IRT Models: On Propensity of Endorsing Item-Options in Multi-category Test Items and Response Categories Analysis

Adetutu O. M. [1] and Lawal H. B. [2]

Department of Statistics, Federal University of Technology, Minna[1].
Department of Statistics and Mathematical Sciences, Kwara State University, Malete[2].
adetutuolayiwola@gmail.com[1] *blawal66@gmail.com*[2]

5th July, 2021.

Abstract

Every year, university teachers face challenge of how to cope with increasing number of examination students together with the validity of test items. Due to this, multiple choice items came to address this problem but little or no attention is paid to the properties of test items, and absent of item analysis in developing these multiple choice items could endanger the integrity of assessment, selection, certification, and placement. In the same way, improper use of item analysis and lopsided test items could lead to the same fate by bringing about wrong award of grade and certificate. A *three-parameter logistic (3PL) model* uses ability to predict the probability of a certain response as a function of student's ability level and item properties in describing suitability of the test items as well as *nominal response (NR) model* which estimates students' propensity to endorse their preferred options are presented as a solution module to the problem. The estimated discrimination, difficulty, pseudoguessing, category difficult, and category discrimination parameters indices of items are presented for illustrations and guide to test developers (e-Center, NECO, WAEC, and JAMB).

Keywords: *logistic, discrimination, difficulty, validity, test, pseudoguessing, students*

Outlines

Introduction I

1. a comprehensive statistical tool for analysing educational test and psychological measurement scale on the onset.
2. interest in studying abilities, personality traits and other unobservable characteristics.
3. examine the relationship between individual items as relates to ability and how the group of items as a whole described the probability of a correct response on a given item as a function of both ability levels and properties of individual item

Aim and Objectives I

1. The aim is to present a framework for evaluating multiple choice items in Nigeria universities' computer based tests.
2. to apply 3PL model with its scoring format to estimate varied difficult, discrimination, and guessing parameters indices for individual item.
3. apply NR model to maximize the precision of ability estimate by making use of information contained in both correct and incorrect options to an item.

Literature Review I

- on inception, research interests were concentrated on binary IRT models.
- low-ability students select a correct response by chance Birnbaum (1968).
- observed asymptote is often lower than the chance level $\frac{1}{m}$, Lord (1974a)
- Bock (1972) proposed a NR model for nominally scored responses which were allocated to mutually exclusive, exhaustive, and unordered categories
- Item analysis should make use of statistics that would reveal important and relevant information for upgrading the quality and accuracy of multiple choice items (Ary et al, 2002)
- van der Linden and Hambleton (1997), and Baker and Kim (2004) applied IRT to educational testing in measuring students' ability using a test that consists of several items.
- two importance of item analysis were the identification of defective test items, and area where students have mastered and not mastered (Suruchi and Rana, 2015)
- a descriptive approach to item analysis of university-wide multiple examinations: The experience of a Nigeria private university (Olukoya et al, 2018).

Materials and Methods I

Materials

Data Description and Coding

- Test items were made up of 35 multiple choice items, each item had 4 options of which one option was a correct while other three options were distractors.
- For binary IRT model, data were coded as 0 = incorrect options; 1 = a correct option.
- For NRM, data was coded as $A = 1$, $B = 2$, $C = 3$, and $D = 4$

Method

Three-parameter Logistic Model

- It can be obtained from 2PL model by adding the third parameter c_i.

Materials and Methods II

- $$P_{ik}(\theta = 1|a_i, b_i, c_i, \theta_k) = C_i + (1 - C_i)\frac{e^{a_i(\theta_k - b_i)}}{1 + e^{a_i(\theta_k - b_i)}} \quad (1)$$

$$: a_i = discrimination \;\; parameter$$
$$: b_i = difficulty \;\; parameter$$
$$: c_i = guessing \;\; parameter$$
$$: \theta_k = student \;\; with \;\; ability \;\; k.$$
$$i : (i = 1, 2, ..., I)$$
$$k : (k = 1, 2, ..., N)$$

- guessing the correct answer for an item may be attempted by the test takers.
- a low trait levels having a non-zero probability of endorsing item correctly.
- It accounts for variability in item discriminating, difficult, guessing parameters.

Materials and Methods III

- the model guarantee students at the upper trait level to endorse item correctly which is not true especially for item that is difficult to endorse.

Method

Nominal Response Model

- Maximize the precision of ability estimate by making use of information contained in both correct and incorrect options to an item.
-

$$Pr\left(Y_{ik} = g|\ a_i, b_i, \theta_k\right) = \frac{\exp\{a_{ig}\left(\theta_k - b_{ig}\right)\}}{\sum_{h=1}^{G}\exp\{a_{ih}\left(\theta_k - b_{ih}\right)\}}, \theta_k \sim N(0,1) \quad (2)$$

where:
$a_i = (a_{i1}, ..., a_{ig}, ..., a_{iG})$,
$b_i = (b_{i1}, ..., b_{ig}, ..., b_{iG})$,
a_{ig} : discrimination index of category g for item i;

Materials and Methods IV

b_{ig} : difficult index of category g for item i;
θ_k :latent ability of student k.
$i : (i = 1, 2, ..., I)$ from student $k : (k = 1, 2, ..., N)$,
all items take on unordered responses, $h : (h = 1, .., g, ..., G)$.

- each item-option characteristics curve represents students ability as a function of probability of endorsing item option.

Analysis I

Analysis

Three-Parameter Logistic Model Parameters Estimates

Table 1: **Some Selected Item Guessing Parameters Estimates for Three-parameters Logistic Model**

Items	Gsg	Diff	SE	Z.Vals	P≻Z	P	95% Conf. Int.	
5	0.7758	0.3029	0.0546	14.2174	0.000	0.8649	0.6589	0.8897
7	0.3644	3.3063	0.0320	11.3828	0.000	0.3711	0.2942	0.3906
8	0.0964	2.0545	0.0259	3.7175	0.000	0.1472	0.0082	0.1059
11	0.0050	-2.7534	0.1839	0.0271	0.950	0.9778	0.0014	0.0194
29	0.0017	-0.9085	NaN	NaN	NaN	0.7772	0.0013	0.0132
Q34	0.0002	-2.5906	0.0151	0.0148	0.000	0.9722	0.0013	0.0132
DISC	1.3727							

Analysis II

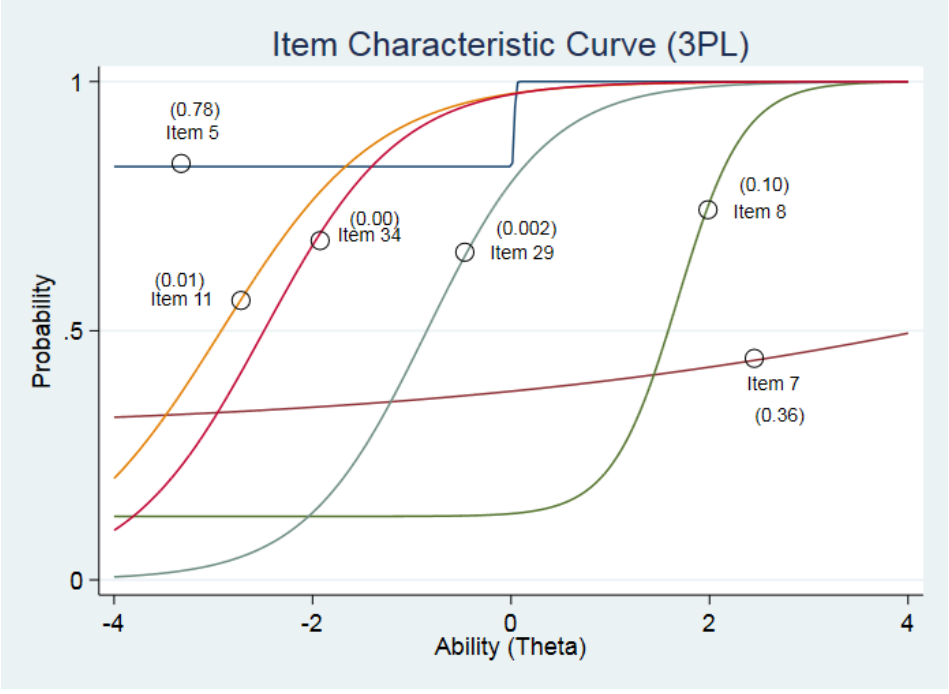

Figure 1: Item Characteristic Curve Displaying Item Psuedoguessing

- Each of the curves in Figure 1 traces the probability of correct response to individual item as a function of student ability level.
- Item 5 has no slope (no discriminating power).

Analysis III

- Its guessing index is about 0.78.
- Probability that an average student endorses correct answer is 0.8649
- It is most likely to be non-informative.
- Item 7 has a negligible slope, can not discriminate among students.
- About 0.36 is probability of guessing item 7 which is perceived to be too difficult as shown in Table 1.
- Average student endorses correct answer with probability 0.3711.
- Figure 1 suggests that most likely, items 5, and 7 are defectives, need attention of test developer.
- Probability of guessing item 34 is 0.0002, means not guessed
- Perceived to be very easy item, about 97% of the students endorse correct option.
- Its shape suggests very little discrimination, and perceives as very easy.
- Probability of guessing correct answer to item 8 is about 0.10.

Analysis IV

- Though is perceived to be difficult, only about 15% of the students answer it correctly.
- Its shape suggests that this item most likely to discriminate students on the upper level of ability scale.
- Probability of average student guessing item 29 answer correctly is 0.0017, most unlikely to guess.
- Perceived as moderately difficult, and most likely to discriminate students' ability.
- Item 11 is perceived as a very easy item, about 98% of the students answer it correctly.
- Unlikely to be guessed by the student.

Table 2: Estimated NRM Parameters

Item		Coef.	Std.Err	z	p >|z|	95% conf.	Int
Q5 Disc	2vs1	0.5820	0.4791	1.21	0.224	-0.3570	1.5210
	3vs1	0.1431	0.4903	0.29	0.770	-0.8179	1.1041
	4vs1	0.6426	0.2307	2.79	0.005	0.1905	1.0948
Diff.	2vs1	6.8316	5.2759	1.29	0.195	-3.5090	17.1722
	3vs1	22.9760	95.6224	0.29	0.770	-159.4404	215.3924
	4vs1	3.7306	1.2033	3.10	0.002	1.3722	6.0890
Q7 Disc	2vs1	0.1162	0.1380	0.84	0.400	-0.1543	0.3867
	3vs1	0.1245	0.1908	0.65	0.518	-0.2504	0.4973
	4vs1	-0.0561	0.1919	-0.29	0.770	-0.4321	0.3200
Diff.	2vs1	0.8147	1.3960	0.58	0.560	-1.9214	3.5508
	3vs1	8.9211	13.768	0.65	0.517	-18.0636	35.9059
	4vs1	-20.7803	70.8478	-0.29	0.769	-159.6394	118.0788
Q8 Disc	1vs4	0.6266	0.1782	3.52	0.000	0.2773	0.9758
	2vs4	1.0253	0.2461	4.17	0.000	0.5429	1.5078
	3vs4	1.2854	0.2888	4.45	0.000	0.7193	1.8517
Diff.	1vs4	-2.0198	0.5149	-3.92	0.000	-3.0291	-1.0106
	2vs4	0.1346	0.2098	0.64	0.521	-0.2766	0.5458
	3vs4	0.5041	0.2075	2.43	0.015	0.0974	0.9108
Q11 Disc	2vs1	0.5797	0.3774	1.54	0.125	-0.1600	1.3193
	3vs1	40.4740	2027.967	0.02	0.984	-3934.269	4015.217
	4vs1	3.3384	1.7409	1.92	0.055	-0.0738	6.7505
Diff.	2vs1	5.9523	3.6086	1.65	0.099	-1.1205	13.0251
	3vs1	2.3669	16.2313	0.15	0.884	-29.4458	34.1797
	4vs1	2.4067	0.4136	5.82	0.000	1.5961	3.2174
Q29 Disc	1vs4	1.2201	0.2559	4.77	0.000	0.7186	1.7216
	2vs4	1.8117	0.4303	4.21	0.000	0.9683	2.6551
	3vs4	2.2877	0.5126	4.46	0.000	1.2830	3.2923
Diff.	1vs4	1.4301	0.2525	5.66	0.000	0.9352	1.9250
	2vs4	1.7821	0.2788	6.39	0.000	1.2356	2.3287
	3vs4	1.7214	0.2138	8.05	0.000	1.3024	2.1404
Q34 Disc	1vs2	0.5540	0.5399	1.03	0.305	-0.5042	1.6122
	3vs2	0.8443	0.5721	1.48	0.140	-0.2769	1.9656
	4vs2	2.3077	0.7106	3.25	0.001	0.9149	3.7005
Diff.	1vs2	7.5736	6.9932	1.08	0.279	-6.1328	21.2800
	3vs2	5.1626	3.0945	1.67	0.095	-0.9025	11.2276
	4vs2	2.4270	0.3617	6.71	0.000	1.7182	3.1359

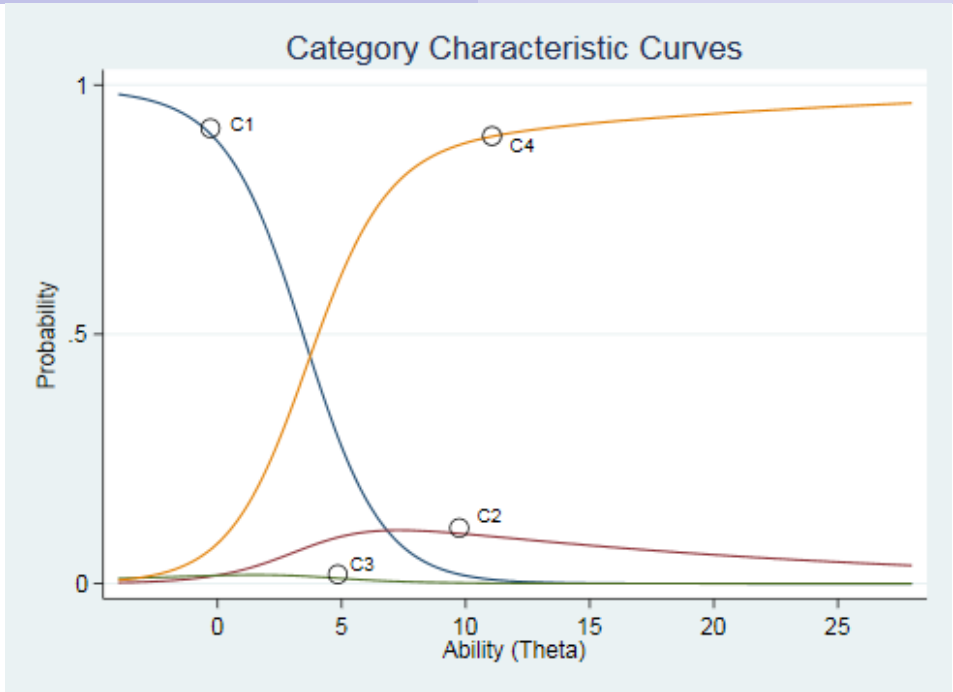

Figure 2: Item 5 Category Characteristic Curves

- Each of the CCCs trace the probability of category response as a function of student level of ability.
- C_1 in item 5 is set as base outcome.
- Difficulty indices represent points at which the base outcome (C_1) intersects with other categories in Figure 2.
- Category 4 is the most discriminating.
- The probability of endorsing C_1 over C_2 is 6.8316, that is $(2v1 = 6.8316)$

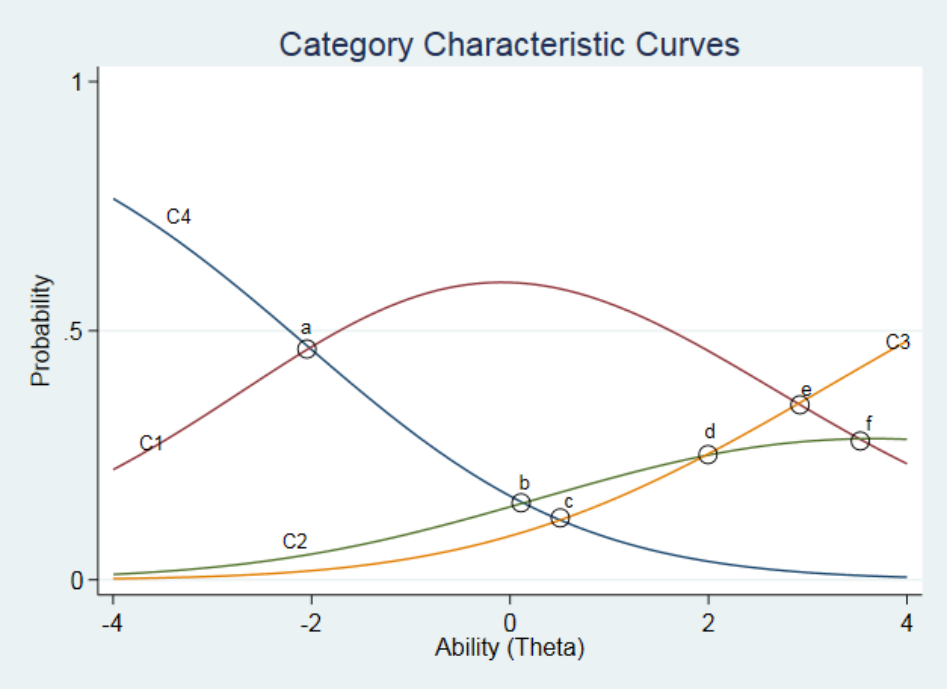

Figure 3: Item 8 Category Characteristic Curves for NRM

- Category 4 is set as base outcome here.
- The propensity of endorsing C_2 over C_4 is 0.1346, that is $(2v4 = 0.1346)$
- C_3 is the most discriminating.
- Perceived category difficult indices for categories are marked a, b, and c.
- Dominant categories are C_1, and C_4 as displayed in Figure 3.
- Similar interpretations apply to other items.

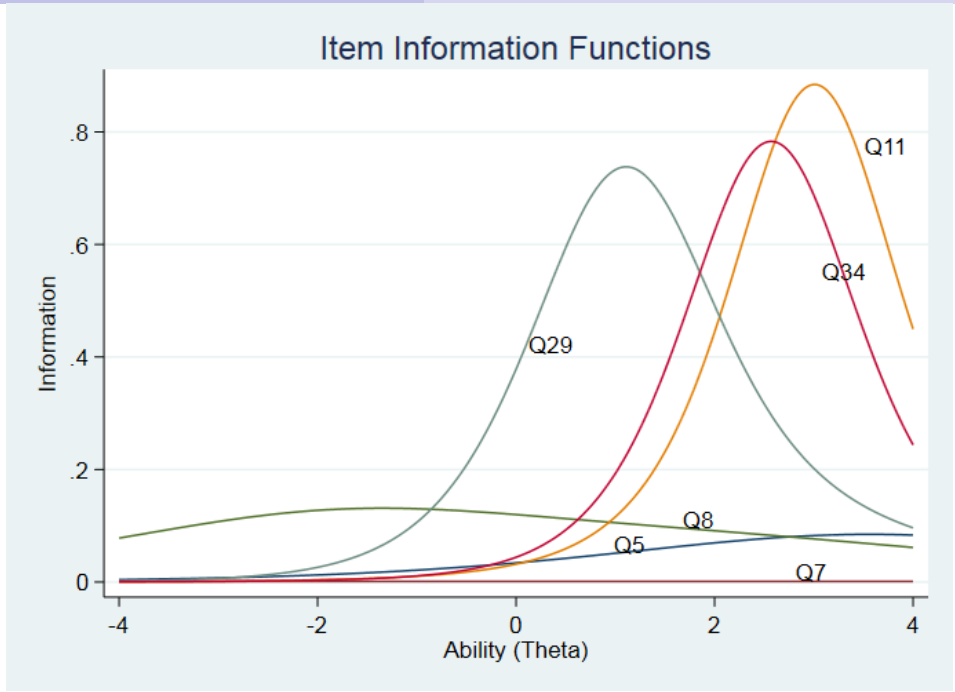

Figure 4: Item Information Functions for Items 5, 7, 8, 11, 29, and 34

- Performances of discuss items are shown graphically in Figure 4.
- Shape of individual item suggests its suitability in terms of information provided.
- Item 29 provided a balanced information on both sides of ability continuum.
- 5, 7, and 8 are suggested to be defectives based on NRM algorithm.
- They are to be discard and replaced.
- 11, and 34 are needed to be critically examined, moderated, and corrected.

Findings, and Conclusions I

A careful use of the IRT statistics tools presented in tables and figures revealed that some items and options need to be either reframed or removed and replaced in order to upgrade the quality and accuracy of MCQ items as follows.

- Item 5 needs to be reframed or changed.
- Item 7 needs to be removed and replaced by another item.
- Items 8, 11, and 34 need moderation, corrections, and reframe (See Figures 1 and 4).
- Item 29 is satisfactory.

Recommendations I

The importance of item analysis is the identification of defective items, as a potent tool in checking flaws in items and finding ways of correcting these flaws before finally administered (Eli-Uri and Malas, 2013). The recommendations to our examination bodies and universities are as follow:

- Item analysis must made compulsory.
- Item moderation .
- a legislation compelling examination bodies, and test developers to have test analysis department.
- Good item analysis culture will boost integrity of assessment, selection, certification, and placement thereby reduce lopsided test items and wrong award of grade and certificate.

Acknowledgement I

I thank the entire staff, Department of Statistics and Mathematical Sciences, College of Pure and Applied Sciences, Kwara State University Malete for granting me access and permission to use the data for illustrative purposes.

References I

- Ary, D., Jacobs, L. C., Razavieh, A. (2002). Introduction to Research in Education, 6th edition edn, Wadsworth, Califonia.
- Baker, F. B., and Kim, S. H. (2004). Item Response Theory: Parameter Estimation Techniques (2nd ed.). New York: Marcel Dekker.
- Birnbaum, A. (1968). Some Latent Trait Models and Their Uses in Inferring an Examinee's Ability. In F. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores. Reading, Mass, Addison-Wesley.
- Bock, R. D.,(1972), Estimating Item Parameters and Latent Ability when Responses are Scored in Two or More Nominal Categories, Psychometrika, **37**, 29-51.
- Eli-Uri, F.I., and Malas, N. (2013). Analysis of Use of Single Best Answer Format in an Undergraduate Medical Examination. Qatar Med.J. **1**, 3-6.
- Lord, F. M. (1974a). Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 39, 247-264.

References II

Olukoya, J., A., Adekeye, O., Igbinoba, A., O., and Afolabi, A. (2018). Item Analysis of University-Wide Multiple Examinations: The Experience of a Nigeria Private University. Qual Quant (2018) 52: 983-997, https.//doi.org/10.1007/s11135.017-0499-2.

Suruchi,Rana, S. R. (2015). Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in biology. Indian J. Res. 3(6), 56- 58.

van der Linden, W. J., and Hambleton, R. K.(Eds.).(1997). Handbook of Modern Item Response Theory. New York: Springer.

Greetings

THANK YOU FOR LISTENING