JIES Journal of Interdisciplinary Educational Studies
March 2021, V (I), 49-63

Construction of Valid and Reliable Achievement Test for Science at grade 7th Level: Applying Item Response Theory

Muqadas Bashir[1*]
Department of Secondary Education
Lahore College for Women University, Lahore.

Abstract

Assessment is the important component in any teaching system to ensure the learning progress of students. This can be done only through good instrument. The purpose of this study was to develop a valid and reliable test of science for grade 7. The test was developed by following all the test construction steps and the alignment with national curriculum was also ensured. There were 72 item in total having nature of MCQs type questions. The test was validated by experts and then pilot on students to check its reliability.

Keywords: *Valid, Reliable, Achievement Test, Item Response Theory*

Introduction

World around us is very much complicated. In order to understand world, scientific and analytical skills need to be acquired. Knowledge of science gives an insight to understand the world. Development in every aspects of human life is due to science. Life without science cannot be imagined. Science helps us in developing problem solving skills Without science we can’t imagine our life so easily and effectively. “Science is a universal part of human culture. Science provides us with a broad range of skills in problem solving, rational reasoning and flexible thinking” (Sharma and Sarita,2018, p. 1037). It is quite natural that children are curious about the world in which they live. Science education at the elementary level is built on this curiosity. Aim of science education at this stage is putting young students on the way of systematic inquiry about the world around them. As their understanding about science increases, they become able to make informed decisions. This awareness enables them to distinguish between scientific facts and fiction as they become adult. Scientific knowledge empowers students to understand social, economic and environmental issues of their world. Economic developments of the countries around the world depends upon their qualified workforce in science and technology. So, it is the demand of modern world and emerging economies to prepare students for advance scientific studies.

But in the context of Pakistan, our students lack behind in above mentioned areas. One of the reasons for this poor performance is inadequate assessment of the students in science subjects. According to Kara and Çelikler (2015) all assessment procedures have certain purposes.

- To measure the progress of the students in that particular subject area
- To identify the deficiencies of the students at the end of the course
- To assess the students' skills at the end of the course
- To determine the effectiveness of the course

Achievement tests are served as assessment tools to determine the cognitive domain of the students (Bhagat and Baliya, 2015). Different kinds of tests are used such as verbal tests (viva), supply type tests (fill-in- the blank, extended response, restricted response) and selection type tests (multiple-choice, matching the column) are used to assess the achievement of the student for students of all grades and in all the stages of higher education (Şimşek, 2009). These test types have strength and weaknesses as compare to one another. Several researches reported that that MCQ's are most commonly used type (Ogan Bekiroğlu, 2004).

There are two kinds of assessment system in Pakistan; institutional assessment and assessment by Board of Intermediate and Secondary Education. Both types of assessment used achievement tests without assessing their psychometric properties. Mostly assessment just measures the knowledge component (cognitive domain) of Bloom's taxonomy. Representation of all other components is far less. Item measuring higher order domains are missing in achievement tests. These tests are constructed without any alignment with national curriculum standards, benchmarks and SLO's. The present study was conducted to construct and standardize an achievement test in Science for VII grade students to measure their achievement.

Test Construction Process

Achievement test is constructed while using the science framework of National Assessment of Educational Progress (NAEP). Specific subject content and skills students need to be acquired and can be defined through framework. It is necessary for the theoretical basis of all assessment and designates the kinds of items that should be constructed and mentioned about design and scoring procedure of that items. Development process of frameworks caters the current requirement of the education. That is why one of the important characteristics of good framework responsiveness and flexibility.

Science framework has two dimensions; content domains and cognitive domains. Content domains include life science, physical sciences, earth science and environmental science while cognitive domains include factual knowledge, conceptual understanding and application of the knowledge in real life situations.

Alignment of National Curriculum to the Framework

Curriculum objectives of science subject is derived from specific need of the country. Content is being selected with help of these objectives. "Content includes concepts, themes, ideas, facts, principles, theories, information and skills that are to be imparted to the students for achievement of curriculum objectives. In the context of subject curriculum, this is the main body of knowledge which students are expected to learn, understand, relate, analyze, and apply. learning outcomes are identified based upon specific needs relating to that particular subject. As such, the objectives of a subject curriculum indicate as to what students should have accomplished after successful completion of curriculum of a subject. Objectives/learning outcomes should preferably be stated in behavioral terms i.e. what changes should take place in the knowledge, skills, and attitudes of students" (curriculum frame work, 2010, p 8).

Validity of the Achievement Test

Validity of an instrument means it measure what it intends to measure.Content validity is most important especially in an achievement test. "Content validity is defined as the degree to which items in an instrument reflect the content universe to which the instrument will be generalized" (Taherdoost, 2017 p.30). Literature revealed that the content validity depends upon the opinions of the domain experts (Kara and Çelikler, 2015). In present study, test was constructed by keeping in view the science framework and SLO's given in the national curriculum for grade VII. 83 items were constructed and send to the eight experts. (Whom? who are expert in the area of an assessment. mention this) 72 items were finalized after expert's review.

Reliability of Achievement Test

The term reliability means the consistency of a measuring instrument (McLeod,2013). Reliability of the test and length of the test was directly proportional to each other.Guessing factor reduce the reliability of the test. According to Ghazali (2016) "reliability on the other hand is defined as 'the extent to which test scores are free from measurement error. It is a measure of stability or internal consistency of an instrument in measuring certain concepts" (p. 149). Items are constructed according to mentioned issue. Item Response theory (IRT) was used to calculate the psychometric properties of the test. "IRT is considered as best assessment tool for construct of reading comprehension. The main difference between CTT and IRT is that CTT emphases on the total test score while

IRT focuses on performance of examinee on each item. IRT statistical models can be approved or disapproved through empirical data” (Arshed and Noureen, 2020, p. 775).

Results

=

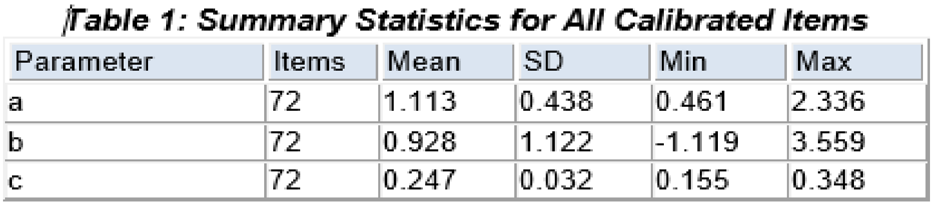


Table 1: Summary Statistics for All Calibrated Items

Parameter	Items	Mean	SD	Min	Max
a	72	1.113	0.438	0.461	2.336
b	72	0.928	1.122	-1.119	3.559
c	72	0.247	0.032	0.155	0.348

Table No.1 shows statistics parameters of all calibrated items. Table no. 2presents the summary of the total scores for the full test only for calibrated items. Table no.3reflects the theta estimates for the whole test. Table no.4gives the overall model fit chi-square(s) for the whole test.

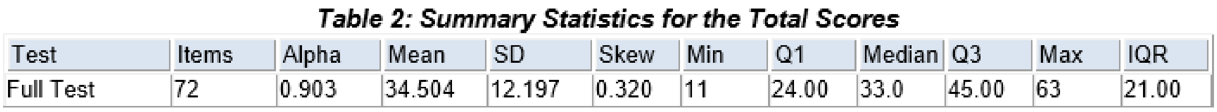


Table 2: Summary Statistics for the Total Scores

Test	Items	Alpha	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	72	0.903	34.504	12.197	0.320	11	24.00	33.0	45.00	63	21.00

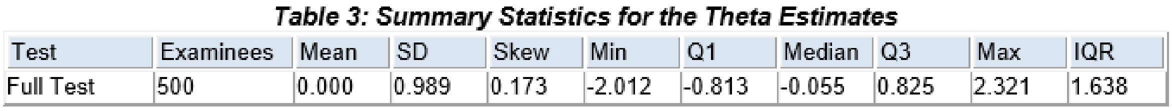


Table 3: Summary Statistics for the Theta Estimates

Test	Examinees	Mean	SD	Skew	Min	Q1	Median	Q3	Max	IQR
Full Test	500	0.000	0.989	0.173	-2.012	-0.813	-0.055	0.825	2.321	1.638

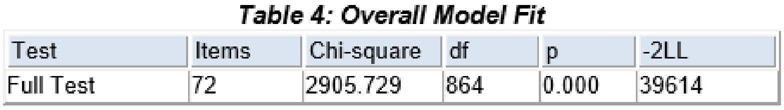


Table 4: Overall Model Fit

Test	Items	Chi-square	df	p	-2LL
Full Test	72	2905.729	864	0.000	39614

Table 5 shows the classical statistics, the item parameters, and any flags for each calibrated item.The K flag specifies that the total score did not have the highest correlation with keyed alternative. The F flag designates that the item fit statistic was significant, and the item did not fit the IRT model.The La, Lb, and Lc flags indicate that the a/b/c parameters were lower than the minimum acceptable value.The Ha, Hb, and Hc flags indicate that the a/b/c parameters were higher than the maximum acceptable value

Table 5 : Item Parameters for All Calibrated Items

eq.	Item ID	P	R	a	b	C	Flag(s)
1	1	0.556	0.255	0.461	0.688	0.277	
2	2	0.376	.270	0.828	1.420	0.241	
3	3	0.476	.421	1.141	0.811	0.301	
4	4	0.570	0.527	1.325	0.184	0.236	
5	5	0.658	0.407	0.869	-0.110	0.260	
6	6	0.566	0.348	0.709	0.330	0.246	
7	7	0.512	0.454	1.284	0.568	0.255	
8	8	0.580	0.425	1.120	0.273	0.259	
9	9	0.552	0.393	0.881	0.379	0.247	
10	10	0.378	0.570	1.961	0.765	0.195	F
11	11	0.540	0.343	1.181	0.619	0283	F
12	12	0.404	0.364	1.324	1.071	0.245	
13	13	0.554	0.531	1.285	0.278	0.239	F
14	14	0.652	0.307	0.613	-0.159	0.251	
15	15	0.594	0.543	1.606	0.158	0.253	F
16	16	0.480	0.040	0.537	1.712	0.304	K, F
17	17	0.388	0.268	0.698	1.483	0.240	
18	18	0.266	0.193	0.915	2.180	0.223	
19	19	0.342	0.243	1.604	1.456	0.249	
20	20	0.434	0.649	2.200	0.499	0.192	
21	21	0.376	0.180	0.548	1.881	0.242	F
22	22	0.398	0.440	1.181	0.936	0.220	
23	23	0.484	0.232	0.703	1.029	0.267	
24	24	0.430	0.078	0.497	1.855	0.270	
25	25	0.506	0.554	1.568	0.435	0.231	
26	26	0.320	0.079	0.785	2.758	0.271	K, F
27	27	0.352	0.415	1.347	1.174	0.221	
28	28	0.438	0.424	1.120	0.869	0.238	
29	29	0.734	0.279	0.606	-0.684	0.254	
30	30	0.312	0.029	0.857	2.558	0.264	K
31	31	0.294	0.220	1.233	1.752	0.231	
32	32	0.434	0.376	1.293	0.976	0.254	
33	33	0.476	0.349	1.119	0.872	0.267	
34	34	0.634	0.595	1.784	-0.101	0.236	
35	35	0.648	0.458	1.152	-0.054	0.259	F
36	36	0.504	0.505	1.193	0.443	0.226	

37	37	0.528	0.479	1.022	0.322	0.222	
38	38	0.346	-0.263	1.253	3.417	0.310	K, F, Hb
39	39	0.088	-0.085	1.298	3.559	0.155	K,Hb
40	40	0.438	0.368	1.514	0.912	0.254	
41	41	0.290	0.457	2.019	1.202	0.195	F
42	42	0.570	0.322	0.679	0.309	0.245	
43	43	0.188	0.216	1.610	1.922	0.184	
44	44	0.718	0.278	0.607	-0.612	0.248	
45	45	0.720	0.304	0.714	-0.514	0.257	
46	46	0.736	0.284	0.609	-0.720	0.250	
47	47	0.576	0.527	1.296	0.188	0.242	
48	48	0.460	0.598	2.336	0.501	0212	
49	49	0.718	0.481	1.237	-0.382	0.260	
50	50	0.798	0.265	0.645	-1.119	0.252	
51	51	0.410	-0.064	1.089	3.092	0.348	K, F, Hb
52	52	0.544	0.582	1.483	0.245	0.227	
53	53	0.580	0.491	1.051	0.206	0.246	
54	54	0.250	0.247	1.000	2.145	0.215	F
55	55	0.472	0.284	1.192	1.072	0.291	
56	56	0.352	0.087	0.839	2.563	0.290	F
57	57	0.584	0.407	0.748	0.168	0.239	
58	58	0.382	0.326	1.006	1.223	0.234	
59	59	0.448	.403	1.217	0.815	0.241	
60	60	0.368	0.021	0.882	2.836	0.310	K
61	61	0.164	-0.252	1.294	3.465	0.197	K, F, Hb
62	62	0.288	0.447	1.453	1.264	0.192	
63	63	0.686	0.347	0.750	-0.280	0.260	
64	64	0.708	0.359	0.743	-0.465	0.251	
65	65	0.506	0.331	0.716	0.704	0.248	
66	66	0.390	0.266	1.898	1.299	0.272	F
67	67	0.528	.271	0.626	0.629	0.250	
68	68	0.418	0.584	2.242	0.623	0.202	
69	69	0.776	0.391	0.906	-0.813	0.249	
70	70	0.236	0.216	0.782	2.665	0.215	F
71	71	0.678	0.316	0.632	-0.362	0.244	
72	72	0.344	-0.203	1.258	3.441	0.309	K, F, Hb

Figure 1: Theta estimates for all calibrated items. (SEE APA)

Table 6shows the frequency distribution for the theta estimates.

Figure 1: Theta Estimates for All Calibrated Items

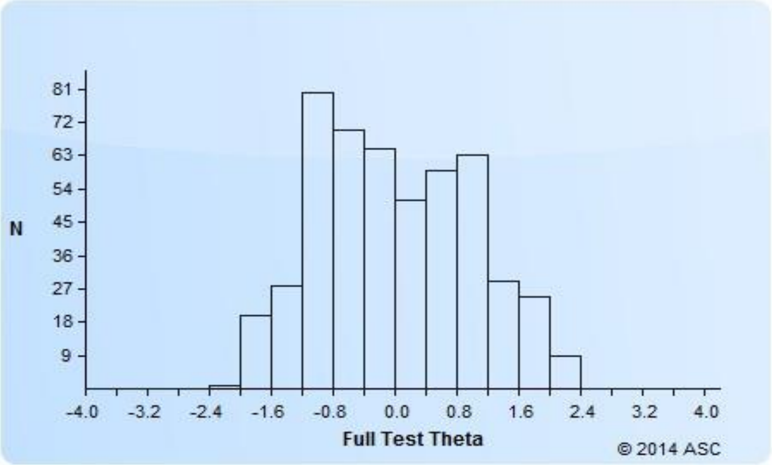
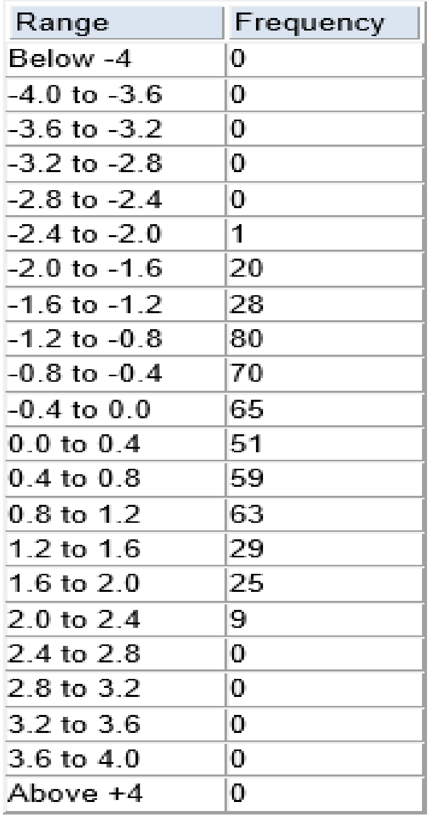


Table 6: Frequency Distribution for Full Test Theta



Range	Frequency
Below -4	0
-4.0 to -3.6	0
-3.6 to -3.2	0
-3.2 to -2.8	0
-2.8 to -2.4	0
-2.4 to -2.0	1
-2.0 to -1.6	20
-1.6 to -1.2	28
-1.2 to -0.8	80
-0.8 to -0.4	70
-0.4 to 0.0	65
0.0 to 0.4	51
0.4 to 0.8	59
0.8 to 1.2	63
1.2 to 1.6	29
1.6 to 2.0	25
2.0 to 2.4	9
2.4 to 2.8	0
2.8 to 3.2	0
3.2 to 3.6	0
3.6 to 4.0	0
Above +4	0

figure: 2 displays the distribution of the “a” parameters….

Table 7 displays the frequency distribution of the a parameters shown in Figure 2.

Figure 2: Histogram of the a Parameters

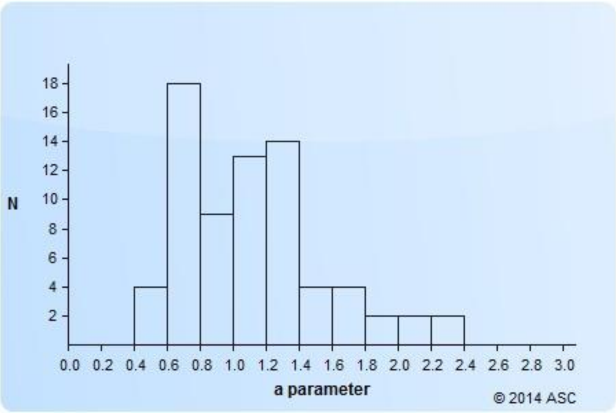


Table 7: Frequency Distribution for the a Parameters

Range	Frequency
0.00 to 0.20	0
0.20 to 0.40	0
0.40 to 0.60	4
0.60 to 0.80	18
0.80 to 1.00	9
1.00 to 1.20	13
1.20 to 1.40	14
1.40 to 1.60	4
1.60 to 1.80	4
1.80 to 2.00	2
2.00 to 2.20	2
2.20 to 2.40	2
2.40 to 2.60	0
2.60 to 2.80	0
2.80 to 3.00	0

Figure 3 displays the distribution of the b parameters.

Table 8 displays the frequency distribution of the b parameters shown in Figure 3.

Figure 3: Histogram of the b Parameters

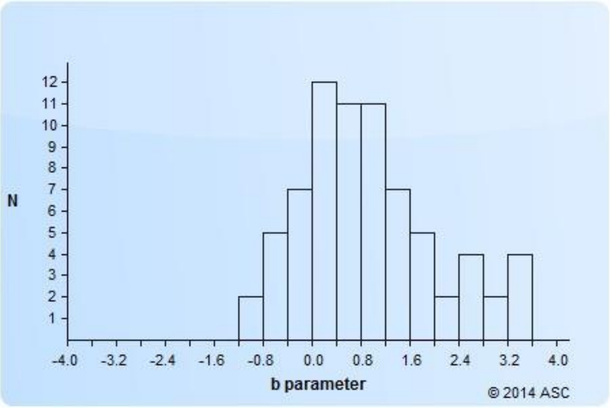


Table 8: Frequency Distribution for the b Parameters

Range	Frequency
-4.0 to -3.6	0
-3.6 to -3.2	0
-3.2 to -2.8	0
-2.8 to -2.4	0
-2.4 to -2.0	0
-2.0 to -1.6	0
-1.6 to -1.2	0
-1.2 to -0.8	2
-0.8 to -0.4	5
-0.4 to 0.0	7
0.0 to 0.4	12
0.4 to 0.8	11
0.8 to 1.2	11
1.2 to 1.6	7

1.6 to 2.0	5
2.0 to 2.4	2
2.4 to 2.8	4
2.8 to 3.2	2
3.2 to 3.6	4
3.6 to 4.0	0

Figure 4 displays the distribution of the c parameters.

Table 9 displays the frequency distribution of the c parameters shown in Figure 4.

Figure 4: Histogram of the c Parameters

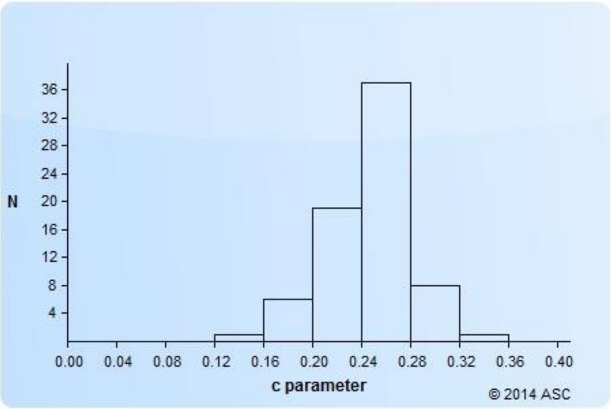


Table 9: Frequency Distribution for the c Parameters

Range	Frequency
0.00 to 0.04	0
0.04 to 0.08	0
0.08 to 0.12	0
0.12 to 0.16	1
0.16to 0.2	6

0.20 to 0.24	19
0.24 to 0.28	37
0.28 to 0.32	8
0.32 to 0.36	1
0.36 to 0.40	0

Figure 5 displays the scatterplot of the b parameter (difficulty) by the a parameter (discrimination) for all calibrated items.

Figure 5: b Parameter by a Parameter

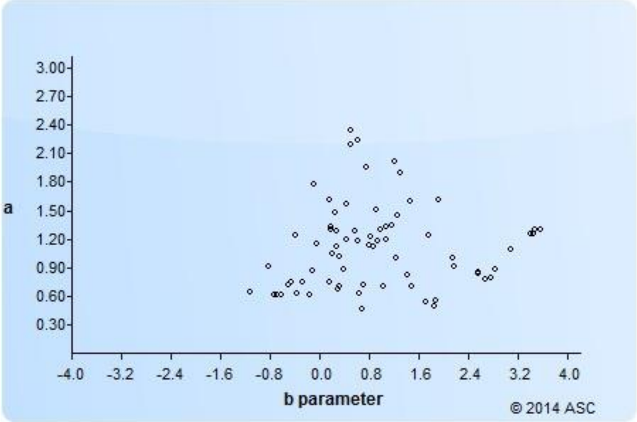


Figure 6 displays the joint distribution of the b parameter by Theta.

Figure 6: b parameter by Theta

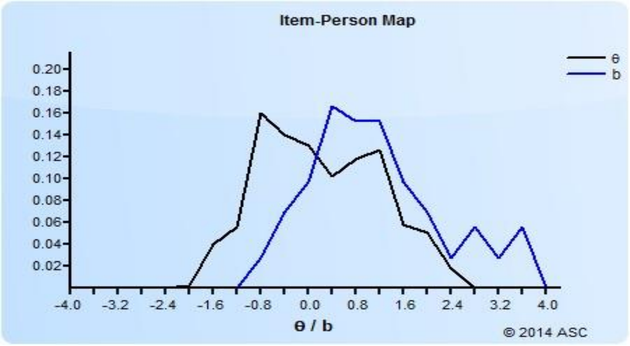


Figure 7 displays a graph of the Test Response Function (TRF) for all calibrated items. The TRF predicts the proportion or number of items that an examinee would answer correctly as a function of theta. The left Y-axis is in proportion correct units while the right Y-axis is in number-correct units.

Figure 7: Test Response Function

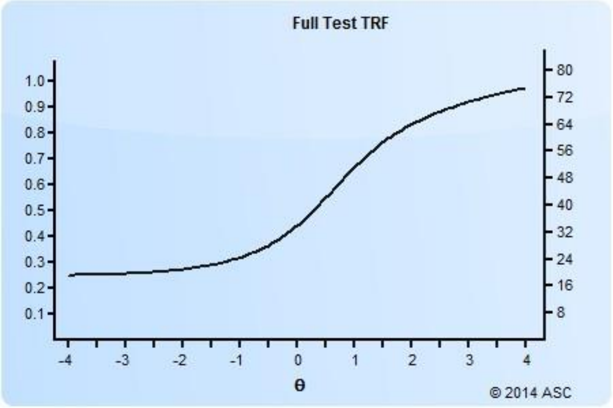


Figure 8 displays a graph of the Test Information Function for all calibrated items. The TIF is a graphical representation of how much information the test is providing at each level of theta. Maximum information was 29.543 at theta = 0.750.

Figure 8: Test Information Function

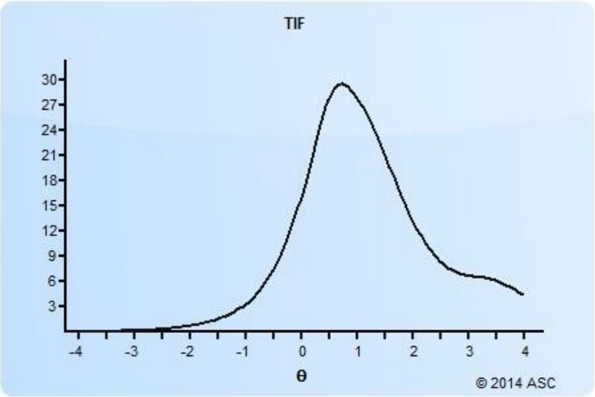
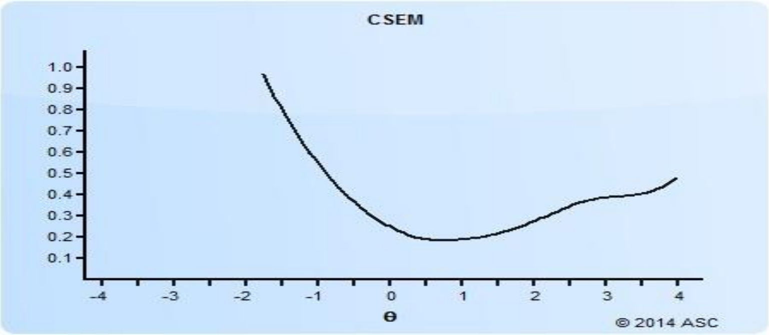


Figure 9 displays a graph of the Conditional Standard Error of Measurement (CSEM) Function. The CSEM is an inverted function of the TIF, and estimates the amount of error in theta estimation for each level of theta. The minimum CSEM was 0.184 at theta = 0.750.

Figure 9: CSEM Function



There swas total 72 items, 54 items selected on the basis of IRT criteria.

Conclusion

Assessment is one of the important aspects of education system. Continuous interaction is required in order teaching and learning. Effective assessment practices are still lacking in Asian countries. Currently, the valid and reliable instruments availability is lacking especially in Pakistan (Ghazali, 2016). Main reason is the non- availability valid and reliable instruments. “In order to carry out a successful assessment, a test with validity and reliability are ensured is required to be used” (Kara and Çelikler, 2016, p.21). Although everyone acknowledges the standing of assessment , very few teachers obtain proper training in assessment design and analysis. In USA, a survey reflected that teachers recruitment agencies mostly not required competence in assessment for licensure as a teacher. “Lacking specific training, teachers rely heavily on the assessments offered by the publisher of their textbooks or instructional materials. When no suitable assessments are available, teachers construct their own in a haphazard fashion, with questions and essay prompts similar to the ones that their teachers used. They treat assessments as evaluation devices to administer when instructional activities are completed and to use primarily for assigning students' grades” (Guskey, 2003,p.2).

Current science teaching and assessment are unable to develop reasoning or intellectual ability among science students. So, it can be concluded specialized training should be arrange for teachers to acquired competence in assessment.

References

Bhagat, P., & Baliya, A. (2016). Construction and validation of achievement test in science. *International Journal of Science and Research*, *5*(6), 2277-2280.

Filiz kara and Dilek Celikler Department of Elementary Science Education, Ondokuz Mayıs University, Samsun, Turkey Journal of Education and Practice www.iiste.org ISSN 2222-1735 (Paper) ISSN 2222-288X (Online) Vol.6, No.24, 2016

Guskey, T. (2003). How classroom assessments improve learning. *Educational Leadership*, *60*(5), 6-11.

Kara, F., & Celikler, D. (2016). Development of achievement test: Validity and reliability study for achievement test on matter changing. *Journal of Education and Practice*, *6*(24).

Simsek, Z. (2009). Organizational ambidexterity: Towards a multilevel understanding. *Journal of management studies*, 46(4), 597-624.