Open Journal of Educational Development (OJED)
ISSN: 2734-2050
Article Details:
DOI: 10.52417/ojed.v3i1.304
Article Ref. No.: OJED0301001-203
Volume: 3; Issue: 1, Pages: 01-19 (2023)
Accepted Date: 10 January, 2022

Open Journals Nigeria (OJN)
Open Access | Bi-annual | Peer-reviewed
www.openjournalsnigeria.org.ng
editorial@openjournalsnigeria.org.ng

RESEARCH ARTICLE **OJED0301001-203**

APPLICATIONS OF ITEM RESPONSE THEORY MODELS TO ASSESS ITEM PROPERTIES AND STUDENTS' ABILITIES IN DICHOTOMOUS RESPONSES ITEMS

[*1]Adetutu, O. M. & [2]Lawal, H. B.

*[*1]Department of Statistics, Federal University of Technology, Minna, Niger State, Nigeria.*
[2]Department of Statistics and Mathematical Sciences, Kwara State University Malete, Ilorin, Kwara State, Nigeria.
***Corresponding Author Phone:** +2348030737153 **Email**: adetutuolayiwola@gmail.com*

ABSTRACT

A test is a tool meant to measure the ability level of the students, and how well they can recall the subject matter, but items making up a test may be defectives, and thereby unable to measure students' ability or traits satisfactorily as intended if proper attention is not paid to item properties such as difficulty, discrimination, and pseudo guessing indices (power) of each item. This could be remedied by item analysis and moderation. It is a known fact that the absence or improper use of item analysis could undermine the integrity of assessment, selection, certification and placement in our educational institutions. Both appropriateness and spread of items properties in accessing students' abilities distribution, and the adequacy of information provided by dichotomous response items in a compulsory university undergraduate statistics course which was scored dichotomously, and analyzed with stata 16 SE on window 7 were focused here. In view of this, three dichotomous Item Response Theory (IRT) measurement models were used in the context of their potential usefulness in an education setting such as in determining these items properties. Ability, item discrimination, difficulty, and guessing parameters as unobservable characteristics were quantified with a binary response test, then discrete item response becomes an observable outcome variable which is associated with student's ability level is thereby linked by Item Characteristic Curves that is defined by a set of item parameters that models the probability of observing a given item response by conditioning on a specific ability level. These models were used to assess each of the three items properties together with students' abilities; then identified defectives items that were needed to be discarded, moderated, and non-defectives items as the case may be while some of these selected items were discussed based on underlining models. Finally, the information provided by these selected items was also discussed.

Keywords: *Ability, Difficulty, Discrimination, Guessing, Response, Test*

INTRODUCTION

The main goal of testing is to collect information to make decisions either about the students' abilities or suitability of test items, and different types of information may be needed depending on the kind of decision is intended to be made. Before Item Response Theory (IRT) was the development of Classical Test Theory (CTT) which was a product of pearsonian statistics intelligence testing movement of the first four decades of 20th century and its attendance controversies (Baker and Kim, 2004). Subsequently, Lord (1968) reformatted the base constructs of CTT using modern mathematical statistical approach where items, and its characteristics played a minor role in the structures of the theory. Earlier, both psychometric theoreticians, and practitioners became dissatisfied over the years with discontinuity between roles of items and test scores in CTT. All were of the opinion that a test theory should start with characteristics of the test items composing a test rather than resultant scores (Brzezinska, 2017).

Two major theories about the development of test are CTT, and IRT (Raykov, 2017). The former is all about reliability with its enormous limitations which includes: estimate of item parameters are group dependent, test item functions that could be either easy or difficult changes as sample changes, ability of students are entirely test dependent, ability of students changes as the occasion changes which result in poor or inconsistency of test, p and r which denote difficulty index and number of students who get item correctly respectively depend on sample of students taking while the latter (IRT) is a bit more complicated than CTT. Rather than looking at the reliability of the test as a whole, IRT looks at each item that makes up the test (Linden, 2018).

ITEM RESPONSE THEORY

An item is a single question or task on a test or an instrument, and Item Response Theory (IRT) is a theoretical frame work organized around the concept of latent trait. It is made up of models, and related statistical methods that define observed responses on instrument to student's level of the ability. It focuses specifically on the items that make up the test, compares the items that make up a test, and then evaluates the extent at which the test measures the student's ability (Raykov and Marcoulides, 2018). IRT models are widely used today in the study of cognitive and personality ability, health responses, items bank development, and computer adaptive testing (Paek and Cole, 2020). For instance, King and Bond (1996) applied IRT to measure anxiety in the use of computer in grade school children, Mislevy and Wu (1996) used IRT in assessing physical functioning in adults with HIV, Boardley *et al.* (1999) used IRT to measure the degree of public policy involvement in nutritional professionals, Olukoya *et al.* (2018) presented a descriptive item analysis of university-wide multiple choice objectives examinations: the experience of a Nigeria private university, Ng *et al* (2016) applied item response theory and Rasch model to develop a new set of speech-recognition tests materials in Cantonese Chinese, Adetutu and Lawal (2020) make a comparisons of frequentist and Bayesian approaches to IRT and discovered that Bayesian approach is better in estimating three item properties along with students' abilities simultaneously, and Zeigenfuse et al. (2020) developed extending dichotomous IRT models to account for test testing behaviour on matching test which violate the assumption of local independence, Bonifay and Cai (2017) findings on the complexity of item response theory models revealed that functional formed of IRT models should be considered not goodness of fit alone when chosen IRT model to be used. However, Suruchi and Rana (2015) identified two uses of item analysis which were the identification of defectives test items, and identifications of areas where students have mastered and not yet mastered. IRT is a potent tool in checking flaws in items and finding ways of correcting them before finally

administering the items hence, item moderation needs to follow item analysis. In cases where item cannot be moderated, such item must be discarded and replaced. Ary *et al.* (2002) asserted that item analysis should make use of statistics that would reveal important and relevant information for upgrading the quality, and accuracy of multiple-choice items. Therefore, IRT plays a central role in the analysis, study of tests and items scores in explaining student test performance, and also provide solutions to test design problems using a test that consists of several items (Baker and Kim, 2004; Baker and Kim, 2017). The potent advantages of IRT over CTT that have propelled us to use IRT are: its treatment of reliability and error of measurement through item information functions which are computed for each item (Hassan, and Miller, 2019),

ASSUMPTIONS OF ITEM RESPONSE THEORY

Undimensionality assumption of IRT implies homogeneity of a test item in the sense of measuring a single ability (Hambleton and Traub, 1973), and the probability of any student's response pattern would be (1's and 0's). On l*ocal independence*, test item response of a given student is statistically independence of another student's response. The implication of this is that test items are uncorrelated for the students of the same ability level (Lord and Novick, 1968). *Monotonicity assumption* focuses on item response functions which model relationship between students' trait level, item properties, and the probability of endorsing the item (Rizopoulos, 2006; De Ayala and Santiago, 2016). Finally, *Item invariance assumption* implies that item parameters estimated by an IRT model would not change even if the characteristics of the student, such as age changes (Peak and Cole, 2019).

STATEMENT OF PROBLEM

Every year, university teachers face the challenge of how to cope with increasing number of examination students, which multiple choice items came to resolve in our educational setting; however, the absence of item analysis in developing these multiple-choice items undermines the integrity of assessments, selection, certification, and placement in our educational institutions. Also, improper use of item analysis leads to same fate while lopsided test items could lead to wrong award of grade, and certificate (Olukoya *et al.*, 2018; Ary *et al.*, 2002). We have seen that hundreds of secondary school students take university entrance examinations, and their results determine the entry into universities, and possible alternatives (Eli-Uri and Malas, 2013; Cechova *et' al*., 2014). Hence, the needs to maintain the validity of tests using IRT models necessitate this study.

RESEARCH JUSTIFICATION

Professional conduct of item analysis that makes use of statistics would reveal important, and relevant information about the item for upgrading the quality, and accuracy of multiple-choice items, its power lies in identifying defective items, areas where students have mastered, and area not yet mastered thereby find ways of correcting them before finally administered them in order to have integrity in assessment, selection, certification, and placement in our educational institutions.

OBJECTIVES

1. Determine the spread and appropriateness of item properties in multiple choice items.
2. Access the distribution of students' abilities and the adequacy of information provided by test items.

METHODOLOGY

DATA DESCRIPTION

Data used in illustrating these (one, two, and three-logistic) models were results of a university semester examination where a total of 403 students took a compulsory general statistics course in the university semester examination for 2017/2018 academic session. The test items (questions) were made up of 35 multiple choice items, where each item had 4 options, each of which had a correct option while the other three options were distractors. The same test items are administered to all the students, and their responses in terms of options chosen are coded into binary, (that is 0 for endorsing any of the incorrect options and 1 for endorsing a correct option) using Stata/SE 16.0 on window 7. Some selected items in supplementary section were discussed.

METHODS

Method I: Rasch/One-parameter logistic Model

The first model employed is basically for accessing how difficult an item is being perceived by the test takers, it was proposed by Georg Rasch, a Danish mathematician in 1966 (Rasch, 1966) similar to One-parameter logistic model (1PL) proposed by Birnbaum (1968). This model is positioned in equation (1) and its described test item in term of only one parameter called difficulty index. The probability that student K with ability (θ_k) will endorse item g with difficulty index (b_g) correctly is presented in equation (1):

$$P_{gk}(\theta_k) = \frac{e^{a(\theta_k - b_g)}}{1 + e^{a(\theta_k - b_g)}} \quad (1)$$

Where:

$\boldsymbol{a}$ is discrimination index denoting how an item discriminates students which is constrained under this model,

$\boldsymbol{b_g}$ is the item difficulty parameter for item $(g = 1, 2, \ldots, n)$, denotes how students perceived the item, and

$\boldsymbol{\theta_k}$ is the student $k's$ ability ($k = 1, 2, \ldots, N$).

Under the model in equation (1), $\boldsymbol{a}$ is constrained ($\mathbf{a} = 1$ for Rasch model, and $\mathbf{a} < 1$ for One-parameter logistic model).

Method II: Two-parameter Logistic Model

The second model employed called two-parameter logistic model in equation (2) measures how well an item discriminates between different ability levels near the infection point of ICC. It estimates varied item difficulty and discrimination indices simultaneously, this model is useful in determining how items segregate students according to their ability levels. Theoretically, it ranges between $-\infty \; to \; \infty$ but in practice negative discriminations are discarded. The model can be obtained from equation (1) by adding varying item discriminating parameters $a_g \; (g = 1,2, \ldots, n)$. The probability that student K with ability θ_k endorsed item g correctly is given in equation (2):

$$P_{gk}(\theta_k) = \frac{e^{a_g(\theta_k - b_g)}}{1 + e^{a_g(\theta_k - b_g)}} \quad (2)$$

Where:

, b_g and θ_k were as defined in equation (1).

Method III: Three-parameter Logistic Model

Finally, the third model is three-parameter logistic model positioned in equation (3) which was used to estimate psuedoguessing indices for the test items. This model described items in term of three parameters which are: difficulty, discrimination, and guessing indices (Lim, 2020). The probability of correct response $P_{gk}(\theta_k)$ to item g by student k with ability θ_k is determined by item discrimination parameter a_g, item difficulty parameter b_g, guessing parameter c_g $(g = 1,2,\dots,n)$, and the student's ability θ_k is as presented in equation (3)

$$P_{gk}(\theta_k) = C_g + (1 - C_g)\ \frac{e^{a_g(\theta_k - b_g)}}{1 + e^{a_g(\theta_k - b_g)}} \tag{3}$$

PARAMETERS ESTIMATION

Parameterization of models position in equations (1), (2), and (3), let y_{gk} be the observed response for Y_{gk} outcome for item g from student k by taking $y_{gk} = 1$ as correct option and $y_{gk} = 0$ as incorrect option. The probability that k^{th} student with ability level θ_k responds correctly to item g is given by the equation (4)

$$\Pr(Y_{gk} = 1 | a_g, b_g, c_g, \theta_k) = C_g + (1 - C_g)\ \frac{e^{a_g(\theta_k - b_g)}}{1 + e^{a_g(\theta_k - b_g)}} \tag{4}$$

Where:

a_g, b_g, and c_g were as defined in equations (1), (2), and (3). When guessing parameter c_g is constrained to be equalled to zero, equation (4) becomes equation (2) (two-parameter logistic model), when both $c_g = 0$ and $a_g = 1$ or $a < 1$, equation (4) turns to be equation (1).

We fit three-parameter model by using slope-intercept of the form in equation (5)

$$\Pr(Y_{gk} = 1 | \alpha_g, \beta_g, \gamma_g, \theta_k) = \frac{e^{\gamma_g}}{1 + e^{\gamma_g}} + \frac{1}{1 + e^{\gamma_g}} + \frac{e^{\alpha_g(\theta_k + \beta_g)}}{1 + e^{\alpha_g(\theta_k + \beta_g)}} \tag{5}$$

and the transformation between these parameterizations is

$$a_g = \alpha_g,\ \ b_g = \frac{-\beta_g}{\alpha_g},\ \ c_g = \frac{e^{\gamma_g}}{1 + e^{\gamma_g}} \tag{6}$$

The γ_g (that is c_g) can be constrained to be the same across all items. Let

$$P_{gk} = \Pr(Y_{gk} = 1 | \alpha_g, \beta_g, \gamma_g, \theta_k) \qquad Q_{gk} = 1 - p_{gk} \tag{7}$$

Conditional on θ_k for student k since item responses are assumed to be independent is given by

$$f(y_k | \Omega, \theta_k) = \prod_{g=1}^{n} P_{gk}^{\ y_{gk}}\ Q_{gk}^{\ 1 - p_{gk}} \tag{8}$$

Where:

$y_k = (y_{1k}, \dots, y_{nk})$,

$\Omega = (\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_n, \gamma_1, \dots, \gamma_n)$, and

n is the number of items. The likelihood for student k is computed by integrating out the latent variable from the joint density in equation (8)

$$L_k(\Omega) = \int_{-\infty}^{\infty} f(y_k | \Omega, \theta_k) \emptyset(\theta_k) d\theta_k \tag{9}$$

$\emptyset(.)$ denotes density function for standard normal distribution. For N students, the sum of the log likelihood in equation (9) is

$logL(\Omega) = \sum_{k=1}^{N} L_k(\Omega)$ (10)

However, the integral for $L_k(\Omega)$ in equation (9) is generally in a closed form, we used numerical methods (Adaptive/Gauss-Hermite Quadrature) implemented with stata 16.SE software on window 7.

RESULTS AND DISCUSSIONS

Knowing fully that IRT models are useful in test development by supplying indices of item difficulty, discrimination and, guessing to match the ability level of a target population. The estimated item difficulty indices in (descending order of item difficulty indices) using equation (1) for selected items were presented in Table 1 together with their indices of precision (SE), the probability of an average student correctly endorsing each of the items (Prob), the confidence interval of the estimates, and remark on the item suitability while Interpretations of difficulty indices are displayed in Table 2.

Items 15, 5, 3, 13, 28, 34, 23, and 11 were identified to be defectives going by item difficult interpretations in the Table 2 (Henning, 1987). Item 8 was perceived to be very difficult; its difficulty index is 2.4988 meaning that a student needs to be at least on ability level 2.5 to answer this item correctly. The probability that an average student answers this item correctly is 0.1713 which implies that most likely; only about 17% of the entire students would endorse correct options, and we are 95% sure that this estimate lies within (2.0154, 2.9821) confidence interval. Follow by item 8 are items; 1, 33, 7, and so on. A careful examination of the results also suggested that Item 11 was perceived by the students as the easiest; an average student would correctly endorse right option to this item with probability 0.9601. This is most likely that about 96% of the students would pass the item. Follow by item 11 are; 23, 34, 28, and so on (as presented in the Table 1).

Table 1: Some Selected Items in Descending Order of Their Difficulty Indices

Item	Diff	S E	Prob	95% Conf. Interval		Remark
8	2.4988	0.2466	0.1713	2.0154	2.9821	Very difficult
1	1.2667	0.1967	0.3140	0.8802	1.6511	Difficult
33	0.9200	0.1872	0.3617	0.5531	1.2869	Moderately difficult
7	0.8455	0.1817	0.3697	0.4893	1.2017	Moderately difficult
9	0.6805	0.1823	0.3965	0.3233	1.0377	Moderately difficult
25	0.6805	0.1823	0.3965	0.3233	1.0377	Moderately difficult
10	0.6261	0.1813	0.4046	0.2707	0.9815	Moderately difficult
27	0.3401	0.1777	0.4477	-0.0081	0.6884	Moderately difficult
19	0.4301	0.1777	0.4477	-0.0081	0.6884	Moderately difficult
2	0.1990	0.1766	0.4694	-0.1472	0.5452	Moderately difficult
32	0.1462	0.1764	0.4775	-0.1995	0.4919	Moderately difficult
35	0.1462	0.1764	0.4775	-0.1995	0.4919	Moderately difficult
24	-0.2048	0.1766	0.5316	-0.5509	0.1412	Moderately difficult
6	-0.2224	0.1766	0.5343	-0.5686	0.1238	Moderately difficult
14	-0.6677	0.1817	0.6107	-1.0238	-0.3116	Moderately difficult
17	-1.0567	0.1901	0.6577	-1.4294	-0.6840	Easy
30	-1.2894	0.1969	0.6892	-1.6754	-0.9035	Easy
26	-1.4100	0.2009	0.7049	-1.8034	-1.0159	Easy
12	-1.5957	0.2076	0.7282	-2.0026	-1.1887	Easy
29	-1.6034	0.2034	0.7333	-2.0032	-1.2036	Easy
31	-1.6381	0.2093	0.7334	-2.0483	-1.2279	Easy
20	-2.2113	0.2350	0.7966	-2.6720	-1.7507	Very easy
18	-2.2889	0.2389	0.8041	-2.7551	-1.8187	Very easy
16	-2.6092	0.2566	0.8335	-3.2747	-2.1063	Very easy
4	-2.7549	0.2652	0.8456	-3.2747	-2.2350	Very easy
22	-2.9729	0.2789	0.8623	-3.5196	-2.4263	Very easy
15	-3.1063	0.2877	0.8719	-3.6702	-2.5425	Poor
5	-3.2486	0.2935	0.8859	-3.8238	-2.6734	Poor
3	-3.5160	0.3168	0.8974	-4.1369	-2.8950	Poor
13	-3.6414	0.3264	0.9045	-4.2811	-3.0017	Poor
28	-4.2313	0.3758	0.9315	-4.9679	-3.4947	Poor
34	-4.7140	0.4200	0.9514	-5.5370	-3.8908	Poor
23	-5.1555	0.4600	0.9579	-5.9679	-4.1646	Poor
11	-5.0360	0.4556	0.9601	-5.9329	-4.1471	Poor
Disc	0.6314	0.0335	---------	0.5653	-4.1471	----------

Table 2: Interpretations of Difficulty Values

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

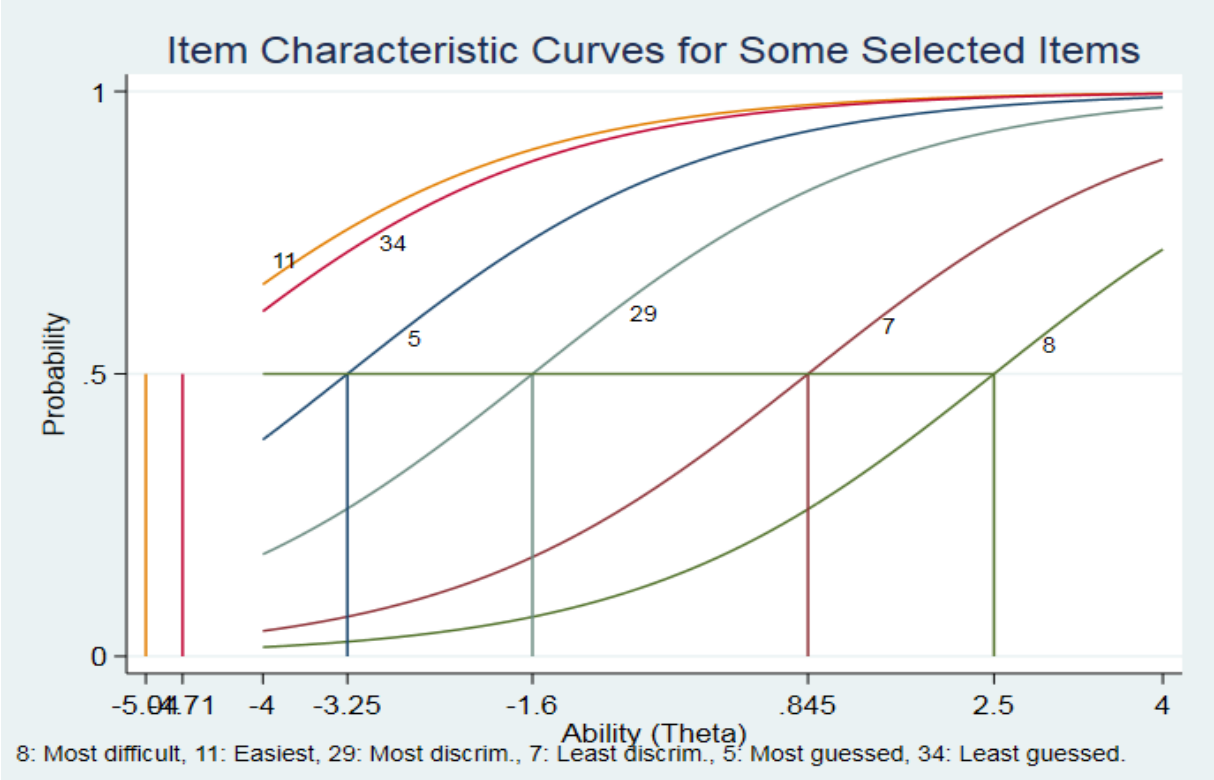

Figure 1: Difficulty of Some Selected Items (1PL)

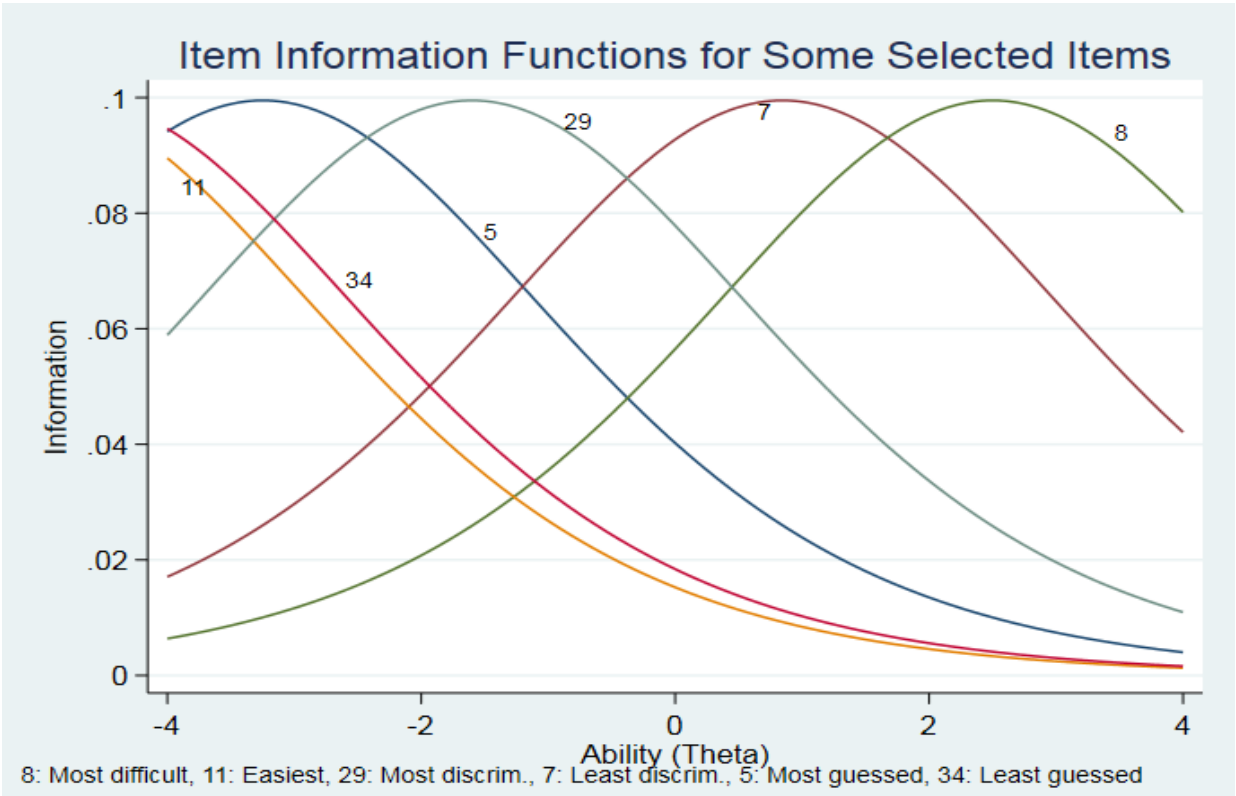

Figure 2: Item Information (1PL)

The graphical evidences in Figures 1, and 2 buttress the fact that Items 11, 34, and 5 were mainly for less able students for the fact that the items needed low level of ability for correct endorsement. Students only need to be from -5.2 to -4.7 on trait scale to correctly endorse items 11, and 34 respectively.

On the basis of equation (1), item 8 provided more of its information on able students while item 7 provided information on students who were on both sides of location point (higher and lower ability students), this is displayed in Figure 2. Difficulty indices of items describe where the item functions along the ability scales, and the model suggests through their difficulty indices that items 15, 5, 3, 13, 28, 34, 23, and 11 displayed in Table 1 needs attention as remarked.

Application of equation (2) which describes test items in terms of two item properties yields output displayed in Table 3, with the confidence interval, probability of correct endorsement by an average student, indices of items discrimination (descending order) that classified students according to their ability as well as the precision of the estimates (SE). Item 29 was identified as the most discriminating $(a = 1.7889)$ though not so difficult (b = -0.8064). An average student would endorse this item correctly with probability 0.7333 which means about 73% of the students are most likely to endorse correct option to the item. Again, students only need to be -0.8 on ability scale to endorse a correct option. Followed by item 29 was item 34, and so on in that order. Another item that draw attention in Table 3 was item 7 which did not discriminate (a = 0.01505) between the students of different ability levels.

Table 3: Some Selected Items in Descending Order of Their Discrimination Indices

Item	Diff	Disc	SE	95% Conf Inter		Remark
29	-0.8064	1.7889	0.2947	1.2113	2.3666	Satisfactory
34	-2. 4237	1.5023	0.3763	0.7648	2.2399	Good item
20	-1.2485	1.3366	0.2387	0.8687	1.8045	Moderate
23	-2.7667	1.3048	0.3700	0.5796	2.0300	Moderate
21	-0.7763	1.2713	0.2130	0.8539	1.6886	Moderate
11	-3.1843	1.0981	0.3410	0.4298	1.7664	Moderate
12	-1.0104	1.1287	0.1969	0.7428	1.5149	Moderate
13	-2.2983	1.0884	0.2524	0.5936	1.5832	Moderate
30	-0.8573	1.0560	0.1826	0.6981	1.4139	Moderate
27	0.2122	1.0508	0.1706	0.7165	1.3851	Moderate
15	-2.1393	0.9727	0.2193	0.5428	1.4026	Moderate
22	-2.1335	0.9221	0.2093	0.5119	1.3323	Moderate
4	-2.0404	0.8866	0.1993	0.4960	1.2773	Moderate
16	-1.9603	0.8716	0.1959	0.4877	1.2554	Moderate
18	-1.8529	0.7933	0.1779	0.4445	1.1420	Moderate
31	-1.4944	0.6875	0.1563	0.3810	0.9939	Moderate
3	-3.1911	0.6874	0.2103	0.2751	1.0996	Moderate
8	2.1361	0.7613	0.1730	0.4223	1.1002	Moderate
14	-0.6417	0.6497	0.1418	0.3718	0.9276	Moderate
26	-3.3734	0.6357	0.1488	0.3440	0.9274	Marginal
32	0.1384	0.6343	0.1410	0.3580	0.9106	Marginal
25	0.6802	0.6121	0.1365	0.3446	0.8795	Marginal
35	0.1523	0.5747	0.1362	0.3077	0.8418	Marginal
5	-3.4127	0.5961	0.1951	0.2138	0.9785	Marginal
17	-1.2942	0.4887	0.1361	0.2220	0.7554	Marginal
28	-5.4431	0.4663	0.2311	0.0133	0.9193	Marginal
10	0.8116	0.4584	0.1278	0.2079	0.7089	Marginal
1	1.7964	0.4159	0.1301	0.1608	0.6710	Marginal
24	-0.3054	0.3918	0.1252	0.1465	0.6372	Marginal
2	0.3338	0.3500	0.1225	0.1099	0.5901	Marginal
33	1.5802	0.3395	0.1246	0.0952	0.5838	Poor
19	0.6821	0.2898	0.1208	0.0531	0.5265	Poor
6	-0.6483	0.1936	0.1180	-0.0376	0.4248	Poor
9	2.5844	0.1505	0.1187	-0.0822	0.3832	Poor
7	32.6302	0.0151	0.1206	-0.2213	0.2514	Poor

Items 33, 19, 6, 9, and 7 are defectives base on standard discriminatory power interpretations displays in Table 4 (Ebel and Frisbie, 1991), not minding their difficult indices, with this model position in equation (2) where items had varied item discriminations, student must be on too extremely high ability level ($\theta = 32.6302$) to endorse a correct option to this item 7. This is an indication that this item needs attention, either the item was poorly written or there was misinformation in the item. The probability that an average student endorsed item 7 correctly was 0.3697; that is about 37% of the students got the item. In the order of least item discrimination were items 7, 9, 6, and so on. If the purpose of using the instrument is to segregate students into those who mastered and not yet mastered, Table 3 suggested those remarked "poor" are likely to be defectives, hence need attention based on their item discrimination indices as remarked.

Table 4: Interpretations of Items Discrimination Indices

Discrimination Indices (a)	Interpretations
$C \geq 1.70$	Item is functioning quite satisfactorily
$1.35 \leq C \leq 1.69$	Good item.
$0.65 \leq C \leq 1.34$	Moderate, little or no revision is needed
$0.35 \leq C \leq 0.64$	Item is marginal and needed moderation
$C \leq 0.34$	Poor item, should be eliminated or moderated

The graphical display in Figures 3 and 4 perfectly agreed with Table 3 in the sense that item 7 did not discriminate and had no sufficient information about students; this was followed by item 5 which was also not informative. Item 29 was much informative about students on both sides of ability continuum while item 8, which was perceived as the most difficult going by equation (1), here only, gave very little information which was only on high ability students.

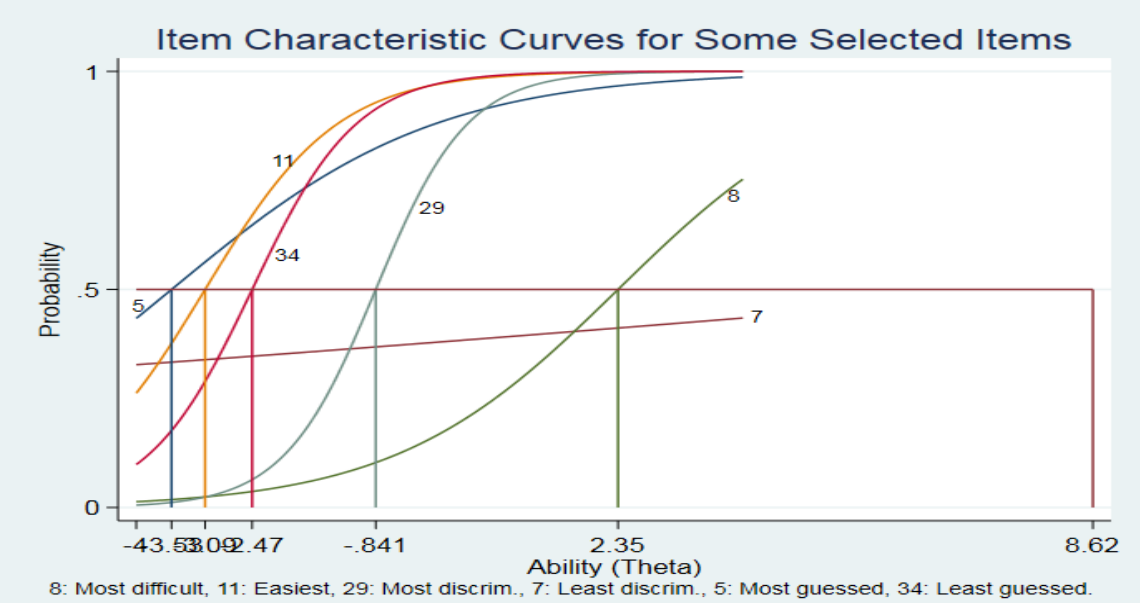

Figure 3: Item Discrimination Indices for Some Selected Items

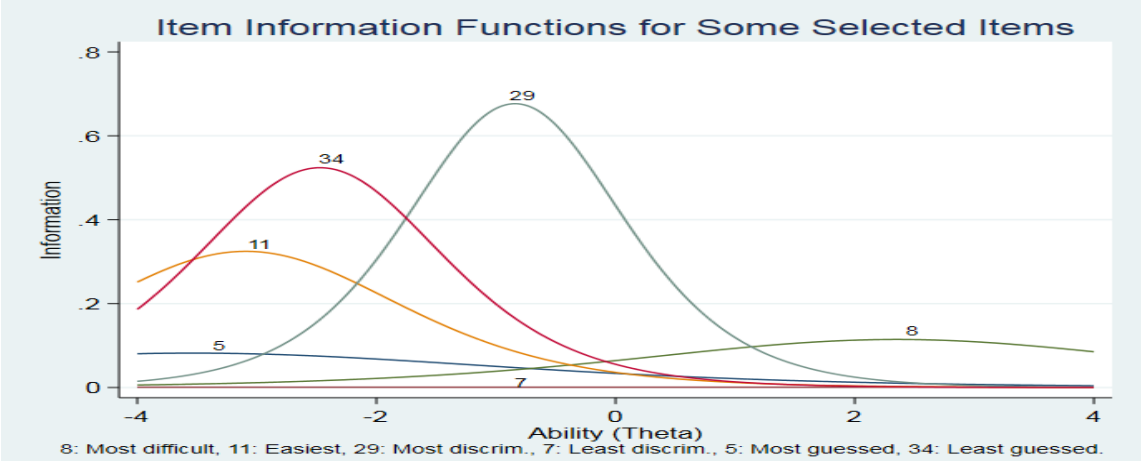

Figure 4: Information Provided by Some Selected Items

A further application of three parameter logistic model in equation (3) to the item responses describes each item in terms of three item properties yields the results in Table 5 along with indices of guessing, difficulty when discriminatory power of all items is held constant. The results suggested that item 5 was the most likely guessed by the students. An average student would endorse this correctly with probability 0.8649, and 0.9855 probability of its guessing for students to select correct option to the item, he or she must be on a high ability level $\theta = 03029$ despite the difficulty of this item, it was guessed correctly. To improve quality of multiple-choice items such items as 5, 23, and 3 remarked "poor" which were suggested to be defectives and must be revisited, moderated due to their high psuedoguessing indices.

Table 5: Some Selected Items in Descending Order of Their Guessing Indices

Item	**Guessing**	**Diff**	**SE**	**95% Conf Inter**		**Remark**
Disc	1.3727		0.0788			
5	0.9855	0.3029	0.0546	0.9719	1.0052	Poor
23	0.7855	-2.7123	0.0658	0.7085	0.9918	Poor
3	0.7377	-0.2184	0.0857	0.0447	1.3104	Poor
4	0.5641	-0.4292	0.1165	-0.3469	0.7762	Marginal
18	0.5245	-0.2026	0.1078	-0.1504	0.8564	Marginal
16	0.5133	-0.4936	0.5133	0.3081	0.8266	Marginal
15	0.5125	-0.8531	0.1665	0.2334	0.6836	Marginal
28	0.5123	0.7337	0.0547	0.6990	0.7210	Marginal
17	0.5008	0.8829	0.0597	0.2218	0.6977	Marginal
31	0.5096	0.3074	0.0840	0.0231	0.6432	Marginal
26	0.4973	0.4721	0.0714	-0.1974	0.8497	Marginal
24	0.4346	1.5546	0.0484	0.0142	0.5581	Moderate
22	0.4313	-0.9702	0.1878	-0.4791	0.7036	Moderate
13	0.4251	-1.4024	0.3356	0.0172	0.5932	Moderate
9	0.3659	2.5314	0.0341	0.5129	0.6487	Moderate
35	0.3287	1.2120	0.0525	0.0910	0.5010	Moderate
2	0.3943	1.8338	0.0414	0.3591	0.5066	Moderate
14	0.3891	0.6821	0.0647	0.1500	0.5571	Moderate
19	0.3830	1.9712	0.0396	0.3606	0.4706	Moderate
7	0.3643	3.3063	0.0320	0.2942	0.3906	Moderate
32	0.3124	1.1142	0.0528	0.2911	0.4108	Good
10	0.2811	1.4529	0.0428	0.2000	0.4416	Good
27	0.1521	0.5975	0.0591	0.2307	0.4039	Good
21	0.1270	-0.5258	0.1250	0.2955	0.5597	Good
33	0.2979	2.0190	0.0370	0.0123	0.5864	Good
12	0.2787	-0.4024	0.1198	-0.6247	0.3760	Good
30	0.2756	-0.2027	0.1055	-0.1582	0.7155	Good
1	0.2385	1.9215	0.0366	0.1504	0.3760	Good
25	0.2582	1.3613	0.0462	-0.1167	0.5529	Good
20	0.1645	-1.0002	0.1837	0.1009	0.6671	Good
8	0.0964	2.0545	0.0259	0.0082	0.1059	Satisfactory
11	0.0050	-2.7534	0.1839	0.0014	0.0194	Satisfactory
29	0.0017	-0.9085	0.1115	0.0013	0.0132	Satisfactory
34	0.0002	-2.5906	0.0151	0.0013	0.0132	Satisfactory

Conversely, items 8, 11, 29, and 34 had negligible guessing indices according to our classification, and interpretation in Table 6 meaning that students most unlikely guessed these identified items. Figures 5 and 6

agreed with Table 5 on this, no wonder these items provided a reasonable amount of information graphically. A careful cursory review of the figures attested that item 5 needs attention as suggested earlier. Their item information functions were too narrow.

Table 6: Interpretations of Items Guessing Indices

Guessing Indices (c)	Interpretations
$c < 0.15$	Item functions quite satisfactorily
$0.15 \leq c \leq 0.35$	Good item
$0.36 \leq c \leq 0.40$	Moderate, little or no revision is needed
$0.41 \leq c \leq 0.60$	Item is marginal, moderation is needed
$c > 0.60$	Poor item, it should be eliminated or moderated

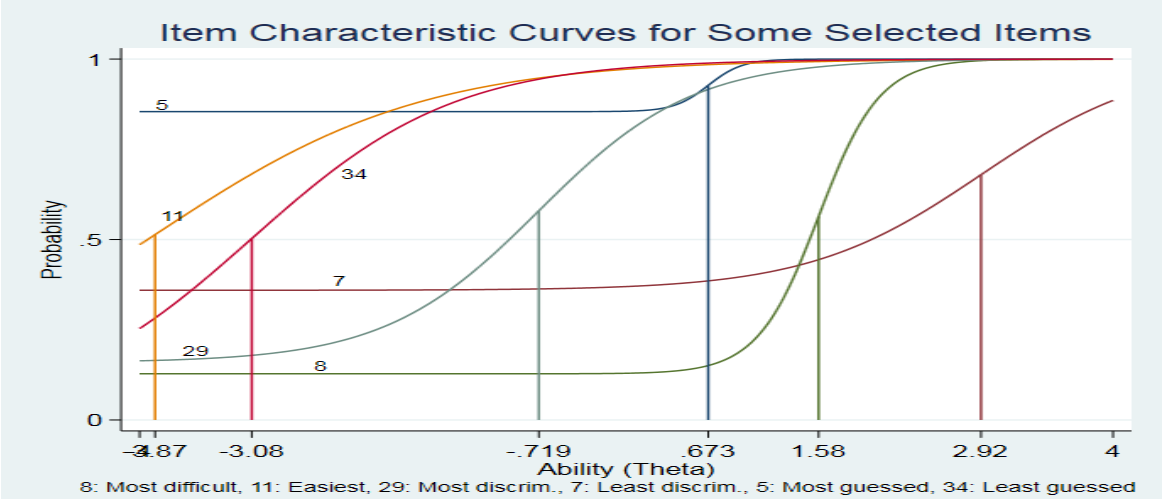

Figure 5: Guessing Indices of Some Selected Items

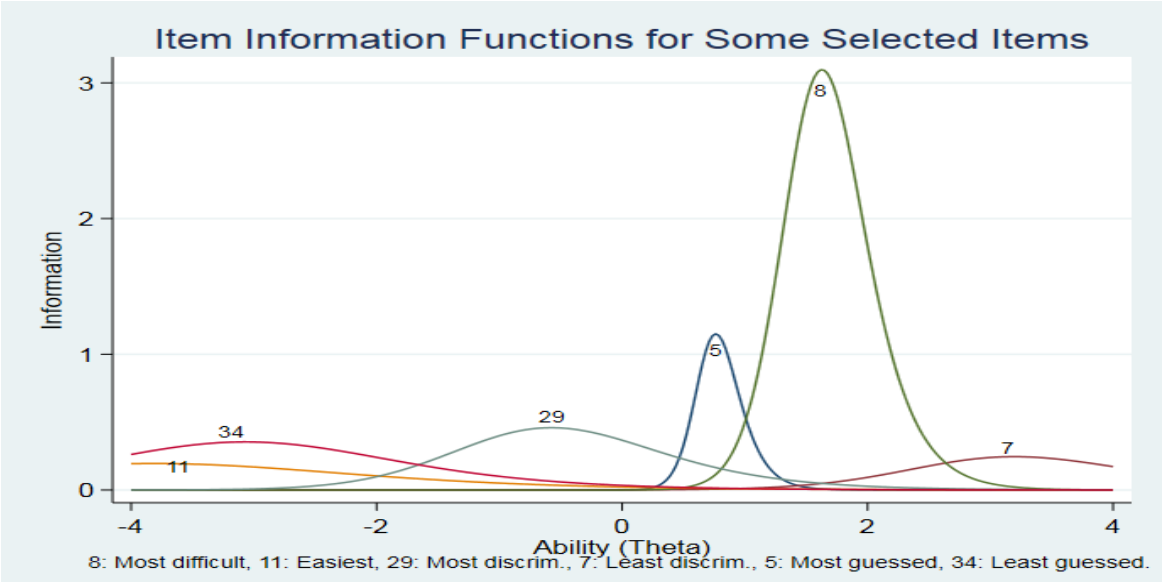

Figure 6: Effect of Guessing on Information Provided by Selected Items

CONCLUSIONS

The major importance of item analysis is the identification of defective test items for the purpose of correcting them by using relevant statistics which revealed important, and relevant information about the items for the purpose of upgrading these multiple-choice items. Dichotomous IRT models applied identified flaws in the selected items as suggested in Tables 1, 3, and 5. Very easy items 5, 34 and 11 need item moderation to upgrade its difficulty indices. Moreover, items 5, and 7 were unlikely to discriminate ability among students as identified by equation (2), and presented in Table 3, these need item moderation as well while item 5 was identified as most guessed where a less ability student most likely to endorse it correctly. These are warning messages that need the attention of test developers, for possible remedies.

Findings that will assist test developers are stated as follows:

1. Students' abilities were modelled under different IRT models as displayed in Tables 1, 3, and 5.
2. A high item difficult index without a commensurate discriminatory power would not serve the intended purposes.
3. Item response analysis of our examination questions at different levels of education is important to identify defectives items.
4. A good test items such as item 29 segregates different ability levels of students as displayed in Figure 4.
5. Good distractors are boaster to quality multiple choice questions
6. Ill prepared item such as 5 will prone to guessing as shown in Table 5.

It has been discovered that for items to be suitable in providing needed information about the students especially in education setting, it must provide information that cut across different levels of ability scales. Too much difficult items (questions) may not serve the purpose intended for. It is very important that items (tests) developers and administrators should take advantages offered by item analysis in order to identify defectives items and in extensive study of tests, item scores, and assessment of students' ability.

Our recommendations to test developers and administrators such as examination bodies and higher institutions of learning are as follows:

1. Item moderation should always precede item analysis. In cases where items cannot be moderated, such item must be discarded and replaced.
2. Item analysis should make use of statistics that would reveal important and relevant information for upgrading the quality and accuracy of multiple-choice questions.
3. Item analysis as a potent tool must be used in checking flaws in items and finding ways of correcting them before finally administering the items.
4. A major element in the quality of a multiple-choice item is the quality of item's distractors, neither the item difficulty nor the item discrimination index considers the performance of incorrect response options, or distractors; hence, a distractor analysis that addresses the efficacy of these incorrect response options should be made mandatory.

In conclusion, all items must be trial tested to identify flaw items and thereby make necessary corrections which would involve collaboration work of test developers, and psychometricians for the purpose of improving quality of selection, certification and graduates in higher institution of learning.

CONFILCT OF INTREST

There is no conflict of interest.

ACKNOWLEDGEMENT

We wish to acknowledge the support provided by the entire staff of Department of Statistics and Mathematical Sciences, Faculty of Pure and Applied Sciences, Kwara State University Malete, Ilorin, Nigeria for their great help and support in data acquisitions.

REFERENCES

1. Adetutu, O. M., and Lawal, H. B. (2020). On Comparisons of Frequentist to Bayesian Estimation for Item Response Theory Models in the Presence of Dichotomous Responses. *Journal of Science, Technology, Mathematics and Education,* **16**(4), 128-137.
2. Ary, D., Jacobs, L. C., and Razavieh, A. (2002). Introduction to Research in Education (6th ed.). California: Wars worth.
3. Baker, F. B., and Kim, S. H. (2004). Item Response Theory: Parameter Estimation Techniques (2nd ed.). New York: Taylor and Francis.
4. Boardley, D. C., Fox, M., and Robinson, K. L. (1999). Public Policy Involvement of Nutrition Professions. *Journal of Nutrition Education,* **31**(5), 248-254.
5. Bonifay, W., and Cai, L. (2017). On the Complexity of Item Response Theory Models. *Multivariate Behavioral Research*, **00**(0), 1-20. doi: 10.1080/00273171.2017.1309262 http://dx.doi.org/10.1080/00273171.2017.1309262
6. Brzezinska, J. (2017). Item Response Theory Models in the Measurement Theory. *Communications in Statistics - Simulation and Computation*, **49**(12), 3299-3313, doi: 10.1080/03610918.2018.1546399 https://doi.org/10.1080/03610918.2018.1546399
7. De Ayala, R. j., and Santiago, S, Y. (2016). An Introduction to Mixture Item Response Theory Models. *Journal of School Psychology*, **60**, 25-40, http://dx.doi.org/10.1016/j.jsp.2016.01.002
8. Ebel, R. L., and Frisbie, D. A. (1991). Essentials of Educational Measurement (5th ed.). Prentice Hall, Engelwood Cliffs.
9. Eli-Uri, F. I., and Malas, N. (2013). Analysis of Use of Single Best Answer Format in an Undergraduate Medical Examination. *Qatar Medical Journal*, **1**, 3-6.
10. Hassan, M. U., and Miller, F. (2019). Discrimination with Unidimensional and Multidimensional Item Response Theory Models for Educational Data. *Communications in Statistics - Simulation and Computation,* **49**(12), 3299-3313,doi: 10.1080/03610918.2019.1705344, https://doi.org/10.1080/03610918.2019.1705344
11. Hambleton, R. K., and Traub, R. E. (1973). Analysis of Empirical Data Using Two-parameter Logistic Latent Trait Models. *British Journal of Mathematics and Statistical Psychology.*, **26**, 195-211.

12. Henning, G. (1987). A Guide to Language Testing: Development, Evaluation Research. New Berry House Publisher.
13. Ng, I. –Y. I., Lee, K. Y. S., Lam, J. H. S., van Hasselt, C. A., and Tong, M. C. F. (2016). An Application of Item Response Theory and Rasch Model in Speech Recognition Test Materials. *American Journal of Audiology*, **25**, 142-152.
14. King, J., and Bond, T. G. (1996). A Rasch Analysis of a Measure of Computer Anxiety. *Journal of Educational Computing Research*, **14**, 49-64.
15. Lim, S. (2020). Review: A Course in Item Response Theory and Modeling with Stata, and Using R for Item Response Theory Model Applications. *Structural Equation Modeling: A Multidisciplinary Journal*. **27**(4), 657-659, doi: 10.1080/10705511.2020.1740886, https://doi.org/10.1080/10705511.2020.1740886
16. Linden, A. (2018). Review of Tenko Raykov and George Marcoulides's A Course in Item Response Theory and Modeling with Stata. *The Stata Journal: Promoting Communications on Statistics and Stata*, **18**(2), 485–488 . https://doi.org/10.1177/1536867X1801800213
17. Lord, F. M., and Novick, M. R. (1968). Statistical Theories of Mental Test Scores. M. A: Addison-Wesley.
18. Mislevy, R. J., and Wu, P. K. (1996, June). Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing [*Research Report- RR-96-30-0NR*]. Available from https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.1996.tb01708.x
19. Olukoya, J. A., Adekeye, O., Igbinola, A. O., and Afolabi, A. (2018). Item Analysis of University-wide Multiple Examinations: The experience of a Nigeria private university., *Quality & Quantity Journal*,. **52**(3), 98-997. Available from https://doi.org/10.1007/s11135.017-0499-2
20. Paek, I., and Cole, K. (2019). Using R for Item Response Theory Model Applications. New York, NY: Routledge.
21. Raykov, T., Dimitrov, D. M., Marcoulides, G. A., and Harrison, M. (2017). On the Connections Between Item Response Theory and Classical Test Theory: A Note on True Score Evaluation for Polytomous Items via Item Response Modeling. *Educational and Psychological Measurement*, **00** (0) 1-12
22. Raykov, T., & Marcoulides, G. A. (2018). A Course in Item Response Theory and Modeling with Stata. College Station, TX: Stata Press College Station.
23. Rizopoulos, D. (2006). Item: An R Package for Latent Variable Modelling and Item Response Theory, **17**, 1-25.
24. Zeigenfuse, M. D., Batchelder, W. H., and Steyvers, M. (2020). An Item Response Theory Model of Matching Test Performance. *Journal of Mathematical Psychology*, **95**, 102327. https://doi.org/10.1016/j.jmp.2020.102327
25. Zogar, E. Y., and Kelecioglu, H. (2017). Examination of Different Item Response Theory Models on Tests Composed of Testlets. *Journal of Education and Learning*, **6** (4), http://doi.org/10.5539/jel.v6n4p113 .

APPENDIX

Selected Items for Discussions and Analysis

3. One of the merits of secondary source of data is that it ----------

(A) is less expensive

(B) is less informative

(C) may not as detail as required

(D) does not give quicker information

5. ---------- consists as of all subjects (human or otherwise) that are being studied.

(A) Population

(B) Sub population

(C) A community

(D) A sample

6. A graphical device for representing qualitative data summarises based on bars is called?

(A) Histogram

(B) Stem-and-leaf display

(C) Pie Chart

(D) Bar Chart

Instruction: Use this information below to answer questions 7, 8, and 9.

A cell, when multiplies can give birth to a maximum of four daughter cells. The probability of X daughter being formed by a cell which has just multiplied is given by the following probability distribution

X	1	2	3	4
$P(X)$	$\frac{1}{4}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{4}$

7. What type of variable is X?

(A) Discrete/Qualitative

(B) Continuous/Quantitative

(C) Binomial

(D) Random

8. The mean of the daughter cell X is

(A) 2.5

(B) 2.0

(C) 1.5

(D) 2.4

9. The variance of daughter cell X is

(A) 1.234

(B) 2.234

(C) 3.234

(D)4.234

11. Data ---------- source are datasets obtained directly from the concerned object.

(A) primary

(B) secondary

(C) tertiary

(D) nursery

13. The following are scales of measurement except ----------

(A) statistic

(B) interval

(C) ordinal

(D) nominal

15. Given the heights of five men in inches as 67, 69, 71, 75, and 80. The range is

(A) 13.

(B) 67.

(C) 71.

(D) 23.

Suppose a sample consists of women age at birth of all live born of all alive born infants at a private hospital in a city during a 1-week period, as shown in the table below; cumulative frequency [CF], Relative frequency [RF], Class interval [CI], Class boundaries [CB], Class Mark [CM]. [Use the information to answer questions 29 and 34].

CI	CB	CM	F	FX	CF	FR
18-20		19		76		0.08
21-23				Q		0.20
24-26				200		0.16
27-29				196		0.14
30-32				155		0.10
33-35		34		136		0.08
36-38				185		0.10
39-41				280	50	0.14
Total				1448		1

29. The arithmetic mean of the women age is

(A) 29.86

(B) 25.96

(C) 14. 48

(D) 28.96

34. The class boundary of the first class is

(A) 38.5-41.5

(B) 17.5-20.5

(C) 23.5-26.5

(D) 20.5-23.5