Q.*

Distance

$\sqrt{(x1 - x2)2 + (y1 - y2)2}$ ………………………………………… *1 T*

$\sqrt{(4 - 2)2 + (3 - (-1))2}$

$\sqrt{(2)2 + (4)2}$

$\sqrt{4 + 16}$

$\sqrt{20}$

$2\sqrt{5}$……………………………………………………………………… *2 T*

1T represent the first threshold; 2T represent the second threshold

The item in Example 2 requires the candidate to find the distance between two points. In solving this problem, a candidate needs to know and be able to first put down the formula for finding distance between two points, then substitute and simplify the equation emanating from the problem. Here, the formula plays the most important role, without which the problem cannot be solved at all. After the equation, then the coordinates are substituted and simplified. In other words, the candidate who is able to correctly put down the formula for calculating distance between two points earns a higher score, than one who is only able to solve or simplify an equation. Therefore, the item in this stance is disordered. This is evident in the fact that the first threshold involves a higher ability than the second threshold, thereby not giving persons with lower ability the opportunity to be rewarded for the little effort they could have put in solving the problem. This could be referred to as ability ceiling, where lower ability candidates do not have a chance of scoring to the limit of their ability because the item is disordered.

Method

The study adopted the survey research design. The population of the study comprised secondary schools' students in Lagos State, Nigeria that registered for the Senior School Certificate Examination (SSCE) in the state. The students in this category have gone through the Junior Secondary School (JSS) and Senior Secondary School (SSS) in an accumulation of six years; have offered Mathematics as a subject for those numbers of years and had registered same in the SSCE for 2015. The sample consisted of 1015 students; these were selected from two Education Districts (EDs) from the six EDs in the state using simple random sampling technique based on availability of federal schools for an inclusive representation of schools; while three schools were selected from each of the EDs. One intact SS III class from each of the schools was selected using simple random sampling technique. The research instrument for the study is titled "Mathematics Achievement Test" (MAT). It is an adapted version of the June/July (2006 - 2014) SSCE General Mathematics Paper 1. The MAT is a 20-item multiple-choice test composed of polytomous responses. It comprised a list of items with 4 response options each. The options comprised (1 partially correct answer, 1 full correct answer and 2 distracters). Data collected were analyzed by using BILOG and IRTPRO.

Analyses

First, preliminary analyses were carried out on the data; first the descriptive analysis of MAT was carried out to describe the frequency of each item response option as well as the mean, maximum, minimum, and standard deviation of each item. The unidimensionality of the test which is in line with the assumptions of IRT was established. According to Nandakumar & Stout, 1993, the assumption of unidimensionality implies that a test measures a single ability and that the responses obey the principle of local independence, which states that item responses are independently conditioned on a particular level of ability. However, unidimensionality can be established when one of two conditions is met from the results of an exploratory factor analysis: first, a factor analysis on the inter-item correlation matrix should show that the first factor accounts for at least 20% of the variance of the unrotated factor matrix or second the eigen values of the first factor should clearly exceed that of the second factor.

The method used to assess unidimensionality in this study was confirmatory factor analysis. Moreover, a scree plot was produced to determine whether uni-dimensionality could be inferred. Scree plots provide a convenient way of visualising a dominant factor in principal component analysis. A dominant factor is evidenced in a scree plot when there is a factor that distinct itself above the “elbow break” of the figure.

In answering the research questions, the item difficulty parameter estimates were examined. Test items with high b-values are normally hard items under IRT model; these are the items that low-ability examinees are unlikely to answer correctly. But items with low b-values are classified as easy items; these are items that most examinees including the low ability will have at least a moderate chance of answering correctly. Therefore, researchers refer to items with b-values greater than 1.0 as difficult items. Furthermore, the discrimination value expresses how well an item can differentiate among examinees with various ability levels. Good items usually have discrimination values ranging from 0.5 to 2.0. In respect of guessing however, in the 3PL model, item discrimination is proportional to the slope of the IRF at the point of inflexion and is equal to 0.25. The parameter c has a theoretical range of 0<=C<=1.0, but in practice, values above 0.35 are not considered acceptable (Baker, 2001; Adedoyin & Adedoyin, 2013). Also, in comparing the discrimination indices of the two methods, paired sample t-test was used.

Results

The mean, the standard deviation, minimum, maximum scores and the frequency of the response options for MAT items are displayed in Table 1.

Table 1. Mean, standard deviation, minimum, maximum Scores and the frequency of the response options for MAT items

	Statistical Properties				Frequency of Response Options		
	M	SD	Min	Max	0	1	2
Item 1	1.40	0.89	0	2	187	31	697
Item 2	0.84	0.83	0	2	442	290	283
Item 3	0.83	0.84	0	2	461	261	293
Item 4	1.01	0.80	0	2	324	354	337
Item 5	1.20	0.88	0	2	309	193	513
Item 6	0.79	0.84	0	2	488	252	275
Item 7	1.09	0.85	0	2	324	268	423
Item 8	0.94	0.87	0	2	417	239	359
Item 9	0.71	0.87	0	2	572	161	282
Item 10	0.77	0.81	0	2	474	293	248
Item 11	1.41	0.82	0	2	222	154	639
Item 12	1.89	0.81	0	2	256	312	447
Item 13	0.89	0.89	0	2	469	183	363
Item 14	0.72	0.92	0	2	582	95	338
Item 15	1.54	0.78	0	2	188	90	737
Item 16	1.07	0.83	0	2	319	302	394
Item 17	1.03	0.90	0	2	400	183	432
Item 18	0.89	0.81	0	2	398	327	290
Item 19	0.72	0.89	0	2	587	123	305
Item 20	0.68	0.83	0	2	559	212	244

To establish the unidimensionality of the test, eigenvalues and total variance explained is presented in Table 2.

Table 2. Eigenvalues and total variance explained of MAT

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.248	16.242	16.242	3.248	16.242	16.242
2	1.287	6.433	22.675	1.287	6.433	22.675
3	1.222	6.112	28.787	1.222	6.112	28.787
4	1.171	5.856	34.643	1.171	5.856	34.643
5	1.048	5.240	39.883	1.048	5.240	39.883

Table 2 showed the factor analysis carried out on the 20 Mathematics test items. It yielded five eigen values greater than one. The first eigen value was 16.242 which were clearly greater than the next eigen value of 6.433, indicating the unidimensionality of the data. The unidimensionality of the data was further confirmed by a scree plot presented in Fig. 1.

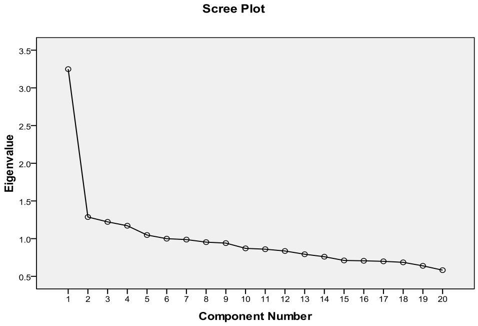

Figure 1. Scree plot of MAT

Research question 1: what is the level of item difficulty of the polytomous items using dichotomous scoring

To answer this research question, the set of 20 polytomous items developed (0, 1, 2) were scored dichotomously. These items were initially developed to have two distractors, one partially correct answer and one full credit answer was intentionally scored as dichotomous in order to examine their item parameters (difficulty and discrimination). The results are presented in Table 3.

Table 3. Item parameters of polytomous items (0, 1, 2) using dichotomous scoring

S/N	Difficulty (α)	Discrimination (λ)
1	-0.73	0.38
2	1.26	1.22
3	1.95	0.62
4	1.09	0.86
5	0.34	0.76
6	1.35	0.72
7	1.23	0.84
8	0.83	1.13
9	1.09	1.72
10	1.67	1.06
11	0.01	1.32
12	1.35	1.09
13	0.92	0.85
14	1.02	1.03
15	-0.59	0.95
16	1.82	0.42
17	0.70	1.36
18	1.64	1.67
19	1.32	1.16
20	0.00	0.52

The item difficulty parameter estimates in the table (α -values) ranged from -0.73 for item 1 to 1.95 for item 3. Evidently, some items could be classified moderately difficult with α -values of 0.83 (item 8), 0.92 (item 13). Out of the twenty (20) items, twelve (12) items were considered difficult; while their discrimination indices (λ) ranged from 0.38 for item 1 to 1.72 for item 9. From the table above, only two (1 and 16) out of twenty (20) items were not able to discriminate among the examinees, since their discrimination indices were lower than 0.5.

Research question 2: what are the transition locations and discrimination indices of the items scored using polytomous scoring

` The research question was examined by subjecting the polytomous items to analysis using the Graded Response Model (GRM). The GRM produced the item parameters of the items in terms of discrimination and transition locations. The transition locations represent the difficulty of the thresholds. The first transition (b1) indicates the difficulty of the lowest threshold i.e between the score of 0 (inability to get the answer correctly) and 1 (ability to get the lowest threshold correctly). The second threshold (b2) on the other hand indicates the difficulty of a higher threshold and the lowest threshold i.e score of 1 and 2. These are presented in Table 4.

Table 4 revealed that six items were considered to had a low difficulty level using the criteria stated earlier in the study; these were (items 1, 5, 11, 12, 15, and 17). Also, five items were classified as moderate (items 4, 7, 8, 13, and 14); while nine items had high difficulty levels (items 2, 3, 6, 9, 10, 16, 18, 19, and 20). The table showed that items 1, 3, 16 and 20 had the lowest discrimination values of 0.35, 0.40, 0.16 and 0.08 respectively.

Table 4. Item parameters and transition locations of GRM (0, 1, 2)

SN	λ	b1	b2	Overall diff
1	0.35	-2.70	-2.28	-2.50
2	1.16	-0.29	1.02	0.37
3	0.40	-0.53	2.31	0.89
4	0.87	-1.03	0.90	-0.06
5	0.84	-1.15	-0.07	-0.61
6	0.59	-0.19	1.75	0.78
7	0.59	-1.42	0.58	-0.42
8	1.01	-0.48	0.66	0.09
9	1.14	0.23	1.00	0.62
10	0.59	-0.29	2.02	0.80
11	1.19	-1.36	-0.60	-0.98
12	0.66	-1.78	0.40	-0.69
13	0.83	-0.25	0.78	0.27
14	0.88	0.34	0.87	0.61
15	0.92	-1.86	-1.25	-1.56
16	0.16	-5.05	2.86	-1.10

17	1.05	-0.55	0.30	-0.13
18	0.68	-0.72	1.48	0.38
19	0.76	0.45	1.24	0.85
20	0.08	2.66	14.99	8.83

There is a need to further assess the distance between transition location parameters. Considering their corresponding distances, item 1 had a distance of 0.42 indicating that the distance was very close, more so, the item was very simple, such that it could not discriminate between the high and low ability examinees. Also, item 3 had a high distance between the transition locations and were unable to discriminate very well resulting in a low value. Item 16 gave an outrageous transition distance, also resulting in a low discrimination value. Moreso, item 20 showed the lowest transition distance and discrimination value. Although, three of the items were classified as highly difficulty items using the transition locations, but the overall difficulties only identified item 20 as been highly difficult. Therefore, it was suggested that such items be modified because of their inability to effectively discriminate among examinees.

Research question three: are there differences in the discriminating properties of the items in both scoring methods

The research question was answered by comparing differences in the means of the discriminating indices of both scoring methods using paired sample t-test. The results are presented in Table 5.

Table 5 showed a t-test value (t = 4.35, df = 19 & p<.05). This indicates that there is a significant difference in the discriminating ability of the two scoring methods. Therefore, the null hypothesis is rejected and the alternative hypothesis is accepted that there is a significant difference.

Table 5. Paired t-test of discrimination indices

Discrimination	**N**	**Mean**	**SD**	**T**	**df**	**p**
Dichotomous	20	0.98	0.37	4.35	19	0.00
Polytomous	20	0.74	0.32			

Discussion

The results showed that when the items were scored using dichotomous scoring, most of the items were considered moderately difficult, while a few were of low difficulty. In general, four items were found to be of low difficulty, although they had averagely good discrimination values. One reason that could have accounted for this was that examinees of average and low ability levels were able to answer the questions correctly. Among the items with high difficulty values however, just one item had a poor discrimination value, implying that despite the high difficulty level of the items most of them were able to discriminate well among examinees. In all, an examination of the item difficulty and discrimination led to the deletion of two items from the test.

In the context of GRM however, when the items were scored using polytomous scoring, four items were considered inappropriate using the criteria for item retention and deletion rules. This may be consequent on the fact that the GRM calibrated all the items as having ordered transition locations (manifested through unequal threshold distances).

This buttressed the submissions of Si, Ching-Fung (2002) when in his study the recovery rate used in establishing the accuracy of ability estimates were lower when the items had categories with unequal threshold distances which were close at one end of the ability/difficulty continuum and were administered to a sample of examinees whose population ability distribution was skewed to the same end of the ability continuum. It was therefore evident that the GRM was mild in its calibration resulting in all ordered transition locations and lower number of items deleted in terms of retention and deletion criteria. This could be due to the nature of the development of GRM which was fundamentally built for ordered rating scales; this finding is in support of previous studies by Churchill & Peter (1984).

Moreover, the overall difficulties of the items were calculated, it showed that thirteen items were of low difficulty, six were moderate while an item was considered high in difficulty. In essence, the low-ability examinees that were able to attempt the first transition location were partially credited for their attempt at the items. This is consistent with the principle of ordered polytomous modeling of which the GRM is one. An examination of the two difficulty parameters showed that the transition locations were in increasing order; this implies that the difficulty level of the full credit scored (2) was higher than the difficulty level of the partial credit scored (1).

In respect of the discrimination indices, the results indicated that the indices of the polyotmous items were lower than in the dichotomous case. In view of the fact that, the purpose of the discrimination indices are to show how well items can differentiate among examinees with various ability levels. Many high discrimination indices in the dichotomous test could be taken to depict wide gap among examinees ability. The lower discrimination indices provide leverage among the examinees. In summary, this makes it evident that some of the items included in the SSCE were of high difficulty levels when they are scored dichotomously as correct or incorrect, which awards the correct option as 1 and incorrect as 0. This suggests that these items when scored dichotomously could often times be answered correctly by the high ability examinees and may be a little proportion of the moderate ability examinees. The result buttressed the submissions of Wilson & Master (1993) that the dichotomous items serves as a stifling limitation, thereby put ceiling on low-ability students from being rewarded for their little effort and will at responding to items.

The study concluded that, the polytomous scoring possess good psychometric qualities and is a good prediction of ability estimate of examinees than the dichotomous scoring.

REFERENCES

Adedoyin, O.O. & Adedoyin, J.A. (2013). Assessing the comparability between classical test theory (CTT) and item response theory (IRT) models in estimating test items parameters. *Herald J. Educ. & Gen. Studies, 2*(3), 107 – 114.

Afolabi, E. R. I. (2012). *Tests and measurement: a tale bearer or true witness: inaugural lecture series 253.* Ile-Ife: Obafemi Awolowo University.

Albanese, M.A. (1993). Type K and other complex multiple-choice items: an analysis of research and item properties. *Educ. Measurements*: *Issues & Practice, 12*, 28 – 33.

Andrich, D. (2002). Understanding resistance to the data-model relationship in Rasch's paradigm: a reflection for the next generation. *J. Appl. Measurement, 3,* 325-359.

Baker, F.B. (2001). *The basics of item response theory*. Washington: ERIC *Clearinghouse on Assessment and Evaluation*.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Briggs, D.C., Alonzo, A.C., Schwab, C. & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educ. Assess., 11,* 33-63.

Churchill, G.A. & Peter, J.P. (1984). Research design effects on the reliability of rating scales: a meta-analysis. *J. Marketing Res., 21,* 360-375.

De Ayala, R.J. (1989). Computerized adaptive testing: a comparison of the nominal response model and the three-parameter model. *Educ. & Psych. Measurement, 49*, 789-805.

De Ayala, R.J. (1992). The nominal response model in computerized adaptive testing. *Appl. Psych. Measurement, 16,* 327-343.

Dodd, B.G. & Koch, W.R. (1985). Item and scale information functions for the partial credit model. *Paper presented at the meeting of the American Educational Research Association,* Chicago.

Jodoin, M.G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *J. Educ. Measurement, 40,* 1-15.

Kaplan, R.M. & Saccuzzo, D.P. (2005). *Psychological testing, principles, applications and issue.* Belmont: Thomson Wadsworth.

Lumsden, J. (1978). Tests are perfectly reliable. *British J. Math. & Stat. Psych., 31,* 19-26.

Minstrell, J. (1992). Facets of students' knowledge and relevant instruction (pp. 110-128). In: Duit, R., Goldberg, F. & Nieddere, H. (Eds.). *Research in physics learning: theoretical issues and empirical studies.* Kiel: Institut für die Pädagogik der Naturwissenschaften.

Minstrell, J. (2000). Student thinking and related assessment: creating a facet-based learning environment (pp. 44-73). In: Raju, N.S., Pellegrino, J.W., Bertenthal, M.W., Mitchell, K.J. & Jones, L.R. (Eds.). *Grading the nation's report card: research from the evaluation of NAEP.* Washington: National Academy Press.

Nandakumar, R. & Stout, W. (1993). Refinement of Stout's procedure for assessing latent trait unidimensionality. *J. Educ. Statistics, 18,* 41-68.

Odeyemi, J.O. (1991). Polices and strategies for the improvement of mathematics instruction at the primary school level. *NMC National Planning and Implementation Committee on Mathematics Education Workshop, Abuja.*

Osterlind, S.J. & Everson, H.T. (2009). *Differential item functioning*. Los Angeles: Sage.

Penfield, R.D. & Bergeron, J.M. (2005). Applying a weighted maximum likelihood latent trait estimator to the generalized partial credit model. *Appl. Psych. Measurement, 29,* 218-233.

Scott, T. (2011). *An evaluation of multiple choice test questions deliberately designed to include multiple correct answers: PhD thesis.* Provo: Brigham Young University.

Si, Ching-Fung B. (2002). *Ability estimation under different item parameterization and scoring models: PhD thesis*. Denton: University of North Texas.

Wilson, M. & Masters, G.N. (1993). The partial credit model and null categories. *Psychometrika*. *58*, 87-99.

✉ Professor E. R. I. Afolabi (corresponding author)
Department of Educational Foundations and Counselling
Obafemi Awolowo University
Ile-Ife, Nigeria
E-Mail: eriafolabi@gmail.com