Sir Syed Journal of Education & Social Research
Vol. 3, Issue 2, 2020 (April – June)
ISSN 2706-6525 (online), ISSN 2706-8285 (Print)
ISSN 2706-9362 (CD-ROM), ISSN 2706-6525 (ISSN-L)
DOI: https://doi.org/10.36902/sjesr-vol3-iss2-2020(426-432)

SJESR
Sir Syed Journal of Education &
Social Research

Rasch Calibration of Achievement Test: An Application of Item Response Theory

* Zunaira Fatima Syeda, PhD
** Uzma Shahzadi, Assistant Professor (Corresponding Author)
*** Ghazanfar Ali

Abstract

Test construction is a fundamental constituent in the field of education. Many researches are being conducted in the field of testing across the academic world. Development of Achievement tests and standardization of achievement tests is a prime area research for academic researchers. In continuous assessment, teachers develop tests to measure students' achievement through the duration of course or study. The main objective of this study was the development of an objective type test in the subject of Measurement and Assessment in Education. Initially, the test comprised 60 items. Test reliability through Cronbach Alpha was 0.77. The test was administered to 385 students enrolled in the department of Education. A convenient sampling technique was used to select the sample. Item analysis was conducted using item response theory (IRT). This study suggests that a parallel form test with more test items and more syllabus coverage might be constructed. Tests might be administered in teacher education departments of all public sector universities. Test responses might be used to diagnose learning difficulties in the subject of measurement and assessment. The study recommended that test construction theories might be the part of course outlines.

Keywords: Assessment, Item Response Theory, Classical Test Theory, Rasch Analysis

Introduction

Student assessment record is an important record in the educational institutes and it is noted that educational data obtained through students assessment is used for several purposes, i.e., improvement in instructional planning, changing the instructional content for students' better understanding, evaluation of learners' knowledge, skills and comparing learners' achievement data (Ojerinde,2013). Bichi and Talib (2018) stated that the assessment of effective instruction is dependent upon the quality of the test constructed and information gathered during assessment settings. Since last decades, rapid changes in instructional system is observed along with changes in assessment planning and procedures i.e., there is a shift from oral to written assessments, subjective to objective type and standardized to teacher made assessments. The change in assessment is to address the change in the paradigm shift of teaching and learning procedures and practices.

It is, therefore, assessment of the educational system is based on testing, though much advancement is being observed its design and delivery of tests. However, the quality of a test is always a need to be addressed in terms of test designing, analysis techniques and interpretation of test scores (Fatima, Tirmazi, Latif & Gardei 2015). To establish test protocols, reliability and validity are necessary to be computed. The quality of any developed test may be questionable without determining the reliability and validity of a test that can be either teacher made or standardized (Kazmi & Tirmizi,2012). Usually, literature, reports that a standardized test development involves five steps; conceptualizing a test, developing items, try-out the test, analyzing and interpreting test and lastly revising items on the basis of test interpretation (Miller, Linn & Gronlund, 2012)

Hence, item analysis follows the initial try-out of the test. Psychometric qualities of test scores are the major focus of item analysis (Boopathiraj & Chellamani, 2013). It is known that the traditional method for item analysis is the application of Classical test theory (CTT). The modern technique, that is used for item analysis is item response theory (IRT). Item response theory is known

* Department of Education, University of Sargodha, 40100, Sargodha, Pakistan
Email: neerasalman@gmail.com
** Department of Education, University of Sargodha, 40100, Sargodha, Pakistan
Email: uzma.shahzadi@uos.edu.pk
*** Department of Education, University of Sargodha, 40100, Sargodha, Pakistan
Email: gazanfar.ali@uos.edu.pk

for its advance and better analysis of the test items. It also addresses the quality of test item and can discuss persons' ranking in the said group of test takers (Deng., 2011).

Therefore, item response theory offers itself as an opulent statistical tool for educational and psychological measurement scales and ,hence, used widely both in the field of education and psychology. The IRT methods were developed in the 1960s to 1980s (MÜlberger, 2017). Binnet and Simon (1916) provided the techniques for planning items for the tests of children's' mental development. Lord (1980) discussed in his book "Statistical theories of mental test scores" a rigorous and uniformed use of IRT through different intelligence test development studies. Egberink (2010) states item response theory as a significant statistical tool that discusses about examinees' responses, test proficiency and describes how performance may be related to the test scores. It is discussed that Item responses may be discrete or continuous. There can be persons with similar or different abilities and many statistical ways are there to stable the relationship between item responses and the fundamental abilities of the individuals.

Meanwhile, Within the broad-spectrum of item response theory, many models are being applied to real test situations. Eleji and Esomonu (2018) stated that item response theory is rooted in latent trait theory. It deals with measurement assumptions about test items difficulty level, examinee performance in the test, relation of performance with knowledge and skills measured by an individual item in a test. Yalcin (2018) discussed the characteristics of IRT: first of all, IRT must specify the relationship between the observed response and the underlying unobservable construct. Secondly, the model must facilitate the ways about observed responses. Thirdly, the scores will provide a sound base to estimate the construct under observation. Lastly, an IRT model is based on the assumption that the performance of an examinee can entirely be predicted or explained by one or more abilities he or she possesses. In IRT, it is generally, assumed that each examinee has some unobservable trait (ability), which cannot be directly studied (Baker, 2001; Hambleton & Swaminathan,2013). The purpose to use IRT in research is to recommend some models that may link latent traits to some observable characteristics of the examinee, especially his/her abilities, by his/her response to a set of items (test).

In IRT, item parameters are based on three aspects of the item i.e., difficulty (location), discrimination (slope), and pseudo-guessing (lower asymptote) of the item. Three IRT models are being used mostly: Firstly, one- parametric-logistic model which only discuss item difficulty parameters. Secondly, the two-parametric-logistic model which discusses the difficulty as well as discrimination of the test and examinee both. Thirdly, Three-parametric-logistic model which not only focuses on item difficulty and discrimination but also the chance of guessing of the examinee to solve an item correctly. It leads toward ICC (item character curve), which represents the probability that examinees with low ability will respond to an item correctly in a test (Koğar & Koğar, 2015; Koçak, 2020).

In a classroom, a teacher made test need address of standardization procedures so as to reach valid performance measurement of students. In the prevailing situation of item analysis in Pakistan, CTT is the common tool for item analysis of different types of tests. Norm-referenced achievement tests are commonly used to evaluate students' knowledge and skills at elementary and secondary levels. To some extent, there is an evidence that items are analyzed through IRT at elementary and secondary levels but it is not addressed in higher education classroom assessments. This study was taken up to construct an objective-type achievement-test in the subject of 'Educational Assessment' i.e. the core course taught in teacher education programs.

Objectives of the Study

The study is carried out to address following objectives i.e., Development of an achievement test in the subject of educational assessment, Application of Item response theory to perform item analysis and Application of Rasch model to calibrate test items.

Procedure of the Study

The population of the study was all enrolled students in the departments of Education in recognized public sector universities of the Punjab (Pakistan). Convenient sampling was used to select a sample size of 385. Research instrument was an achievement test developed by researchers and comprising 60 multiple-choice items.. For the development of achievement tests, a table of specifications was prepared in the subject of educational assessment in education. The content validity of the test was addressed by the table of specification and expert opinion. The reliability of the test valued at 0.77.

Results

An answer key was prepared, keeping in view the test protocols. Test was scored by the researchers: one mark is for the correct answer and 0 is for incorrect and scores were tabulated and recorded for analysis. There was no negative marking for incorrect answers. The item calibration model "IRT" was used as a tool to analyze achievement test data. PROX item calibration and PROX person measurements were also calculated. ICC (item character curve) and PCC (person character curve) were drawn to discuss item difficulty and person measure. Method PROX was used to the solution of the probability of items and persons.

Item Calibration:

Proportion correct and incorrect of the item scores were calculated for each item separately. Logits incorrect and the mean of these Logits were calculated. The variance of distribution from this mean was the initial item calibration. Table 1 represents the detail of this final item calibration.

Table 1: PROX Item Calibration

No. of items	**Scores of items**	**Proportion of Correct items** "$P_i = \frac{S_i}{N}$**t**	**Proportion of Incorrect items** "1- P_i"	**Final Calibration** $d_i = Y \times d$	**No. of items**	**Scores of items**	**Proportion of Correct items** "$P_i = \frac{S_i}{N}$**t**	**Proportion of Incorrect items** "1- P_i"	**Final Calibration** $d_i = Y \times d$
33	223	0.58	0.42	-1.16	54	125	0.32	0.68	0.03
5	204	053	0.47	-0.94	59	124	0.32	0.68	0.03
7	189	0.49	0.51	-0.76	53	122	0.32	0.68	0.03
2	186	0.48	0.52	-0.72	27	121	0.31	0.69	0.09
29	182	0.47	0.53	-0.67	11	119	0.31	0.69	0.09
41	179	0.46	0.54	-0.63	51	117	0.30	0.70	0.15
10	177	0.46	0.54	-0.63	49	116	0.30	0.70	0.15
4	169	0.44	0.56	-0.54	42	115	030	0.70	0.15
39	164	0.43	0.57	-0.49	58	109	0.28	0.72	0.25
1	163	0.42	0.58	-0.45	26	106	0.28	0.72	0.25
56	163	0.41	0.58	-0.45	37	103	0.27	0.73	0.30
34	158	0.41	0.59	-0.40	28	93	0.24	0.76	0.48
47	158	0.41	0.59	-0.40	24	89	0.23	0.77	0.55
57	158	0.41	0.59	-0.40	20	88	0.23	0.77	0.55
3	157	0.41	0.59	-0.40	22	88	0.23	0.77	0.55
18	157	0.41	0.59	-0.40	60	87	0.23	0.77	0.55
46	156	0.41	0.59	-0.40	17	82	0.21	0.79	0.67
21	154	0.40	0.60	-0.35	25	75	0.19	0.81	0.82
55	153	0.40	0.60	-0.35	12	74	0.19	0.81	0.82
40	151	0.39	0.61	-0.30	15	74	0.19	0.81	0.82
35	149	0.39	0.61	-0.30	31	70	0.18	0.82	0.90
9	147	0.38	0.62	-0.26	13	67	0.17	0.83	0.97
44	146	0.38	0.62	-0.26	16	65	0.17	0.83	0.97
19	144	0.37	0.63	-0.21	8	63	0.16	0.84	1.05
43	143	0.37	0.63	-0.12	6	62	0.16	0.84	1.05
38	142	0.37	0.63	-0.21	14	62	0.16	0.84	1.05
32	141	037	0.63	-0.21	54	125	0.32	0.68	0.03
48	141	0.37	0.63	-0.21	59	124	0.32	0.68	0.03
50	137	0.36	0.64	-0.16	20	88	0.23	0.77	0.55
36	133	0.35	0.65	-0.11	22	88	0.23	0.77	0.55
52	131	0.34	0.66	-0.07	60	87	0.23	0.77	0.55
23	128	0.33	0.67	-0.01	17	82	0.21	0.79	0.67
45	127	0.33	0.67	-0.01	25	75	0.19	0.81	0.82
30	125	0.32	0.68	0.03	12	74	0.19	0.81	0.82
54	125	0.32	0.68	0.03	15	74	0.19	0.81	0.82
59	124	0.32	0.68	0.03	31	70	0.18	0.82	0.90
53	122	0.32	0.68	0.03	13	67	0.17	0.83	0.97
27	121	0.31	0.69	0.09	16	65	0.17	0.83	0.97

11	119	0.31	0.69	0.09	8	63	0.16	0.84	1.05
51	117	0.30	0.70	0.15	6	62	0.16	0.84	1.05
49	116	0.30	0.70	0.15	14	62	0.16	0.84	1.05
42	115	030	0.70	0.15	53	122	0.32	0.68	0.03
58	109	0.28	0.72	0.25	27	121	0.31	0.69	0.09
26	106	0.28	0.72	0.25	11	119	0.31	0.69	0.09
37	103	0.27	0.73	0.30	51	117	0.30	0.70	0.15
28	93	0.24	0.76	0.48	49	116	0.30	0.70	0.15
24	89	0.23	0.77	0.55	42	115	030	0.70	0.15

Persons' Measurement:

As discussed earlier, the total number of items in the achievement test was 60 and administered on 385 sampled students. Every person probably has either zero as a minimum score or sixty as a maximum score.

Table 2: Person Measurement

Freq.	Block Title	Possible values (Score)	Proportion of Correct items	Proportion of Incorrect items	Ability level (final)	Freq.	Block Title	Possible values (Score)	Proportion of Correct items	Proportion of Incorrect items
0	AT_1	1	0.02	0.98	-4.08	9	AT_{26}	26	0.43	0.57
0	AT_2	2	0.03	0.97	-3.66	15	AT_{27}	27	0.45	0.55
0	AT_3	3	0.05	0.95	-3.09	10	AT_{28}	28	0.47	0.53
0	AT_4	4	0.07	0.93	-2.72	6	AT_{29}	29	0.48	0.52
0	AT_5	5	0.08	0.92	-2.56	7	AT_{30}	30	0.50	0.50
0	AT_6	6	0.10	0.90	-2.31	1	AT_{31}	31	0.52	0.48
0	AT_7	7	0.12	0.88	-2.09	3	AT_{32}	32	0.53	0.47
0	AT_8	8	0.13	0.87	-2.00	2	AT_{33}	33	0.55	0.45
3	AT_9	9	0.15	0.85	-1.82	1	AT_{34}	34	0.57	0.43
1	AT_{10}	10	0.17	0.83	-1.67	1	AT_{35}	35	0.58	0.42
3	AT_{11}	11	0.18	0.82	-1.60	0	AT_{36}	36	0.60	0.40
10	AT_{12}	12	0.20	0.80	-1.46	1	AT_{37}	37	0.62	0.38
11	AT_{13}	13	0.22	0.78	-1.33	0	AT_{38}	38	0.63	0.37
18	AT_{14}	14	0.23	0.77	-1.27	0	AT_{39}	39	0.65	0.35
19	AT_{15}	15	0.25	0.75	-1.16	0	AT_{40}	40	0.67	0.33
25	AT_{16}	16	0.27	0.73	-1.04	0	AT_{41}	41	0.68	0.32
28	AT_{17}	17	0.28	0.72	-0.99	0	AT_{42}	42	0.70	0.30
33	AT_{18}	18	0.30	0.70	-0.89	0	AT_{43}	43	0.72	0.28
39	AT_{19}	19	0.32	0.68	-0.79	0	AT_{44}	44	0.73	0.27
30	AT_{20}	20	0.33	0.67	-0.75	0	AT_{45}	45	0.75	0.25
27	AT_{21}	21	0.35	0.65	-0.65	0	AT_{46}	46	0.77	0.23
23	AT_{22}	22	0.37	0.63	-0.56	0	AT_{47}	47	0.78	0.22
23	AT_{23}	23	0.38	0.62	-0.51	0	AT_{48}	48	0.80	0.20
21	AT_{24}	24	0.40	0.60	-0.43	0	AT_{49}	49	0.82	0.18
15	AT_{25}	25	0.42	0.58	-0.34	0	AT_{50}	50	0.83	0.17
0	AT_{51}	51	0.85	0.15	1.82	0	AT_{56}	56	0.93	0.07
0	AT_{52}	52	0.87	0.13	2.90	0	AT_{57}	57	0395	0.05
0	AT_{53}	53	0.88	0.12	2.09	0	AT_{58}	58	0.97	0.03
0	AT_{54}	54	0.90	0.10	2.31	0	AT_{59}	59	0.98	0.02
0	AT_{55}	55	0.92	0.08	2.56	0	AT_{56}	56	0.93	0.07

Table 2 represents the person measurement and reflects that Blocks AT1 to AT59" were allocated keeping in view maximum and minimum scores from (zero to sixty). Persons from block A1 to A59 having their probable scores from 1 to sixty, According to blocks, proportions of correct and incorrect were determined. These preliminary person measurement scores go with each possible score on the test. Further, Table 2 signifies final person measurement.

Item Characteristic Curve

The item characteristic curve provides a thorough representation of item functioning across the proficiency level of the candidates. This curve provides a relationship between the observable performance of the candidate and the magnitude of the probability of correct or incorrect responses (p). So the curve was drawn between final item difficulty (di) and the magnitude of probability (p).

Table 3: Item Characteristic Curve

	Ability Level (of Person)							
Difficulty Level (of Item)	**-3.66**	**-1.04**	**-0.29**	**-0.08**	**0.29**	**1.04**	**1.46**	**3.09**
-0.72	0.05	0.41	0.6	0.65	0.73	0.85	0.9	0.98
1.05	0.01	0.09	0.18	0.21	0.28	0.45	0.55	0.86
0.03	0.02	0.22	0.38	0.43	0.52	0.7	0.78	0.95

Table 3 reflects the values for calculated DI and P for the items and individuals. The table, further, reflects the values for item functioning across the proficiency level of the individuals participating in the study.

With the help of calculations in table 3, item characteristics curve was drawn, the curve is shown below in figure 1

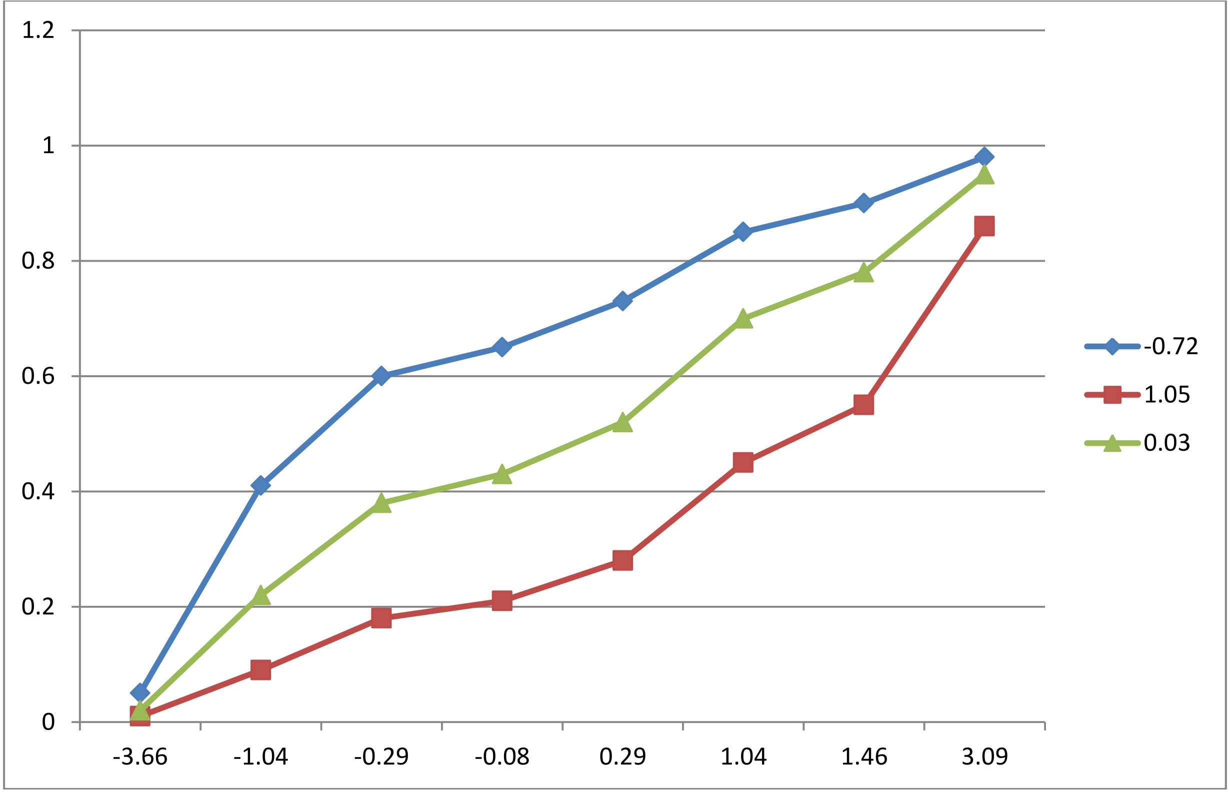

Figure 1: Item Characteristic Curve

Figure 1 represents the item characteristics curve and reflecting the steep in the middle of the curve that represents the great item discrimination of the items of the test.

Person Measurement Curve:

The curve was drawn between the person measurement value (br) and the magnitude of probability (p).

Table 4: Person Character Curve for Assessment in Education (Form A & B)

	Difficulty Level (of Item)											
Ability Level (of Person)	**-1.16**	**-0.76**	**-0.63**	**-0.40**	**-0.35**	**-0.21**	**-0.01**	**0.03**	**0.30**	**0.67**	**0.82**	**0.97**	**1.05**
2.72	0.98	0.97	0.96	0.94	0.92	0.9	0.81	0.79	0.72	0.68	0.58	0.51	0.41
-0.51	0.86	0.8	0.78	0.7	0.68	0.53	0.46	0.41	0.39	0.34	0.29	0.21	0.15
-3.66	0.71	0.51	0.45	0.39	0.34	0.29	0.23	0.2	0.18	0.11	0.09	0.04	0.03

Table 4 represents thirteen values for person measurement(br) taken randomly. The table also reflected the calculated value of the magnitude of probability. Table, further, represents the detail of the relationship.

With the help of table 4, the curve PCC was drawn to represent the relationship between person measurement (br) and magnitude of probability i.e., P

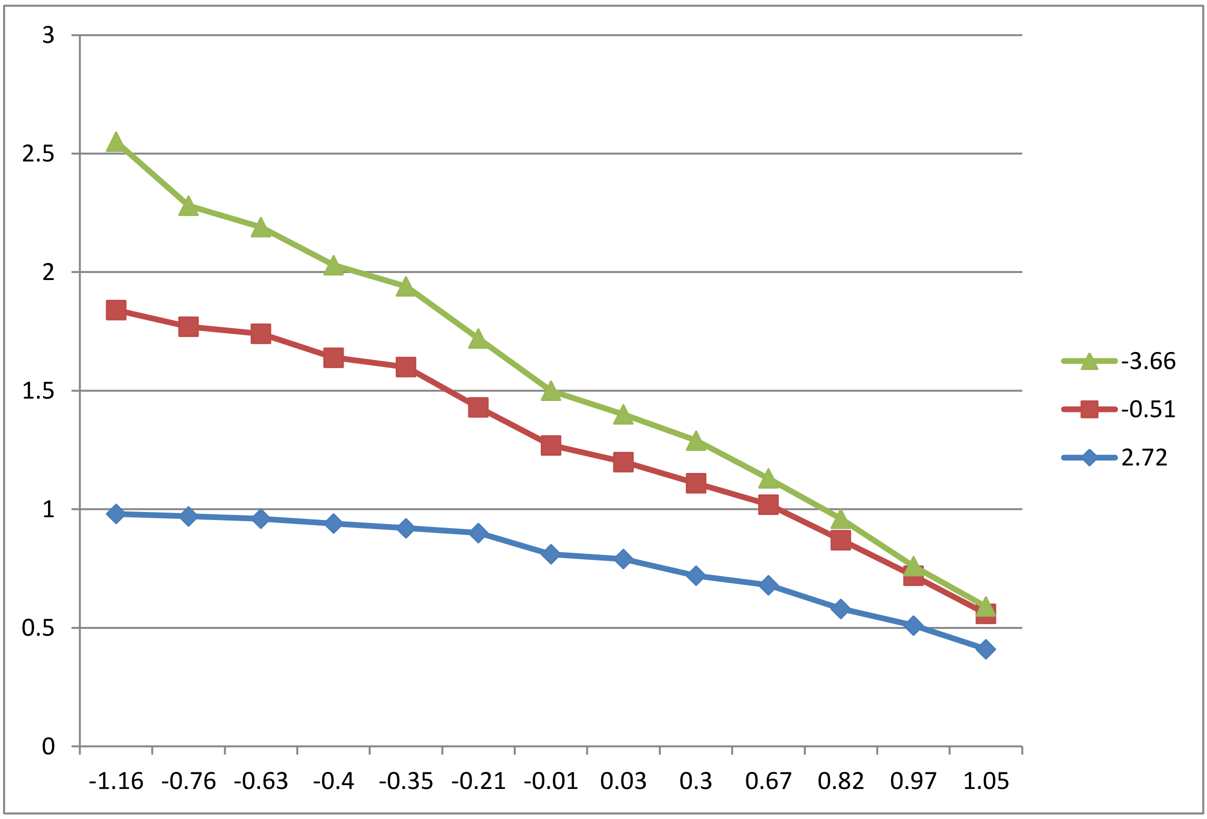

Figure 2: Person Character Curve

The curve 1 was drawn between brand p values. Curve 1 is drawn between br$_1$ and P$_1$, Curve 2 between br$_2$ and P$_{2,}$ and curve 3 between br$_3$ and p$_3$. The slope of the curve shows the discrimination between the persons of low and high ability.

Conclusions and Discussions

The basic application of the Rasch model results in items consistency on the latent continuum from the sample of respondents. Moreover, ordering item values can be compared from a defined group of persons' abilities, attitudes, and other characteristics, to ultimately define the variable. It is assumed that the items share a single dimension (Yang, 2014) or trait and the defined group is homogeneous (Utesch, Bardid, Huyben, Strauss, Tietjens, De Martelaer, & Lenoir, 2016). Previous studies mostly conducted in Pakistan calculating item calibration and person measurement discussed respondents of secondary school mostly. Due to this reason, their results were different. For example, Nafees, Farooq, Tahirkheli and Akhtar (2012) conducted a test of General Science on grade seven which showed high reliability at .22 and most of the items were the easiest items as most of the respondents were able to solve them. Major findings of the data represented that table with "item calibration" exhibited easy items had negative values while hard items depicted positive values. As item difficulty gets high it also increased from negative to positive. "Person calibration" showed the first half of the person's block had negative values While, last half had positive values. ICC discusses that probability for solving an individual item: which may have a specific difficulty level, may increase by the increase of ability level of the persons. As the difficulty level of the item increases, the probability of solving items may surely decrease. PCC showed, persons with specific abilities may have some problem if item difficulty increases. Similarly, as the ability level of the persons increases the probability of solving an item may also increase.

Recommendations

This study indicated that Rasch analysis is a useful tool to check the achievement and the current ability status of the respondents. There should be more research on test standardization. CTT and IRT should be included as a fundamental component in the course outlines of assessment in Education for teacher educators in Pakistan. IRT should be introduced to the pre-service teachers so that they can batter evaluate their self-constructed test items. This will provide a guideline for the betterment of the examination system of Pakistan. Different software of IRT may also be used to make the analysis convenient.

References

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae. net/irt/baker..

Bichi, A. A, & Talib, R. (2018). Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development. *International Journal of Evaluation and Research in Education*, *7*(2), 142-151.

Binet, A., & Simon, T. (1916). New methods for the diagnosis of the intellectual level of subnormals. (L'Année Psych., 1905, pp. 191-244).

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. *International journal of social science & interdisciplinary research*, *2*(2), 189-193.

Deng, Z. (2011). Revisiting curriculum potential. *Curriculum Inquiry*, *41*(5), 538-559.

Egberink, I. J., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality*, *44*(2), 232-244.

Eleje, L. I., & Esomonu, N. P. (2018). Test of Achievement in Quantitative Economics for Secondary Schools: Construction and Validation Using Item Response Theory. *Asian Journal of Education and Training*, *4*(1), 18-28.

Fatima, Z., Tirmizi, S. H., Latif, M. I., & Gardezi, A. R. (2015). Development and Rasch analysis of an achievement test at the master level (philosophy of education). *Pakistan Journal of Commerce and Social Sciences (PJCSS)*, *9*(1), 269-281.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media.

Kazmi, U., and Tirmizi, S., (2012), Calibration of Parallel Forms of an Achievement Test in Applied Psychology at B. Ed Level, Pakistan Journal of Social Sciences (PJSS), 33(1).

Koçak, D. (2020). Investigation of Rater Tendencies and Reliability in Different Assessment Methods with Many Facet Rasch Model. *International Electronic Journal of Elementary Education*, *12*(4), 349-358.

KOĞAR, E. Y., & KOĞAR, H. Investigation of scientific literacy according to different item types: PISA 2015 Turkey sample. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *19*(2), 695-709.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Miller, M. D., Linn, R. L. & Gronlund, N. E. (2008). Measurement and assessment in teaching. 10thedition. New Jersey: Pearson Education, Inc., pp. 1–287

Nafees, M., Farooq, G., Tahirkheli, S. A., & Akhtar, M. (2012). Effects of instructional strategies on academic achievement in a high school general science class. *International Journal of Business and Social Science*, *3*(5).

Ojerinde, D. (2013). Classical test theory (CTT) vs item response theory (IRT): an evaluation of the comparability of item analysis results. *A guest lecture presented at the Institute of Education, University of Ibadan on 23rd May*.

Utesch, T., Bardid, F., Huyben, F., Strauss, B., Tietjens, M., De Martelaer, K.,. & Lenoir, M. (2016). Using Rasch modeling to investigate the construct of motor competence in early childhood. *Psychology of Sport and Exercise*, *24*, 179-187.

Yalcin, S. (2018). Determining Differential Item Functioning with the Mixture Item Response Theory. *Eurasian Journal of Educational Research*, *74*, 187-205.

Yang, F. M. (2014). Item response theory for measurement validity. *Shanghai Archives of Psychiatry*, *26*(3), 171.