Journal of Education and Vocational Research
Vol. 3, No. 3, pp. 98-106, Mar 2012 (ISSN 2221-2590)

Rasch Calibration of General Science Test at Grade-VIII in Pakistan

*Muhammad Javid Qadir[1], Iram Gul Gilani[1], Abdul Hameed[2]
[1]Bahauddin Zakarya University Multan, Pakistan
[2]School of Social Sciences and Humanities, University of Management and Technology, Lahore, Pakistan
*javid_qadir123@yahoo.com

Abstract: The present study aimed at calibration of General Science achievement test for grade (VIII) through Rasch Model. For this purpose a General Science achievement test comprising 45 items was constructed from the text book of General Science for class VIII. Finally the test was administered to 300 students (M/F) in different high schools for boys and girls in Multan District. The answer sheets were scored and results were tabulated. Eleven (11) items were rejected on the basis of F, D and ϕ. Fifteen (15) items were to be improved on the basis of F, D and ϕ. Remaining all items were good items. Rasch Model indicates that overall test is good to measure the achievement of the students class (VIII) in the subject of General Science. On the basis of findings, major conclusions were drawn: One item was rejected on the basis of facility index (F). Twelve (12) items need improvement on the basis of facility index (F). Thirty two (32) items were very good items on the basis of facility index (F). Seven items were rejected on the basis of discrimination index (D). One item needed improvement on the basis of discrimination index (D). Thirty seven (37) items were good items on the basis of discrimination index (D). Three items were rejected on the basis of phi-co-efficient (ϕ). Two items needed improvement on the basis of phi-co-efficient (ϕ). Forty items were good on the basis of phi-co-efficient (ϕ). Test has high positive test reliability value.22 distracters were to be rejected as attempted by less than 5%.The distracters D(27), B(28), A(31), D(37), B(41), A(43) and B(10) were distracters attracted by high achievers more than low achievers, so they were rejected.

Keywords: *Calibration, General Science, Achievement test, Rasch, Item analysis*

1. Introduction

Science is knowledge about the structure and behavior of the natural and physical world, based on facts that one can prove by experiments. It is a progressive activity that constitutes a world view and permeates almost every aspect of modern life. It offers a method by which the universe and all the beings therein, may be examined to discover the artistry in God's creation, thereby communicating it to mankind. Bhatt and Sharma (1993) have described science as "The knowledge, tested, controlled and authoritatively approved."Because of the important nature of science, it has been given the status of a compulsory subject in the educational system in the form of General Science. So it is being taught even to the students of Arts, Literature and Social Sciences. It is due to this large scale demand of science teaching that we need a large number of science teachers. These science teachers must be equipped with an in depth knowledge, with full understanding and well developed skills to communicate the concepts, principles, theories and laws of science at elementary and secondary level. They must have competence and command over the content of science from class I-X as well as over the latest available methods, techniques and approaches towards science teaching. They must keep in mind the objectives of science teaching and its need and importance in the advancement and progress of the society and civilization.The objectives of science teaching are different than that of other subjects, therefore the purposes of science examination is also different. Das (1985) described: "In science examination, the purpose of a question is to test the student's knowledge, skill and understanding of science concepts and not his handwriting spellings, grammar or his linguistic ability. These are not to be assessed in science examination."The aims and objectives of education are formulated in every society. For achieving the set objectives, specific curriculum and specific methods are used. Thus a nation is quite entitled to ask whether schools are indeed equipping young people with the knowledge, skills and attitudes necessary for life in the modern world.

Knowledge of science at Elementary level provides base for higher education. General Science is taught as compulsory subject from class I. Due to its importance as a base for further education and as being a science graduate researcher decided to conduct a study. Many studies have been conducted related to evaluation and assessment but the subject of "General Science" at elementary level is given no importance in educational research especially in Pakistan. The purpose of the study was to practice a better, new and useful approach to the measurement of students' achievement. So the researcher used the simple logistic Rasch model for test calibration. As two parameters (person ability and item difficulty) have been identified r knowledge the present study was designed.This research study was focused on traditional and Rasch analysis of achievement test in the subject of general science at grade eight.Followings were the major objectives of the study:

- To construct an achievement test in the subject of General Science at grade eight (Class VIII).
- To analyze the test items through traditional methods of item analysis.
- To determine the difficulty level of each item.
- To analyze test items through Rasch Model.
- To compare the results of both Rasch analysis and traditional method of item analysis.
- To determine test reliability.

2. Literature Review

A systematic process is used to determine the extent to which pupils achieve instructional objectives. This process is called assessment. According to Venn (2000) "Assessment is the process using tests and other measures of students' performance and behavior to make educational decision."Evaluation is a process which consists of the sub processes of measurement and assessment. Evaluation usually refers to making judgment about students' performance and behavior. According to Ebel and Frisbie (1991) "Evaluation is an information gathering process that results in judgments about the quality or worth of performance, product or activity."A better evaluation and assessment can play an important role to improve education standards. Assessment of pupil learning requires the use of a number of techniques for measuring pupil achievement. It is a process that plays a significant role in effective teaching. Assessment and evaluation provides information that is used for a variety of educational decisions. There are various instruments which are being used to collect data for assessment and evaluation. The test is a most popular and widely used instrument among these. A test is a means of measuring the knowledge, skill, feeling and intelligence of aptitude of an individual or group. According to Sax (1997) "A test is a task or series of tasks used to obtain systematic observation presumed to be representative of educational or psychological attributes."Tests are divided into two general categories: The objective item and essay type questions. For some instructional purposes, the objective items may be most efficient. Owing to their usefulness the weightage given to objective types items at all levels is 40 to 60%. In objective type items the most commonly used type is the multiple choice items. The multiple choices item consists of a problem and a list of alternative solutions. The pupil responds by selecting the alternative that provides the correct or best solution to the problem. The incorrect alternatives are called distracters.Different types of tests are used regarding to what they measure. Tests are designed to measure one of the several characteristics, learning ability, achievement, aptitude, interest or personality. Aptitudes tests are used to predict how well someone is likely to perform in a future situation while the achievement test measure the past learning. They tell us the current situation of the knowledge which has been achieved by individuals.

According to Gay (1996) "Achievement tests measure the current status of individuals in a given area of knowledge or skills." Achievement tests may be teacher made or standardized. Teacher made tests evaluate the learning outcomes and content unique to a particular class or school. In formal education we cannot ignore the importance of achievement tests. In these day's world is changing and new techniques and technologies are introduced in many fields. In our schools, achievement tests are prepared by class teachers but these tests are not reliable. These tests do not fulfill the purposes of learning. It is the need of age that valid and reliable tests should be prepared. Tests should be prepared at all levels of education.The teacher made tests can be improved by calibrating these tests. Calibration is the determination of accuracy of a measurement. In Oxford Advanced Learner's Dictionary (2002) the meaning of calibration is described as:

"To mark units of measurement on an instrument such as a 'thermometer' so that it can be used for measuring something accurately." Any test can be calibrated by item analysis method. Item analysis is a set of procedures that provides us with the estimates of validity of each item. Ebel and Frisibi (1991) have defined item analysis as: "Item analysis indicates which items are difficult, easy and moderately easy." A test construction is fruitful when a test is a reliable and valid. According to Gay (1996) "Validity is the degree to which a test measures what it is supposed to measure." There are different methods to determine the validity of a test. Unclear directions, inappropriate level of difficulty of the items, ambiguity and improper arrangement of items are factors which affect the validity of a test.To find the validity of test; item analysis is an important process. In traditional item analysis item difficulty, discrimination index and effectiveness of distracters is calculated. The Rasch model was developed by a Prof. of Math in 1960. The Rasch approach of item analysis is independent of the sample and the item. The item calibrations are sample free and the person measurement is iteming free. This quality gives the Rasch model a specific objectivity of measurement implies that the comparison of selection of the items from the relevant universe and that the comparison of items is independent of the selection of subjects. Rasch model may be effectively used with questions which are right or wrong and is applicable to multiple choice items.Rasch method of item analysis is one of the methods in which results do not depends upon the sample and item. The Rasch model was developed by a Danish Prof. of Mathematics in 1960. In Rasch Model, the probability of getting an item correct depends on the difference between person ability and item difficulty.Basically Rasch calibration sets out to place the measurement of person attainment and item difficulty on the same scale and use the same unit for both.

3. Methodology

Population: The population for the proposed study was the students of all government schools at secondary level of Multan District.

Sample: Six boys and six girls' high schools were selected to collect data. Researcher adopted the method of simple random sampling to draw the sample for the study from the population. With the help of simple random sampling 300 students were selected: 158 Male and 142 Female Students were selected as sample.

Data Collection: A 45 item test in General Science (VIII) was constructed keeping in view the structural objectives to be measured by the test. The item types, selected for the test was multiple choice objectives items. The test was administered to randomly selected students of (VIII) class. Scoring was done on a principle one item one mark. The test was administered to three hundred (300) students.

Analysis of Data: The responses were collected on the answer sheet. Answer sheets were scored by awarding one mark for each correct response. F% (Facility Index), D (Difficulty Level), ϕ (Phi-co-efficient) was calculated and items were calibrated through Rasch model. Reliability of the test was also calculated by using Kuder Richardson method.Rasch Model was applied on collected data. In this method PROX item calibration and PROX person measurement was calculated. The data were arranged in item-person tables. Item and person position was identified. Item characteristics curve was drawn between item difficulty (di) and magnitude of probability (p). Similarly person characteristic curve was drawn between person measurement (br) and magnitude of probability (p). Item calibration and person measurement was done with the help of the procedure named PROX and results were tabulated in tables.

4. Results and Discussion

To determine the validity and effectiveness of individual item, item analysis was conducted. There were 45 items in the test that was administered to 300 students. Each item was analyzed on the basis of (F) Facility index, (D) Discrimination Index and phi-coefficient (φ)

Table 1: Traditional Item Analysis

No of items = 45
No of students = 300

Item No.	F%	D	φ
1	76%	0.46	0.55
2	42%	0.52	0.53
3	80%	0.33	0.42
4	64%	0.38	0.40
5	79%	0.33	0.41
6	64%	0.61	0.64
7	65%	0.50	0.91
8	84%	0.29	0.40
9	90%	0.18	0.32
10	40%	0.29	0.30
11	57%	0.10	0.10
12	50%	0.49	0.49
13	58%	0.82	0.84
14	83%	0.33	0.45
15	68%	0.53	0.52
16	60%	0.58	0.60
17	83%	0.30	0.41
18	68%	0.54	0.69
19	62%	0.68	0.70
20	32%	0.36	0.38
21	64%	0.54	0.57
22	64%	0.56	0.58
23	84%	0.30	0.42
24	80%	0.34	0.43
25	72%	0.46	0.52
26	54%	0.64	0.64
27	38%	0.28	0.29
28	42%	0.45	0.66
29	59%	0.52	0.53
30	62%	0.58	0.60
31	34%	0.36	0.38
32	64%	0.65	0.68
33	57%	0.50	0.51
34	51%	0.12	0.12
35	74%	0.26	0.31
36	79%	0.38	0.46
37	47%	0.41	0.41
38	28%	0.41	0.38
39	64%	0.42	0.44
40	59%	0.36	0.37
41	44%	0.41	0.42
42	50%	0.32	0.32
43	44%	0.41	0.42
44	65%	0.56	0.59
45	41%	0.37	0.38

Item Calibration: Proportion correct and incorrect of item scores was calculated for each item. Logits incorrect and mean of these log its incorrect was determined. The variance of distribution from this mean was initial item calibration. The table 2 shows the detail of initial calibration.

Table 2: Prox Item Calibration

No of items = 45

No of student = 300

Item #	Item Score Si	Proportion		Item final calibration di = y.di^{o}
		correct Pi = $\frac{Si}{N}$	Incorrect 1 – Pi	
1	227	0.76	0.24	– 0.73
2	115	0.38	0.62	1.18
3	228	0.76	0.24	– 0.73
4	169	0.56	0.44	0.33
5	236	0.79	0.21	– 0.93
6	191	0.64	0.36	– 0.07
7	211	0.70	0.30	– 0.38
8	234	0.78	0.22	– 0.89
9	262	0.87	0.13	– 1.62
10	113	0.38	0.62	1.19
11	170	0.57	0.43	0.27
12	146	0.49	0.51	0.66
13	268	0.89	0.11	– 1.89
14	160	0.87	0.13	– 1.62
15	217	0.72	0.28	– 0.50
16	162	0.54	0.46	0.42
17	265	0.88	0.12	– 1.71
18	206	0.69	0.31	– 0.33
19	193	0.64	0.36	– 0.07
20	75	0.25	0.75	1.90
21	178	0.59	0.41	0.18
22	214	0.71	0.29	– 0.44
23	258	0.86	0.14	– 1.55
24	242	0.81	0.19	– 1.12
25	210	0.70	0.30	– 0.38
26	153	0.51	0.49	0.57
27	94	0.31	0.69	1.56
28	128	0.43	0.57	0.96
29	173	0.58	0.42	0.22
30	196	0.65	0.35	– 0.12
31	92	0.31	0.69	1.56
32	196	0.65	0.35	– 0.12
33	172	0.57	0.43	0.27
34	175	0.58	0.42	0.22
35	211	0.70	0.30	– 0.38
36	225	0.75	0.25	– 0.70
37	125	0.42	0.58	0.99
38	99	0.33	0.67	1.44
39	183	0.61	0.39	0.08
40	176	0.59	0.41	0.18
41	118	0.39	0.61	1.13
42	175	0.58	0.42	0.22
43	131	0.44	0.56	0.90
44	193	0.64	0.36	– 0.07
45	127	0.42	0.58	0.99

Table 3: Prox Person Measurement

Person Frequency	Block	Possible score r	Proportion		Final Measure br= Xbr^{o}
			Correct Pr = $\frac{r}{L}$	Incorrect 1 – Pr	
0	A1	1	0.02	0.98	– 4.43
0	A2	2	0.04	0.96	– 3.62
0	A3	3	0.06	0.94	– 3.20
0	A4	4	0.09	0.91	– 2.75
0	A5	5	0.11	0.89	– 2.42
0	A6	6	0.13	0.87	– 2.25
0	A7	7	0.16	0.84	– 1.89
0	A8	8	0.18	0.82	– 1.78
0	A9	9	0.20	0.80	– 1.58
0	A10	10	0.22	0.78	– 1.45
0	A11	11	0.24	0.76	– 1.30
0	A12	12	0.27	0.73	– 1.13
0	A13	13	0.29	0.71	– 1.01
5	A14	14	0.31	0.69	– 0.91
8	A15	15	0.33	0.67	– 0.81
10	A16	16	0.36	0.64	– 0.66
4	A17	17	0.38	0.62	– 0.56
11	A18	18	0.40	0.60	– 0.46
5	A19	19	0.42	0.58	– 0.38
13	A20	20	0.44	0.56	– 0.29
21	A21	21	0.47	0.53	– 0.14
10	A22	22	0.49	0.51	– 0.05
21	A23	23	0.51	0.49	0.05
10	A24	24	0.53	0.47	0.14
20	A25	25	0.56	0.44	0.27
7	A26	26	0.58	0.42	0.37
19	A27	27	0.60	0.40	0.46
16	A28	28	0.62	0.38	0.56
6	A29	29	0.64	0.36	0.66
14	A30	30	0.67	0.33	0.81
8	A31	31	0.69	0.31	0.91
7	A32	32	0.71	0.29	1.02
10	A33	33	0.73	0.27	1.13
13	A34	34	0.76	0.24	1.31
7	A35	35	0.78	0.22	1.45
6	A36	36	0.80	0.20	0.58
12	A37	37	0.82	0.18	1.73
19	A38	38	0.84	0.16	1.89
7	A39	39	0.87	0.13	2.17
1	A40	40	0.89	0.11	2.38
5	A41	41	0.91	0.09	2.63
4	A42	42	0.93	0.07	2.95
1	A43	43	0.96	0.04	3.63
0	A44	44	0.98	0.02	4.43

Person Measurement: As there were 45 items the possible score of any person was zero (0) to 45 and the test was conducted on 300 students, blocks (A_1 to A_{44}) were allocated according to score ranging from 1 to 44. The maximum and minimum score (zero and 45) were excluded. The persons having score (1) were named

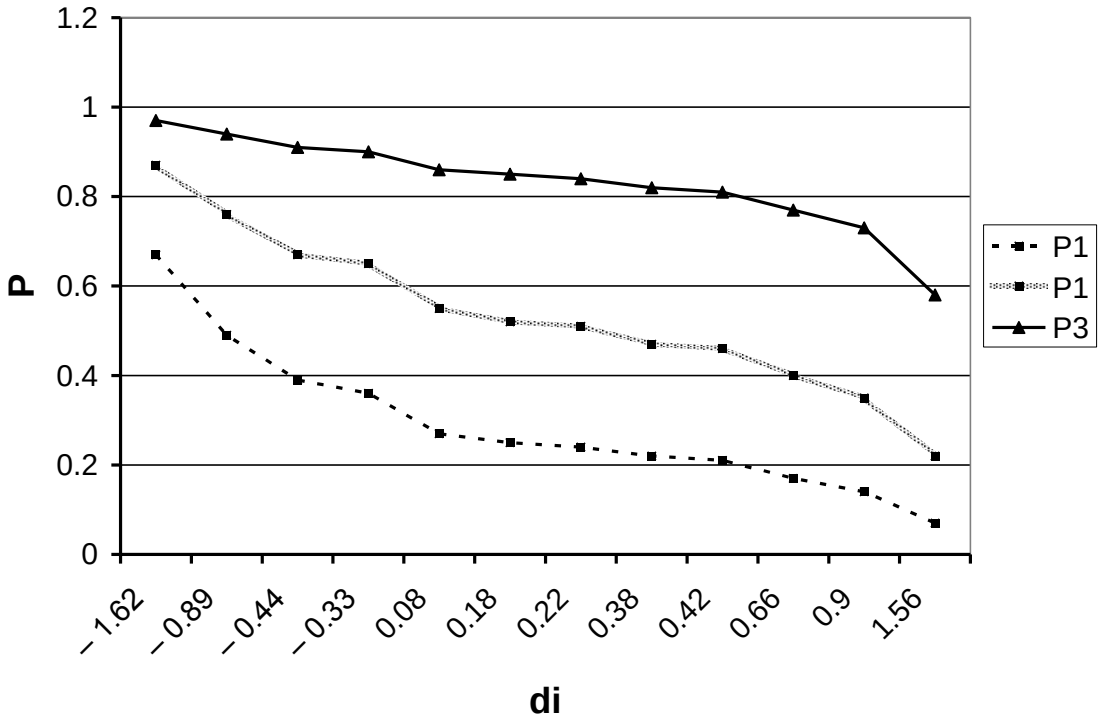
block as A_1, possible score 2 as A_2, score 3 as A_3 and so on. Proportion correct and incorrect according to blocks was determined and logits correct were calculated. These were initial person measurement values that go with each possible score on the test. Table 3 shows the detail of person measurement.

Item and Person Characteristics Curves: An ICC provides a detailed map of item functioning across proficiency level. ICC specifies a relationship between observable examinee item performance (correct and incorrect responses and the magnitude of probability of correct responses (P).A curve was drawn between final itemdifficulty (di) and magnitude of probability (P). Only 12 values from table 2 were (randomly) taken and their magnitude of probability was determined. The table 4 gives the detail of these relations. ICC curve was drawn by use of this table.

Table 4: Magnitude of Probability of Correct Response for Item Difficulty

Item No.	Di	P1 (Blok A14) br = – 0.91	P2 (For Blok A25) br = 0.27	P3 (Blok A38) br = 1.89
9	– 1.62	0.67	0.87	0.97
8	– 0.89	0.49	0.76	0.94
22	– 0.44	0.39	0.67	0.91
18	– 0.33	0.36	0.65	0.90
39	0.08	0.27	0.55	0.86
21	0.18	0.25	0.52	0.85
29	0.22	0.24	0.51	0.84
35	0.38	0.22	0.47	0.82
16	0.42	0.21	0.46	0.81
12	0.66	0.17	0.40	0.77
43	0.90	0.14	0.35	0.73
27	1.56	0.07	0.22	0.58

Figure 1: Item Characteristics Curve



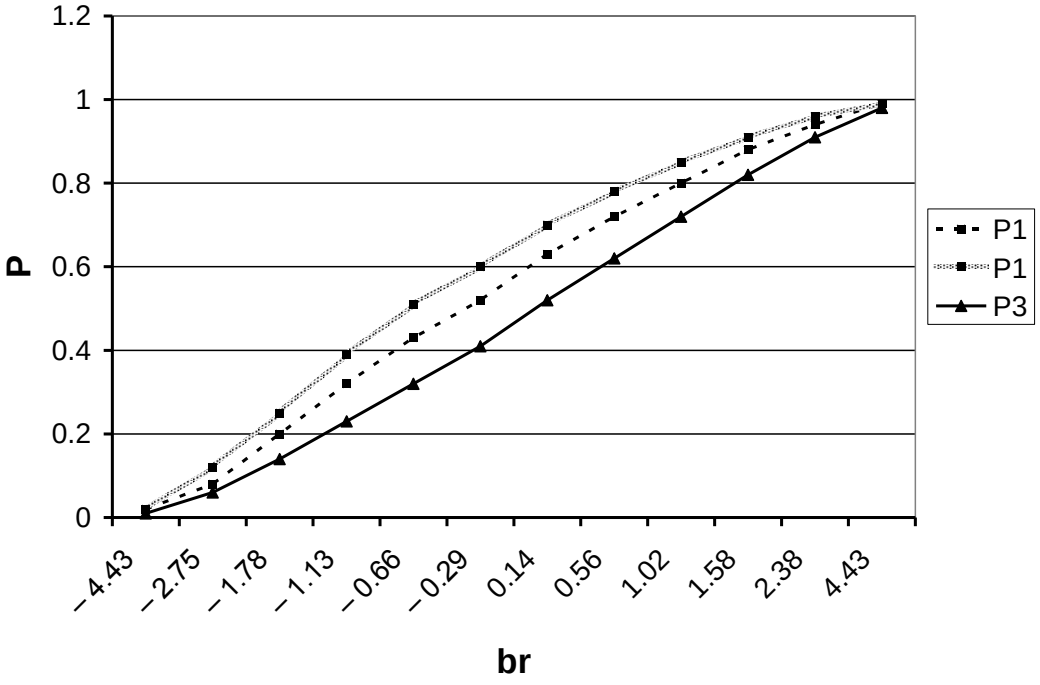
The curve is drawn between di and P_1, Curve 2 between di and P_2 while curve 3 ,was drawn between di and P_3.As the curve is s-shaped and rather steeps in its middle section therefore item discrimination is greater than moderate.

Person Characteristic Curve (PCC): A curve was drawn between person measurement values (br) and the magnitude of probability (p).There were taken only 12 values (randomly) of person measures (br) from table.3 and magnitude of probability was determined. The same formula was used to calculate (p) as in case of ICC. The table 5 gives the detail of these relations. PCC curve was drawn by use of this table.

Table 5: Magnitude of Probability for Person Measurement

Block No.	br	P (For item 25 di = – 0.38)	P (For item 36 di = – 0.70)	P (For item 44 di = – 0.07)
A1	– 4.43	0.02	0.02	0.01
A4	– 2.75	0.08	0.12	0.06
A8	– 1.78	0.20	0.25	0.14
A12	– 1.13	0.32	0.39	0.23
A16	– 0.66	0.43	0.51	0.32
A20	– 0.29	0.52	0.60	0.41
A24	0.14	0.63	0.70	0.52
A28	0.56	0.72	0.78	0.62
A32	1.02	0.80	0.85	0.72
A36	1.58	0.88	0.91	0.82
A40	2.38	0.94	0.96	0.91
A44	4.43	0.99	0.99	0.98

Figure 2: Person Characteristic Curve



The curve 1 was drawn between br and P_1 values, curve 2 is between br and P_2 and curve 3 is drawn between br and P_3.As the curve is sloppy it differentiates well between those with low and high ability persons.The results obtained indicated the following major findings:

- Item no. 2, 4, 6, 7, 10, 11, 12, 13, 15, 16 , 18, 19, 20, 21, 22, 26, 27, 28, 29, 30, 31, 32, 33, 34, 37, 39, 40, 41, 42, 43, 44, 45 are good items on the basis of the value of F (30%-70%).
- Item no. 1, 3, 5, 8, 9, 14, 17, 23, 24, 25, 35, 36 are to be improved on the basis of F (more than 70%)
- Item no. 38 is rejected on the basis of F (below 30%)
- Item no. 1, 2, 3, 4, 5, 6, 7, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 30, 31, 32, 33, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45 are good on the basis of D (0.30 – 0.70)
- Item no. 8, 9, 10, 11, 27, 34, 35, are to be rejected on the basis of D (below. 30)
- Item No. 13 is to be improved on the basis of D (above 0.70)
- Item no. 1, 2, 3, 4, 5, 6, 8, 9, 10, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29, 20, 31, 32, 33, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45 are good items on the basis of ϕ (0.30 – 0.70)
- Item no. 7, 13 are to be improved on the basis of ϕ (above 0.70)
- Item no. 11, 27, 34, are rejected on the basis of ϕ (below 0.30)
- Reliability of the test was calculated by KR # 20 and KR # 21 method, which was (0.82) and (0.85)
- The distracters B(2), B(3), B(4), D(5), D(7), C(7), D(8), B(9), C(9), C(11), D(11), A(13), A(14), D(14), D(17), D(20), C(23), A(24), A(25), C(29), D(33), C(34), D(34), C(35), D(36), D(40), A(44), B(45) were rejected on basis of that they were attracted less than 5%.
- Difficulty of items was ranging from (-1.89) to (1.90) which is calculated by Rasch Model.
- Item 13 is easiest and item 20 is the hardest item on the basis of Rasch calibration.

5. Conclusion

On the basis of findings, following major conclusions were drawn.One item was rejected on the basis of facility index (F).Twelve (12) items need improvement on the basis of facility index (F).Thirty two (32) items were very good items on the basis of facility index (F).Seven items were rejected on the basis of discrimination index (D).One item needed improvement on the basis of discrimination index (D).Thirty seven (37) items were good items on the basis of discrimination index (D).Three items were rejected on the basis of phi-co-efficient (ϕ).Two items needed improvement on the basis of phi-co-efficient (ϕ).Forty items were good on the basis of phi-co-efficient (ϕ). Test has high positive test reliability value.22 distractors were to be rejected as attempted by less than 5%.The distractors D(27), B(28), A(31), D(37), B(41), A(43) and B(10) were distracters attracted by high achievers more than low achievers, so they were rejected.

Recommendations: On the basis of the results mentioned above, following recommendations are made:Standardized tests should be modified according to our national and cultural norms.Tests should be administered at the end of the year.Item analysis techniques and Rasch model should be included into the courses of teacher training programmes.The Rasch model should be introduced to the students, and examiners through seminars, debates and workshops.Teachers should use 'The Rasch Model' for the calibration of their tests in addition to the traditional methods of item analysis and test calibration.Software's for item analysis should be developed and used for different situations.

References

Bhatt, B. D. & Sharma, S. R. (1993). Methods of Science Teaching, Delhi: Kanishka publishing, 3.

Das, R. C. (1985). Science Teaching in Schools, New Delhi: Sterling Publishers (Pvt.) Ltd., 193.

Ebel, L. & Frisbie, A. (1991). Essential of Educational Measurement, New Delhi: Prentice Hall of India (Pvt.) Ltd.23-26.

Gay, L. R. (1996).Educational Evaluation and Measurement, New York: McMillan Publishing Company, 130-153.

Oxford Advanced Learner Dictionary. (2002). Oxford: Oxford University Press, Great Clarendon Street.

Sax, G. (1997). Principles of Educational and Psychological Measurement and Evaluation, Washington: International Thomson Publishing Company, 15.

Venn J. (2000). Assessing Students with Special Needs, London: Prentice Hall International. P: 2 Available on http://www.hyperdictionary.com/test. Retrieved on 12-10-2011.