International Journal of Operational Research in Management, Social Sciences & Education | IJORMSSE
p-ISSN: 2536-6521 | e-ISSN: 2536-653X
Volume 5 Number 2 October, 2019

Examining the Relationship Between 1-PL of Rasch and 3-PL Models of IRT in Item Selection in a Constructed Test for Assessment

[1]Akuche, Ukamaka E. & [2]Aliyu, R. T.
Department of Science Education (Measurement and Evaluation Unit),
Faculty of Art and Education, Lead City University, Ibadan

Abstract

Rasch Measurement Model is a probabilistic model used to examine and validate the psychometric proportion of measurement instruments and test forms. This study used the Rasch Measurement Model and 3-parameter logistic model of item response theory to examine, validate and analyze person and instrument relating to the Physics Aptitude Test (PAT). A 50 items instrument with a reliability value of 0.82 was developed by the researchers using Classical Test Theory (CTT). The infit and outfit of the mean square score (MNSQ) and standardized score (ZSTD) of fitness of Winsteps and 3-PL model of Bilog-Mg3 were used to investigate how well the PAT fit the Models. Eventually, forty-three (43) items whose parameters were known scaled through the Rasch model and were confirmed to measure the same construct (uni-dimensionality) while 11 items were not significant and fit into the 3-PL model at $p< 0.05$. Rasch shows that only 43 items fit into the model while 3-PL shows that 11 items fit into its model. This shows that a great disparity occurs between Rasch and 3-PL model, this could be as a result of the number of test items and the sample size used from the population. The study shows the hierarchy of items which are difficult and easy to attempt by students based on the line of inquiry of the logit model and the replicability of the test items and therefore recommends the use of Rasch model over 3-PL model since item fit shows unidimensionality of the test hence banking the calibrated items for reference and future use.

Keywords: *Item Response Theory (IRT), Rasch model, 3-PL model, Physics Aptitude Test (PAT)*

Corresponding Author: Akuche, Ukamaka E.

Background to the Study

It is generally recognized that examinations determine the extent to which educational goals have been achieved as well as the extent to which educational institutions have served the needs of community and society (Shah, 2002). Examinations are not limited to measure educational or societal goals and needs but blend in a way of coping with the educational system (Havens, 2002). Rehmani (2003) opines that, examinations play a significant role in determining what goes on in the classroom in terms of what, and how teachers teach and students learn and can have an impact on both teaching and learning. Wikipedia used test or examinations as alternative terms of assessment and defined it as; test or an examination (or exam) is an assessment indeed to measure a test-takers knowledge, skill, aptitude, physical, fitness or classification in many other topics.

The psychometric methods that allow the scores of test-takers attempting different sets of items to be compared directly are based either on the Classical Test Theory model, (Ogbebor, 2017), Rasch model (Odili, Osadebe, & Aliyu, 2015) or on item response theory (IRT) models (Wagner-Menghin & Mater, 2013). The Rasch model postulates that the probability of a person giving a correct response to an item is governed only by the person's ability and the item's difficulty, both of which can be represented as locations on the same underlying measurement scale. A Person's ability is estimated from that individual's response to a set of items with previously estimated difficulties.

Rasch Model also known as one parameter model uses only a single parameter, namely item difficulty to estimate an unobservable trait of a particular examinee. The two-parameter and three-parameter models are widely used especially in large scale assessment (Downing, 2003 and Odili, Osadebe, & Aliyu, 2015). The two models add item discrimination and guessing parameters to the item difficulty.

Model appropriateness is determined by the type of test items and their scoring (Aliyu, 2015). But, in practice, the choice of models depends on the amount of data available. The larger the number of the parameter, the more data are needed for parameter estimation, thus requiring more complex calculation and interpretation. In this situation, The Rasch Model has some special properties that make it attractive to users. Rasch Model involves fewest parameters; therefore, it is easier to work with (Aliyu, 2013). Wright (1990) gives a more influential explanation in favor of the Rasch Model compared to a three-parameter model. These two models are opposite in philosophy and practice. The three-parameter model will adjust to adapt whatever type of data (includes invalid responses). The Rasch model, however, has tight standards in controlling the data. Unlike the three-parameter model, invalid responses such as guessing on an item will not be accepted. It is described as an unreliable person reliability. Critics of the Rasch Model often regard the model as having strong assumptions that are difficult to meet. However, these are values that make the Rasch Model more appropriate in practice than the two and the three-parameter models.

In any mathematical model, it is important to assess the fit of data to the model. If item misfit with any model is diagnosed as due to poor item quality, for example confusing distractors in a multiple-choice test, then the items may be removed from that test form and rewritten or replaced in future test forms. If, however, a large number of misfitting items occur with no apparent reason for the misfit, the construct validity of the test will need to be reconsidered for curriculum development and the test specifications may need to be rewritten. Thus, misfit provides invaluable diagnostic tools for test developers, allowing the hypotheses upon which test specifications are based to be empirically tested against data. Assessment is an essential component of learning and teaching, as it allows the quality of both teaching and learning to be judged and improved. It determines the priorities of education, influences practices and affects learning in general. Changes in curricula and learning objectives are ineffective if assessment practices remain the same as learning and teaching tend to be modelled against the test. To this end, the researchers want to examine the relationship between the Rasch model and the 3-PL model of IRT.

There are several methods of assessment for assessing fits, such as a chi-square statistic, or a standardized version of it. Two and three-parameter IRT models adjust item discrimination, ensuring improved data-model fit, so fit statistics lack the confirmatory diagnostic value found in one-parameter models, where the idealized model is specified in advance.

Data should not be removed based on misfitting the model, but rather because a construct relevant reason for the misfit has been diagnosed. One parameter IRT measures are argued to be sample-independent and are not population independent, so misfit such as this is construct relevant and does not invalidate the test or the model. Such an approach is an essential tool in instrument validation. In two and three-parameter models, where the psychometric model is adjusted to fit the data, future administrations of the test must be checked for fit to the same model used in the initial validation to confirm the hypothesis that scores from each administration generalize to other administrations. If a different model is specified for each administration to achieve a data-model fit, then a different latent trait is being measured and test scores cannot be argued to be comparable between administrations.

Objectives of the Study

The main objective of this study is to examine the relationship between the Rasch and 3-PL model of IRT using the PAT items. The specific objectives are to:

i. Find out the difficulty index of each item in the constructed Physics Aptitude Test (PAT) using the Rasch model
ii. Determine the difficulty index of each item in the constructed Physics Aptitude Test (PAT) using the 3-PL model
iii. Find the total number of items that fit into the Rasch model and 3-PL models of IRT

Research Questions

The following research questions were used for this study:

I. What are the difficulty indices of each item in the constructed Physics Aptitude Test (PAT) items using the Rasch model?

ii. What are the difficulty indices of each item in the constructed Physics Aptitude Test (PAT) items using the 3-PL model of IRT?
iii. What is the total number of PAT items that fit into the Rasch model and 3-PL of IRT?

Research Method & Design

This study focuses on the relationship between the Rasch and 3-PL in a developed multiple choice Physics Aptitude Test for curriculum development. The instrumentation research design was adopted.

The target population for this study consists of all senior secondary school two students (SSII) in Oyo State. Ten (10) senior secondary schools were sampled. The simple random sampling techniques of balloting were used for the selection of the ten (10) senior secondary schools. The sample size for the study was 755 respondents with 75 testees each from nine schools using non-proportionate stratified random sampling technique while 80 was taking from one out of the ten selected secondary schools.

Instrument of the Study

The Physics Aptitude Test (PAT) developed by the researcher contained 100 items. The test content consists of three components. Test content was based on a well-designed Test Blue Print convening the six levels of the cognitive domain of learning. It consists of three components of aptitude test which include: Verbal Aptitude test with the highest number of fifty (30) items; Abstract Aptitude Test which contains forty-three (27) items and Numerical/Quantitative Aptitude Test with fifty-seven (43 items). This shows how the 100 test items in the PAT were distributed among the content areas as well as the instructional objectives.

A total of 50 items that formed the PAT were drawn using the Classical Test Theory (CTT) procedure after the experimental try-out and revision of the test items. The difficulty and the discrimination indices found were used in selecting a total of fifty test items.

Reliability of the Instrument

The reliability of the PAT was established with the use of Kuder-Richardson formula 20 (KR-20). The calculated coefficient of reliability was 0.82 which indicated that the test items could be administered to the targeted audience. The research questions were analyzed using Winsteps and BILOG-MG3 statistical software to determine the: difficult level of PAT using the Rasch and 3-PL models of IRT. In WINSTEPS, the measures are determined through iterative calibration of items using the PAT. Research questions 1 was answered using the winsteps software, research question 2 was answered using the Bilog-Mg3 software while question 3 was answered using both software.

Analysis and Presentation of Result

The results obtained in this study are presented and discussed here. The Winsteps 3.75.0 and Bilog-Mg3 were used to answer the research questions. The following are the stated research questions:

Research Question 1: What is the difficulty index of each item in the constructed Mathematics Aptitude Test (PAT) using the Rasch model?

Table 1: Difficulty indices of PAT using infit and outfit of MNSQ and ZSTD indices of Rasch

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	Item
42	68	755	1.28	.13	1.04	.4	1.39	2.8	.03	.17	91.2	91.1	I0042
19	94	755	.90	.11	.98	-.2	.93	-.6	.22	.19	87.8	87.7	I0019
6	96	755	.88	.11	.99	-.1	1.01	.2	.19	.19	87.5	87.4	I0006
9	114	755	.67	.10	1.03	.4	1.17	1.9	.12	.20	85.1	85.0	I0009
43	121	755	.59	.10	1.06	.9	1.19	2.2	.08	.20	83.8	84.1	I0043
37	130	755	.50	.10	1.04	.6	1.05	.6	.15	.21	82.9	83.0	I0037
35	134	755	.46	.10	1.03	.5	1.11	1.4	.14	.21	82.9	82.4	I0035
18	135	755	.45	.10	.93	-1.2	.90	-1.3	.32	.21	82.8	82.3	I0018
44	139	755	.42	.10	1.02	.3	1.04	.6	.17	.21	81.8	81.8	I0044
33	147	755	.34	.09	1.01	.2	1.01	.2	.20	.22	81.0	80.8	I0033
34	147	755	.34	.09	1.04	.8	1.05	.8	.15	.22	80.5	80.8	I0034
48	149	755	.33	.09	1.01	.2	1.05	.7	.19	.22	81.6	80.5	I0048
7	157	755	.26	.09	1.05	1.0	1.17	2.5	.10	.22	79.4	79.5	I0007
30	157	755	.26	.09	1.12	2.2	1.15	2.2	.03	.22	77.9	79.5	I0030
11	161	755	.22	.09	.99	-.3	.98	-.2	.24	.22	79.2	79.0	I0011
47	161	755	.22	.09	1.06	1.2	1.07	1.1	.13	.22	77.6	79.0	I0047
31	166	755	.18	.09	1.01	.2	1.03	.4	.21	.22	78.8	78.4	I0031
49	166	755	.18	.09	1.03	.7	1.04	.6	.17	.22	79.0	78.4	I0049
16	167	755	.17	.09	.98	-.3	1.04	.6	.23	.22	78.4	78.2	I0016
41	167	755	.17	.09	1.01	.2	1.01	.1	.21	.22	78.6	78.2	I0041
14	168	755	.16	.09	.93	-1.4	.88	-2.0	.34	.22	78.2	78.1	I0014
50	168	755	.16	.09	.93	-1.4	.88	-2.0	.34	.22	78.2	78.1	I0050
29	178	755	.09	.09	1.02	.4	1.01	.1	.20	.23	77.1	76.8	I0029
46	183	755	.05	.09	.98	-.5	1.00	.0	.25	.23	76.9	76.2	I0046
20	187	755	.02	.09	1.03	.7	1.10	1.7	.16	.23	75.6	75.7	I0020
38	187	755	.02	.09	1.00	-.1	.98	-.3	.24	.23	76.1	75.7	I0038
45	191	755	-.01	.09	1.02	.5	1.14	2.5	.18	.23	75.9	75.2	I0045
15	197	755	-.06	.09	.91	-2.3	.87	-2.6	.38	.23	75.6	74.5	I0015
26	198	755	-.07	.09	1.01	.4	1.02	.4	.21	.23	74.7	74.4	I0026
22	200	755	-.08	.09	1.05	1.2	1.03	.6	.17	.23	72.3	74.1	I0022
17	203	755	-.10	.08	.96	-1.1	.93	-1.3	.31	.23	74.7	73.7	I0017
1	211	755	-.16	.08	1.01	.2	.97	-.6	.24	.24	71.0	72.7	I0001
12	211	755	-.16	.08	.94	-1.7	.94	-1.3	.33	.24	73.9	72.7	I0012
40	211	755	-.16	.08	1.01	.4	1.04	.7	.21	.24	71.8	72.7	I0040
36	213	755	-.17	.08	1.03	.9	1.03	.7	.19	.24	72.5	72.5	I0036
21	215	755	-.19	.08	.99	-.2	.97	-.6	.26	.24	71.0	72.3	I0021
10	217	755	-.20	.08	1.03	.9	1.04	.9	.18	.24	70.7	72.0	I0010
32	218	755	-.21	.08	.95	-1.4	.92	-1.8	.33	.24	72.9	71.9	I0032
13	219	755	-.21	.08	1.00	.0	.99	-.2	.24	.24	72.0	71.8	I0013
23	224	755	-.25	.08	.98	-.5	1.02	.4	.26	.24	70.0	71.1	I0023
5	226	755	-.26	.08	.96	-1.0	.95	-1.1	.30	.24	71.5	70.9	I0005
39	248	755	-.41	.08	.99	-.2	.98	-.6	.26	.25	68.0	68.3	I0039
4	284	755	-.63	.08	1.00	-.1	.98	-.7	.26	.25	63.7	64.7	I0004
2	290	755	-.66	.08	1.01	.4	.99	-.4	.25	.25	61.0	64.1	I0002
24	303	755	-.74	.08	.96	-1.9	.94	-2.0	.32	.25	67.2	63.1	I0024
25	322	755	-.85	.08	.99	-.3	1.00	.0	.26	.26	64.3	61.8	I0025
27	323	755	-.86	.08	.99	-.6	.98	-.7	.28	.26	61.8	61.8	I0027
28	323	755	-.86	.08	.96	-2.1	.96	-1.5	.32	.26	64.2	61.8	I0028
3	341	755	-.96	.08	.92	-4.5	.91	-3.8	.39	.26	71.8	60.9	I0003
8	359	755	-1.07	.08	.96	-2.0	.97	-1.4	.31	.26	65.0	60.3	I0008
MEAN	195.9	755.0	.00	.09	1.00	-.2	1.02	.0			75.5	75.1	
S.D.	65.4	.0	.49	.01	.04	1.1	.09	1.4			6.8	7.3	

In answering the RQ 1, Winsteps software programme was used to calibrate the responses of the 755 testees to the 50 PAT items. The table 1 above shows the difficulty indices in the fourth column, item 42 is the most difficult item in the test. The difficulty of this item is estimated to be 1.28logits with the standard error of 0.13 while item 8 is the easiest with -1.07 logits and standard error of 0.08.

Research Question 2: What is the difficulty index of each item in the constructed Mathematics Aptitude Test (PAT) using the 3-PL model?

Table 2: Estimates of b, a and c parameter of PAT |

ITEM	INTERCEPT S.E.	SLOPE(a) S.E.	THRESHOLD(b) S.E.	LOADING S.E.	ASYMPTOTE(c) S.E.	CHISQ (PROB)	DF
ITEM0001	-1.093	0.410	2.663	0.380	0.159	118.8	6.0
	0.220*	0.098*	0.490*	0.091*	0.041*	(0.0000)	
ITEM0002	-0.977	0.444	2.200	0.406	0.254	106.7	7.0
	0.278*	0.126*	0.465*	0.115*	0.054*	(0.0000)	
ITEM0003	-0.624	1.193	0.523	0.766	0.167	144.0	5.0
	0.209*	0.213*	0.121*	0.137*	0.046*	(0.0000)	
ITEM0004	-3.972	2.418	1.643	0.924	0.339	36.3	6.0
	1.422*	0.900*	0.117*	0.344*	0.021*	(0.0000)	
ITEM0005	-4.097	2.810	1.458	0.942	0.243	23.4	6.0
	1.250*	0.941*	0.090*	0.316*	0.019*	(0.0007)	
ITEM0006	-3.175	0.927	3.426	0.680	0.126	16.0	7.0
	0.851*	0.365*	0.890*	0.268*	0.015*	(0.0255)	
ITEM0007	-2.131	0.507	4.206	0.452	0.188	5.4	7.0
	0.507*	0.165*	1.047*	0.147*	0.025*	(0.6132)	
ITEM0008	-0.715	0.820	0.872	0.634	0.267	61.2	7.0
	0.267*	0.182*	0.202*	0.141*	0.058*	(0.0000)	
ITEM0009	-3.807	1.144	3.327	0.753	0.153	16.1	7.0
	1.236*	0.501*	0.888*	0.329*	0.014*	(0.0239)	
ITEM0010	-3.303	1.913	1.726	0.886	0.247	26.1	7.0
	0.878*	0.522*	0.115*	0.242*	0.020*	(0.0005)	
ITEM0011	-2.907	1.582	1.838	0.845	0.172	26.1	7.0
	0.810*	0.509*	0.145*	0.272*	0.019*	(0.0005)	
ITEM0012	-1.755	1.243	1.412	0.779	0.172	27.2	7.0
	0.416*	0.295*	0.116*	0.185*	0.029*	(0.0003)	
ITEM0013	-1.792	0.575	3.116	0.499	0.248	36.7	7.0
	0.459*	0.201*	0.734*	0.175*	0.032*	(0.0000)	
ITEM0014	-4.861	3.638	1.336	0.964	0.148	26.5	6.0
	2.495*	2.068*	0.066*	0.548*	0.016*	(0.0002)	
ITEM0015	-3.541	2.561	1.382	0.932	0.189	18.4	6.0
	1.450*	1.175*	0.088*	0.427*	0.020*	(0.0054)	
ITEM0016	-3.394	2.303	1.474	0.917	0.156	4.3	6.0
	0.857*	0.625*	0.079*	0.249*	0.017*	(0.6411)	
ITEM0017	-1.412	0.659	2.144	0.550	0.178	42.0	7.0
	0.332*	0.191*	0.310*	0.159*	0.036*	(0.0000)	
ITEM0018	-3.206	2.180	1.470	0.909	0.108	8.7	6.0
	0.730*	0.536*	0.075*	0.223*	0.016*	(0.1928)	
ITEM0019	-2.410	0.834	2.891	0.640	0.102	31.7	7.0
	0.516*	0.294*	0.566*	0.226*	0.018*	(0.0000)	
ITEM0020	-2.660	0.845	3.149	0.645	0.235	15.2	7.0
	0.739*	0.337*	0.777*	0.257*	0.021*	(0.0341)	
ITEM0021	-1.231	0.487	2.527	0.438	0.183	93.5	7.0
	0.278*	0.140*	0.482*	0.126*	0.041*	(0.0000)	

ITEM0022	-1.682	0.380	4.422	0.355	0.224	59.5	7.0
	0.414*	0.120*	1.266*	0.113*	0.034*	(0.0000)	
ITEM0023	-1.899	1.204	1.578	0.769	0.213	14.9	7.0
	0.457*	0.296*	0.139*	0.189*	0.028*	(0.0369)	
ITEM0024	-1.898	1.512	1.255	0.834	0.300	4.4	6.0
	0.514*	0.380*	0.120*	0.210*	0.030*	(0.6265)	
ITEM0025	-0.981	0.715	1.373	0.581	0.278	36.8	7.0
	0.318*	0.190*	0.230*	0.155*	0.053*	(0.0000)	
ITEM0026	-2.913	0.920	3.165	0.677	0.254	9.2	7.0
	0.865*	0.383*	0.816*	0.281*	0.020*	(0.2393)	
ITEM0027	-0.760	0.672	1.130	0.558	0.231	64.1	6.0
	0.243*	0.151*	0.220*	0.125*	0.056*	(0.0000)	
ITEM0028	-1.417	1.155	1.227	0.756	0.307	21.6	6.0
	0.456*	0.341*	0.147*	0.224*	0.040*	(0.0014)	
ITEM0029	-2.887	0.782	3.692	0.616	0.232	32.4	7.0
	0.872*	0.300*	1.128*	0.236*	0.020*	(0.0000)	
ITEM0031	-2.845	0.992	2.868	0.704	0.208	15.7	7.0
	0.815*	0.413*	0.609*	0.293*	0.019*	(0.0280)	
ITEM0032	-3.588	2.393	1.500	0.923	0.232	6.6	6.0
	1.006*	0.713*	0.090*	0.275*	0.020*	(0.3569)	
ITEM0033	-3.055	0.965	3.167	0.694	0.189	15.5	7.0
	0.874*	0.398*	0.776*	0.286*	0.018*	(0.0301)	
ITEM0034	-3.317	1.078	3.077	0.733	0.194	35.0	7.0
	1.059*	0.473*	0.763*	0.322*	0.017*	(0.0000)	
ITEM0035	-3.210	1.030	3.117	0.717	0.173	14.4	7.0
	0.931*	0.434*	0.742*	0.302*	0.017*	(0.0449)	
ITEM0036	-2.014	0.584	3.452	0.504	0.254	19.6	7.0
	0.520*	0.209*	0.904*	0.180*	0.028*	(0.0065)	
ITEM0037	-3.497	0.988	3.539	0.703	0.174	18.2	7.0
	1.108*	0.400*	1.034*	0.285*	0.016*	(0.0112)	
ITEM0038	-1.758	0.535	3.288	0.472	0.203	42.2	7.0
	0.416*	0.181*	0.784*	0.160*	0.032*	(0.0000)	
ITEM0039	-2.136	0.731	2.923	0.590	0.301	11.5	7.0
	0.593*	0.280*	0.673*	0.226*	0.027*	(0.1195)	
ITEM0040	-1.390	0.642	2.166	0.540	0.188	26.4	7.0
	0.297*	0.134*	0.278*	0.113*	0.036*	(0.0004)	
ITEM0041	-3.220	0.978	3.291	0.699	0.218	8.0	7.0
	0.990*	0.412*	0.893*	0.294*	0.018*	(0.3333)	
ITEM0042	-6.178	1.527	4.046	0.837	0.110	14.4	7.0
	2.502*	0.659*	0.717*	0.361*	0.012*	(0.0439)	
ITEM0043	-5.321	0.945	5.629	0.687	0.182	7.9	7.0
	1.577*	0.382*	1.951*	0.278*	0.014*	(0.3421)	
ITEM0044	-3.135	0.904	3.469	0.671	0.182	9.4	7.0
	0.916*	0.361*	0.960*	0.268*	0.017*	(0.2249)	
ITEM0045	-2.362	0.712	3.319	0.580	0.235	14.3	7.0
	0.615*	0.269*	0.842*	0.219*	0.023*	(0.0456)	
ITEM0046	-5.611	3.543	1.584	0.962	0.203	15.1	6.0
	1.559*	1.035*	0.088*	0.281*	0.017*	(0.0193)	
ITEM0047	-2.283	0.695	3.287	0.571	0.192	22.7	7.0
	0.536*	0.196*	0.498*	0.161*	0.022*	(0.0019)	
ITEM0048	-3.587	1.281	2.801	0.788	0.193	29.1	7.0
	1.215*	0.608*	0.574*	0.374*	0.016*	(0.0001)	
ITEM0049	-4.806	1.248	3.850	0.780	0.239	9.6	7.0
	1.899*	0.532*	1.118*	0.333*	0.016*	(0.2094)	
ITEM0050	-4.861	3.638	1.336	0.964	0.148	26.5	6.0
	2.495*	2.068*	0.066*	0.548*	0.016*	(0.0002)	

* STANDARD ERROR

LARGEST CHANGE = 2.655057 1489.4 328.0 (0.0000)

PARAMETER	MEAN	STN DEV
ASYMPTOTE	0.206	0.052
SLOPE	1.270	0.849
LOG(SLOPE)	0.061	0.583
THRESHOLD	2.537	1.099

QUADRATURE POINTS, POSTERIOR WEIGHTS, MEAN AND S.D.:

To answer this research question, the BILOG MG-3 software program was used to calibrate the responses of 755 testees to the 50-items of Physics Aptitude Test. Table 2 above shows the item parameter estimates obtained using the three-parameter model (3-PL model); Difficulty indices are in column 4, i.e. the, b, threshold.

Research Question 3: What are the total number of items in the constructed Physics Aptitude Test (PAT) that fit into the Rasch and 3-PL of IRT models?

In answering the RQ 2, the infit and outfit columns for both MNSQ and ZSTD in table 1 above were equally used while Bilog MG-3 was used in table 2 for the 3-PL model. Difficulty index (b) ranged from -1.07logit to 1.28logit using Winsteps. The table indicates that 43 items fit into the Rasch model, the listed items that are not fit are item 42, 43, 7, 30, 45, 15 and 3. These seven (7) items are items that fell outside the recommended value ranging between 0.6 - 1.2 and -2 & +2 of MNSQ and ZSTD respectively. Also, all the items showed a positive correlation with the reliability of .97 which indicated a high-quality data. The forty-three items should be kept for future use while the seven (7) highlighted items should be omitted, deleted or revised because of lack of fit to the model. These items are measuring something other than the intended content and construct. Therefore, 43 items met the model assumption which was an indication of their unidimensionality. The 43 items showed the construct validity of PAT.

BILOG MG-3 software program was used to calibrate the responses of 755 testees to the 50 items of the PAT. Chi-square probability table of the Bilog MG was used in determining the fitness of the item at a 0.05 level of significance. The difficulty index (b) of the PAT items is in the fourth column on the estimates of b parameters of the PAT table above with threshold (b) highlighted. Difficulty index (b) ranged from .523 to 5.629 using the Bilog Mg3. This shows that generally, the items are difficult for the respondents. By implication, thirty-nine (39) items were scientifically and statistically significant and do not fit into the 3-PL model of IRT. Items that do fit into the model but not scientifically and statistically significant are 7, 16, 18, 24, 26, 32, 39, 41, 43, 44 and 49. Therefore, by interpretation 11 items *fit* into the 3-PL model. All item fit/misfit were determined at a 0.05 level of significance.

Discussion of Findings

Difficulty indices of the PAT items using the Rasch model

The means of the infit and outfit MNSQ was 1.00 and 1.02 respectively and the means of the infit and outfit ZSTD of -.2 and 0.0 respectively, were very close to the expected value by the model (1.00 for MNSQ and .0 for ZSTD). The most difficult item of this test is item 42 which is estimated to be 1.28logits with a standard error of 0.13 while item 8 is the easiest with -1.07logits with a standard error of .08. The standard deviation of both the infit and outfit MNSQ and ZSTD (.04 & .09 and 1.1 & 1.4,) respectively were insignificant compared with the expected value, these difference discrepancies were not too many and showed that most data demonstrated fit from the Rasch Model expectation, the seven (7) items that were not fit showed overfit to the Rasch model expectation. This showed that

the reliability of the items was very good with .97. That is, the chances that the difficulty ordering of the items is repeated if the test were given to another group is extremely high. This is because there is a widespread of difficulty in the items as the separation index is 5.31.

Item difficulty measures spread in approximately .00*logits* (from -1.07logit to 1.28logit). The mean for item difficulty was .00logit (standard error = .01logit), while the standard deviation is 0.49. The main difference in mean measures of the testees and the items indicated that the PAT targeted the testees well. Therefore, the items distribution on the scale showed that the items were adequate in accessing important features of the constructed PAT.

The separation index of the person is 1.34, which translates to a person strata index of 2.10. The person strata index indicates the number of distinct ability levels which can be identified by the test. The minimum person strata index is 2, which means that the test is capable of distinguishing at least 2 strata of persons namely, highly-ability and low-ability persons.

Difficulty indices of the PAT items using the 3-PL model of IRT

The difficulty index (b) ranged from .523 to 5.629. This shows that generally, the items are difficult for the respondents. By implication, thirty-nine (39) items were scientifically and statistically not significant and do fit into the 3-PL model of IRT and by interpretation 11 items did not fit into the 3-PL model. All item fit/misfit were determined at a 0.05 level of significance. Among the items that fit into the 3-PL model were observed not to fit into the Rasch model.

Conclusions and Recommendations

From the data analyzed and described in the study, the 50 Items constructed showed that only a few of the items scaled through the 3-PL model while a large number scaled through the Rasch model objectively. It was noted that a few of the 43 items that fit into the Rasch model were not recognized by the 3-PL model since only 11 items were recognized by the model. This implies that the Rasch and the 3-PL models have functioned differently on some of the constructed PAT items. This shows the disparity between the two models which may be as a result of sample size. According to Bergan (2010) and Aliyu & Akinoso, 2017. "In the Rasch approach, data that do not fit the theory expressed in the mathematical model are ignored or discarded. In the scientific [IRT] approach, a theory is discarded or modified if it is not supported by data. Bergan admits that "Adherence to a scientific [IRT] approach does not imply that there are no bad items. Indeed, measurement conducted under the scientific approach facilitates effective item evaluation and selection. Generally, an important aspect of the IRT approach is the selection of an IRT model to represent the data". The researcher's conclusion "is that for this assessment, the Rasch model is preferred over the 3-PL models because the model offers a significant improvement in the fit of the data to the model over the alternative models. In other words, the additional parameters estimated in the Rasch model are

justified because they help provide a better fit to the data." This could be the result of the objectivity of the Rasch in item selection of fitness. Only items, 43 and 7 were all recognized by both models. They are therefore suggested to be removed from the test instrument.

More interesting are the fit statistics for the simulated items from the Rasch analysis. All the items have acceptable fit statistics! The most under-fitting item is item 42 (highest difficulty value of Rasch) with an outfit mean-square 1.39. The most over-fitting item is item 43 (with the highest 3-PL difficulty) with an outfit mean-square value of 1.19 in Rasch. Both items (42 & 43) are underfit in Rasch showing item redundancy in the test. This shows that the item does not adequately differentiate between the high and low ability examinees. The most difficult item should be able to differentiate between high and low ability examinees, with a high discrimination value whereas item 3 with difficulty index of .523 has a higher discrimination value of 1.193 than item 43 in 3-PL. Therefore, the 3-PL model did not show the true picture of items 43 and 3 in the model. Therefore, generally, the most appropriate model (i.e. the model involving the least number of estimated parameters with objectivity measure) is preferred to represent the data" and this would motivate the selection of Rasch over 3-PL.

This paper therefore, recommends the use of the Rasch model over the 3-PL model since items fit show the unidimensionality of the test. Also, item measure order in Rasch reduces any bias of any form according to literature. It does not discriminate between samples and also, shows high content and construct validity.

References

Ahmad, Z.K. & Nordin, A. (2012). Advance in Educational Measurement: A Rasch Model Analysis of Mathematics Proficiency Test. *International Journal of Social Science and Humanity, 2*(3).

Akuche, U. E. & Aliyu, R. T. (2018). Assessment of Psychometric Qualities of Gender Performance in Basic Education Certificate Examination in Mathematics Multiple Choice Test Items. *Journal in the press of Lead City University, Ibadan*

Aliyu, R. T. & Akuche, U. E. (2019). Assessment of Differential Item Function (DIF) in Mathematics Multiple Choice Test Items in Basic Education Certificate Examination in Oyo State. *Journal of the Evaluation of the Association of Educational Researchers and Evaluators of Nigeria, 4*(1), 125-139

Aliyu, R. T. & Akinoso, S (2017). Development and validation of Mathematics Aptitude Test using the Rasch and 2-PL Models of IRT. *An unpublished Journal in the University of Lagos press.*

Aliyu, R.T. (2015). *Construct Validity of Mathematics Test Items using the Rasch Model. An International Journal of Social Science and Humanities Research, 3*(2), 22-28

Aliyu, R. T. (2015). *Development and validation of Mathematics Achievement Test using the Rasch Model.* An Unpublished Ph.D thesis in Delta State University, Abraka

Aliyu, R. T. & Ocheli, O.E. (2013). Development and Validation of College Mathematics with Item Response Theory (IRT) Models in Attaining Quality Education in Nigeria. *Journal of Educational Research and Development (DJERD), 12*(1), 130-140

Andrich, D. (1992). The application of an unfolding model of the PIRT type to measurement of attitude. *Applied Psychological Measurement, 12,* 33-35.

Baghaei, P. & Amrahi, V. (2011). Rasch Model as a construct validation tool, in *Rasch Measurement Transaction*, *22* (1),1145-1146

Bergan, J.R. (2010) *Assessing the Relative Fit of Alternative Item Response Theory Models to the Data.* Tucson AZ: Assessment Technology Inc. http://ati-online.com/pdfs/researchK12/AlternativeIRTModels.pdf

Bond, T. G. & Fox, C. M. (2001). Applying *the Rasch model: Fundamental Measurement in Human Sciences, 1st ed.* Mahwah, NJ: Lawrence Erlbaum

Chen, S.Y., Ankenmann, R.D. & Chang, H.H. (2000). A comparison of item selection Rules at the early stage of computerized adaptive testing. *Apply Psychological Measurement, 24,* 241-255

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* Orlando, FL: Holt, Rinehart and Winston Inc.

Downing, S. M. (2013). Item response theory: Applications of Modern Test Theory, *Medical Education*, *37,* 739-745.

Embretson, S.E. & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah, NJ: Lawrence-Erlbaum.

Green, K. E. & Frantom, C. G. (2002). *Survey Development and Validation with the Rasch model*. A paper presented at the international conference on questionnaire, development, evaluation and testing, Charleston, SC, November 14-17, 3-8

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston: Kluwer, Nijhoff.

Havens, A. (2002). *Examinations and learning: An Activity – Theoretical Analysis of the Relationship between Assessment and Learning.* Retrieved December 03, 2010 from http://www.leeds.ac.uk/educol/documents/00002238.htm

Linn, R. L. (2000). Assessment and Accountability. *Educational Researcher, 29* (2) 4-6

Linacre, J. M. (2012). *A user's guide to Winsteps*

Nenty, H. J. (2005). *The application of Item Response Theory in strengthening assessment role in the implementation of national education policy*.

Nitko, A. J. (1996). *Educational Assessment of Students.* The wright map. 2nd. ed. Merrill: Englewood Cliffs, NJ.

Odili, J. N., Osadebe, P. U. & Aliyu, R. T. (2015). Assessment of Stability of Item Parameter in a Mathematics Achievement Test Under the Rasch Model. *Journal of Association of Educational Researcher and Evaluators of Nigeria (ASSEREN), 1*(1), 1-8

Olaleye, O. O. & Aliyu, R. T. (2013). *Development and Validation of Mathematics Achievement Test Items Using Item Response Theory (IRT) Models in Attaining Quality Education for National Development.* A paper presented and published in the Proceedings of Mathematics Association of Nigeria (MAN) at the 50th Anniversary of the Annual National conference of MAN, 82-95

Opasina, O. C. (2009). *Development and validation of alternative to practical Physics test using item response theory model.* An unpublished Ph.D thesis, University of Ibadan.

Osadebe, P. U. (2010). *Construction and validation of Test Items.* An unpublished lecture note, Delta state university.

Rehmani, A. (2003). *Impact of public examination system on teaching and learning in Pakistan*. Retrieved December 24, 2010 from http://www.aku.edu/AKUEB/pdfs/pubexam.pdf

Reza, P., Baghaei, P. & Ahmadi, H. S. (2011). Development and validation of English Language Teacher Competency Test using Item Response Theory. *The International Journal of Education and psychological assessment, 8*(2),54-68.

Shah, J. H. (2002). *Validity and credibility of public examinations in Pakistan*. An unpublished Ph. D., in the Department of Education, Islamia University Bahawalpur, Pakistan.

Thissen, D., & Orlando, M. (2001). Test Scoring. Mahwah, NJ: Lawrence Erlbaum Associates. 3PL, Rasch, Quality-Control and Science. J.M. Linacre. *Rasch Measurement Transactions, 27*(4) 1441-4

Wang, T & Vispoel, W. (1998). Properties of Abilities Estimation Method in Computer Adaptive Testing. *Journal of Educational Measurement, 35,* 109-135

Wiberg M. (2004). *Classical Test Theory Vs Item Test Theory: An Evaluation of the Theory Test in the Swedish driving-license Test*. http:www.eedusci.umn.se/digital Assets159/5929-em-no-50p.d.f.cited 4/1/2016 1.43pm