Pakistan Social Sciences Review
www.pssr.org.pk

RESEARCH PAPER

Development of Students' Learning Outcomes Based Standardized Achievement Test of Mathematics

Muhammad Jafar Ali [1] Abdul Hameed [2]

1. PhD Scholar, Department of Education, University of Management and Technology, Lahore, Punjab, Pakistan
2. Professor, Department of Education, University of Management and Technology, Lahore, Punjab, Pakistan

PAPER INFO

Received:
January 09, 2020
Accepted:
March 18, 2020
Online:
March 31, 2020

Keywords:
Standardized, Mathematics, Achievement test, Learning Outcomes, Classical Test Theory (CTT), Item Response Theory (IRT), Rash Model

Corresponding Author:
mjafar33@yahoo.com

ABSTRACT

The present study was undertaken to promote a research on standardization of tests items for testing learning outcome in the subject of Mathematics at secondary level. An item bank consisting of 140 dichotomous items was developed using modern test theory-Item Response Theory (IRT). Each item was aligned, on the basis of Blooms' taxonomy, with a unique students' learning outcome SLO of National Curriculum for 9th Grade Mathematics. Three Item's booklets were piloted over 960 students from four strata male, Female, rural and urban. Pilot data was analyzed using Classical Test Theory (CTT) and Item Response Theory (IRT) based software and sixty items were selected for final test based on item difficulty and discrimination indexes. Total 320 students were participated from 09 districts of the Punjab and four strata—male, Female, rural and urban. Data was analyzed by using SPSS, Conquest, and Multilog software. IRT based person item map showed the test item covered -3 to +3 range of abilities. Test characteristics and item characteristics curves and factor analysis supported the reliability and validity of the test. It explored that Math achievement test is appropriately constructed. One may replicate this study by using 2PLM and 3PLM of Item Response Theory (IRT) models.

Introduction

The testing and evaluation is one of the essential pillars of an education system. In any effective educational system, testing and evaluation complements teaching. Siddiqui (1998) mentions education as a recursive process which starts from teaching and ends at testing which is undertaken to assess the learning during this process. The test is a unique method of measurement and evaluation. It is often used to assess the achievement of the students with respect to retrieved knowledge and skills. Cal loch and Crook (2008) declare tests as tools to assess the learners in any system of education. Earl (2003) confirms the scientifically image of standardized tests and claims that the same have dominated the horizon of education since last 50 years. Payne (2003) confirms that the standardization of tests and their popularity have prevailed for many decades because these the experts has maintained quality, technicality and convenience though this system. Undoubtedly, we may find numerous professionals engaged in developing standards of testing and standardizing the system throughout the history. Sharma (2005) declares it the most prominent movement and tendency of the present age that we have focused on the standardization of the tests. Singh (2005) suggests focusing upon the intended outcome of learning while deciding items for a standardized test. Jha (2005) determines it the responsibility of the test developer to ensure the provision of useful information and supporting evidence to choose, administer and interpret standardized tests. Bradfield (2007) reveals that standardized norm-referenced testing motivates students and distinguishes the levels of individual test takers when selection of the suitable is an objective. Munn and Butler (2006) have the point of view that curriculum outlines process for education and knowledge. It comprises of ideas, standards and benchmark that describes set of parameters of student achievement or students learning outcomes. Therefore, it is essential to understand concepts, standards and benchmark of the curriculum. The evaluation taken under the current system of examination is considered as smaller part of the results of teaching and learning process which promotes only rote learning rather than the learning and understanding of the learners associated to their real-life situations and application. It keeps the students away from any effort to improve their learning. Thus, students lack behind in developing application, critical thinking and positive attitude or study habits during their study period (Memon, 2007, Government of Pakistan, 1996). Day by day, educational outcomes measures/instruments are developed or revised from previous instruments/measures to improve more reliability, validity, sensitivity and interpretability. This increasing demand of powerful psychometric measures demands for better analytical measuring tools beyond the scope of classical test theory (CTT) (traditional measurement theory) methods can provide. IRT is a model for exploring the association between an individual's responses to an item and “ability” or “trait” of individuals being measured by the instrument.

Material and Method

How many number of items are appropriate are needed to true measurement of psychological construct is a question of reliability and validity. Classical Test Theory (CTT) based measurement models depend upon number of items in a test while Item Response Theory (IRT) based measurement models depend upon responses on test takers. The most famous (IRT) Rasch model recognizes test takers in terms of accomplishment/capability and recognizes the items included in the test from their difficulty perspectives. The model uses the same scale for measuring the both aspects. The model discovers the chances of the candidates' particular answers to a question (Item) as a function of the responders' aptitude (Ɵ) through unfolding the relationship of the item to Ɵ. (Thissen Steinberg and Mooney, 1989; Sireci, Thissen, and Wainer, 1991; Fennessy, 1995). Item response theory is based on a set of fairly strong and testable assumptions. If these assumptions are not met the usefulness of item response theory is compromised.

These are the following assumptions:

1. Uni-dimensionality: It is called "one factor" model. This assumption tells that the test of interest measures only one construct e.g; reading skills, math achievement.
2. Local independence: This assumption states that item responses are independent of one another.
3. Invariance: This assumption gives information that item characteristics are constant among various subgroups.
4. Nature of the ICC: for dichotomously scored test items (0 or 1) logistic functions are used to model the probability of ''success'' (i.e.0 or 1).

In this study, both CTT and IRT test theories are used for test construction.

Population and Sampling

This study is survey in nature in the field of test development. It is related to educational measurement and psychometrics. All those students of class 9th enrolled in (science group) for the academic session 2013-14, who were studying mathematics (as an elective subject) in high school (9th& 10th classes) and higher secondary schools (HSS) or colleges of Punjab, are included in the population for the study (see Table No.3.1)

Table 1
Location and Gender Wise Target Population

Class	Rural		Urban		Total	
9th	Boys	Girls	Boys	Girls	Boys	Girls
	151860	157200	211873	157980	363733	315180

(Department of Education, PMIU, PESRP, Govt. of the Punjab, 2014)

There were 678,913 students enrolled in class 9th in state government schools in the Punjab province that comprise the population of the study. Representative sample is necessary for IRT based models because IRT based model are based on responses of test takers. In education testing, existence of different strata such as rural-urban, gender, age or residence in the population may give premise for including a stratified sampling (Best, 2005). For this study, the schools included in sampling were considered as clusters because it is the most suitable technique when the individual subjects (Population) are broadly dispersed in terms of geographical distribution (Best, 2002). The detail of sampling is as under:

The Province of Punjab is comprises of 36 administrative units called districts. Punjab can be divided into three geographic regions: Northern, Central and Southern Punjab. From each region/part three districts were taken randomly. From Northern Punjab; Rawalpindi, Gujranwala and Lahore; from central Punjab; Sahiwal, Pakpatan and Toba; from Southern Punjab;D.G. Khan, Multan and Rahim Yar Khan.

Eight secondary schools were selected through stratified random sampling technique on the basis of location and gender from all Government secondary schools situated in each sample district. From each selected school 40 students were randomly included in the sample with the help of Random Number Table developed by NEAS (2007). A sample of 320 students was obtained in each district. So, a sample of 2880 was obtained from 72 schools to whom the test was administered.

Instrumentation

Bloom taxonomy of learning objectives is adopted for this study. Bloom taxonomy is implemented for determining taxonomical levels of the SLOs of Pakistan National Math Curriculum for 9th class are constructed.

The test result can provide more accurate information about what students have actually learned. Test items developed are SLOs based. The following preferences were adopted to develop the achievement test.

The objectives of national curriculum were being kept in view.

- A table of specification is prepared for distribution of content selection to construct the test item from whole text book.
- While preparing the test, only the three levels of cognitive domain of Bloom's taxonomy are considered more appropriate i.e. Knowledge, Comprehension and Application.
- The researcher selected four optioned multiple-choice item tests to provide better test setting, marking and standardization. It can measure inference and reasoning understanding and judgment. It can be scored rapidly. Most of the test specialists believe that it is the best test of objective test items.

The main purpose of an item is to create intended answers by the students from which inferences can be made (Fulcher and Davidson, 2007). The researcher developed 140 items in this math achievement test. Before piloting all the 140 items were reviewed by the experts presented in the department. These all 140 items have association with the math achievement framework and the national math curriculum 2006.

For the pilot study, 140 items were randomly distributed in three booklets with 20 items common in all three tests. The three tests were piloted in eight schools. Tests were administered on 960 students (480 from urban secondary school and 480 from rural secondary schools).The pilot study was conducted in the district Sahiwal of the Punjab. Piloting of the test (Math achievement test) was done by the researcher himself to confirm the protected and quality based analysis. Initially, the data entry was done using SPSS program and then it was shifted to IRT based computer program (Conquest) and CTT based computer program (ITMAN).The selection criteria for final 60 items were as under:

Initially all the 20 common items were chosen. Only 40 out of 120 items were according to the criteria based on the values of IRT: range for infit and outfit, CIT difficulty range, and the value as per Webb's alignment criteria. Key indices used for finalization of items for final data are given in table 3.5.

Table 2
Item Selection Criteria

IRT Based	CTT Based
Range of "b" -3 to +3	Difficulties index 0.30 to 0.8.
Mean – square Range of "Infits" 0.80 to 1.30	Discrimination 0.20to 0.80
Mean square Range of "Outfits" 0.80 to 1.30	Point – based > 0.8

Only the items were finalized which were assumed to be allied with the six conditions with high priority. And the items which were allied with 3 conditions including two IRT based conditions were included in the finalized items with secondary preference.

Development of Rubrics / Scoring Guide

The SCRQs were dichotomous in nature. The right answer choice or responses of some items/questions may be formed numerically (in figures) or textually (in words). For proper marking of the participants' answers a rubric was created for marking SCRQ. The rubric was endorsed by the experts who designed the items (question), subject experts, and testing & assessment experts.

Conduct of Test

In addition to the development of math achievement test, the researcher also developed a manual for test administrators with the assistance of faculty members of University of Management and Technology, Lahore. The manual provided the test administrators with all the procedural guidance from receiving the test material upto the final stage. The piloting was carried out for the manual and tryout test. The feedback for the piloting was collected through the feedback Performa from the administrators, also supplied with the test material. The feedback was analyzed and manual for test administrators was modified accordingly. The manual was also piloted along with the piloting of the math achievement test. During piloting, the test administrators recorded their observation and suggestions in the attached feedback Performa. In the light of received feedback, the test administrators' manual was further improved and finalized for the large-scale testing. Due to financial constraints, the researcher was forced to select already trained test administrators. The most of the selected test administrators had test administration training from NEAS and PEAS. Training for untrained test administrators was arranged at University of Management and Technology, Lahore, and Government Colleges of Elementary Teachers (GCETs) of Punjab. Prior to administration of the test, each test administrator was provided with a copy of the test administrators' manual. All the test administrators were linked with the researcher via a control-room, run by two M.Phil students, established at UMT, Lahore. The control-room heads asked each test administrator if he/she had any reservation regarding the test administration. Two days prior to test administration, the control-room was satisfied with the arrangements. The researcher was also present in the control-room on the day of the test. The researcher applied in written, through proper channel, to the Education Department, Govt. of Punjab, and seeking official permission to collect data for this study from students of grade 9, enrolled in public secondary and higher secondary schools. A bonafide permission letter was issued by the Chairman of department of education UMT Lahore for granting permission to collect the data. The copy of the letter was sent to the heads of all sampled schools, the respective District Education Officers, and all the test administrators. Services of the Pakistan Post were used to send the test materials to all the test administrators. Receiving of the test material was confirmed by the control room two days earlier.

The test was administrated in the third month of 2014 (March, 2014). The test administrators selected 20 students each according to the random table provided along with the manual. All the selected students were provided a pencil each for the test. The test administrators told and explained the instructions and answered the students' queries. Before starting the test, the participating students were asked to fill up the initial page asking demographic and other required information. Then, the participants were asked to write their responses for each question item in the test. The maximum time limit was 60 minutes for the 60 questions in the test. The test administrators sent back the administered test copies to the control room through Pakistan Post as registered parcels. The postal charges borne by the test administrators were refunded in the shape of prepaid credit on their mobile SIMs.

The test copies were received by the researcher. The test material was sent 72 sampled schools. In addition to the sampled schools, test material was also sent to 7 schools (10 percent of the sampled school) to compensate data in case a test administrator failed to conduct the test, or in case missing a parcel in during delivery and receiving of the test material. The data was received from 72 schools in which a total of 2880 students had taken the test. The reason behind slightly lower number of students who took the test was because in few schools, the total number of grade 9 students was less than the required 40. All 2880 students were included in the analysis on the basis of their responses to the test items.

The Data was compiled using a computer spreadsheet program (Microsoft Excel). Two operators were assigned the data entry task, they were also instructed for the test rubric of SCRQ test items. The validity of any test denotes the level to which it truly scales what it purposes to measure. It is also the degree to which the decisions are made, implications, inferences and conclusion based on the output scores of the test are suitable and meaningful. The notion of validity is uniform entity, and it can be regarded through observing major evidences of validity such as "content" validity, "construct" validity, and "consequential" validity. These types sometimes incorporate supplementary notions of validity. "Face validity" and "Curricular validity" ought to be accomplished to observe "content" validity of any test. "Convergent" validity can be investigated for the "construct" validity. It can also be established through criterion-related validity and sometimes also through the evidence from content of the test. Whereas, the validity of a test is a decision made subjectively which solely rely on personal experiences rather than empirical indicators.

The expert opinion from the committee of test developers and the results of the analysis on the basis of IRT were brought under consideration for the validation of the test used in this study. The evidences for the validation were extracted from statistical scores (infit & outfit), characteristics curves (ICC), test characteristics curves (TCC), test information function (TIF) curve, and factor analysis.

In IRT models, item characteristic Curve ICC represents the location parameter (b) on θscale where the curve passes for p=0.5. so persons whose trait value exceeds the location parameter of the item having greater than a 50% chance of a positive response, while the persons whose θ values lies below the location have less than50% chance of a positive response. In the context of achievement tests, the location of an item corresponds to its difficulty: the higher the location parameter, the more achievement is required before the examinee has a 50% chance of a correct response. The sample size required for useful item calibration varies widely, depending on the format of the response and the strength of the relationship between the item response and the trait. IPL Rasch's model only requires a few hundred examinees for the test calibration. For the test calibration, there is only the requirement of few hundred test takers. There are two ways (Fixing the θ value or selecting it randomly) of carrying out item parameter calibration.

The fit test for the model may be carried out for the whole test or individual items with the availability of the data collected from huge sample population. Fit testing method is dependent upon the number of items included in the constructed test. EAP estimation is adequate for a test containing more than 20 items to carry out items calibration. At the last phase 'X^2 test' statistics are used for the comparison of frequencies of improper and proper answers in the interval with the interval means of those from the model which is expected to be fit. If the values of

'X^2' are significantly high, the response model is considered not fit in terms of the provided number of items. Unlike the analysis on individual basis, the combined level allows a difficult test of fit of the answer (response) pattern for the cluster.

A likelihood ratio of Pearson X^2 test may examine the fit of the model within each stratum for the reason that the frequencies of the answers for the items (Questions) of a measured component are independent and distributed in binomial nature.

Since, the size of sample and the pattern of responses highly sensitize the statistics of person X^2 therefore it is assumed as the only one feature of fit statistics. The adjoining estimate to reliability present from perspective of levels of θ is stated through MULTILOG, which is marginal value indicating the average reliability. If the condition in which the information of the test is uniform, the above-mentioned description of accuracy for the scale is considerable. For the selection and calibration of items as per Rasch's model, during the estimation for a reliable and valid test, maximum likelihoods estimations (<LE), separation reliability coefficient, RK # 20 reliability coefficient, item's discrimination power, items difficulty index and marginal reliability were brought under consideration. Test score of students were calibrated with a scale 0-1000 with mean 500 using the following formula.

Achievement Score = $\boldsymbol{\theta_{target} = A.\theta_{observed} + B}$

Where

$\boldsymbol{\theta_{observed}}$ **=** achievement observed = latent variable = value of $\boldsymbol{\theta}$ on item person scale

A = multiplying factor = 100, B = mean of scale (500)

$\boldsymbol{\theta_{target}}$ = calibrated value of $\boldsymbol{\theta_{observed}}$

Using "bookmarks" method, following levels of achievement were identified.

Level of Achievement	Minimum Scale Score
Basic	≤ 400
Proficient	400-650
Advance	≥ 651

Limitations

In spite of running the math achievement test smoothly, there are certain limitations of this study.

- Lack of funds.
- Lack of enough human resources.
- Exclusion of Practical geometry portion due to insufficient material availability.

- Non-co-operation of school staff in conducting test due to chance of identification of poor student's performance.

The limitation listed above might not be manageable by the researchers. These can be resolved through enough research grants.

Computer programs used for both piloting and the final analysis include ITMAN, Conquest, and Multilog. Items, for the final test used for the study, were selected on the basis of piloting results. The features of the test items and hypotheses of the study were checked through analyzing the final data collected at large scale

Results and Discussion

The amount of 3.62% of the total data, not brought under consideration for the analysis, was missing. The statistical measures and tests given below are considered to be applied for the current design of the study which also suits the data for containing minimal data amount which is missing. The results given below are computed to make sure, each item the provided response was a single.

Table 3
Descriptive Statistics of Math Achievement Test

Statistics	Value
N	2880
Mean	37.08
Standard Deviation	11.73
Variance	137.61
Skewness	-0.29
Kurtosis	-1.00
Standard error of mean	0.22
Standard error of measurement	3.43
Coefficient Alpha	0.86

Table 4
Reliability Indices Generated by Different Software

Reliability	Index	Software
Item Separation reliability	0.891	Conquest
MLE Person separation Reliability:	0.875	Conquest
WLE Person separation Reliability	0.884	Conquest
Cronbach Coefficient Alpha	0.860	Conquests
Cronbach Coefficient Alpha	0.883	SPSS
Marginal Reliability (1PL)	0.848	Multilog
Marginal Reliability (2PL)	0.903	Multilog
Marginal Reliability (3PL)	0.998	Multilog

Person separation reliability, which is marginal reliability in accordance with item Response Theory, was analyzed through Multi log software. Maximum Likelihood

Estimation (MLE) reliability and Weighted Likelihood Estimation (WLE) are somewhat dissimilar but are used almost for the same purpose. The high reliability score (0.891) for item separation recommended that there was adequate number of items with diverse difficulty points in the math achievement test. High reliability score advocates that the tested math achievement test (MPT) covers diverse levels of achievement because the difficulty level of items is directly associated with scaled achievement. It also gives insights into the validity of the test contents. The reliability score for person separation, which is measured high (0.87), interprets that the students with diverse capabilities are included in the sample taken. It proves to be suitable in terms of sampling. According to the reports of (Multilog analysis), 3PL model indicates the consistency among the items of the test higher in comparison with 2PL and 1PL (Rasch model). Items are analyzed in terms of difficulty levels by applying 1PL (Rasch Model) and discriminating power is added as extension in analysis by 2PL model while 3PL model uses the three features; difficulty levels of items, guessing of the test items and their discrimination. Items guessing and discrimination power are excluded in the interpretation of the data, therefore the overall analysis is delimited to items difficulty levels only.

Concluding all, the screened reliability scores were generally high for the sixty items of the constructed math achievement which reports the presence of internal consistency across items. The reliability scores for Items and Person separation indicates that inclusively the test contains satisfactory reliability and validity irrespective of the model fit.

Table 5
Item Difficulty and Discrimination Analysis

Sr. No.	Difficulty Level	Discrimination	Remarks
1.	0.65	0.37	Appropriate
2.	0.48	0.39	Appropriate
3.	0.47	0.45	Best Item
4.	0.57	0.52	Best Item
5.	0.62	0.5	Best Item
6.	0.87	0.33	Most easy
7.	0.76	0.49	Most easy
8.	0.79	0.31	Most easy
9.	0.83	0.30	Most easy
10.	0.90	0.41	Most easy
11.	0.79	0.47	Most easy
12.	0.49	0.47	Best Item
13.	0.62	0.32	Appropriate
14.	0.72	0.35	Appropriate
15.	0.77	0.37	Most easy
16.	0.69	0.53	Appropriate
17.	0.60	0.49	Best Item
18.	0.63	0.49	Appropriate

19.	0.77	0.37	Most easy
20.	0.64	0.41	Appropriate
21.	0.70	0.59	Appropriate
22.	0.35	0.49	Appropriate
23.	0.60	0.57	Best Item
24.	0.81	0.48	Most easy
25.	0.73	0.36	Appropriate
26.	0.51	0.44	Best Item
27.	0.82	0.45	Most easy
28.	0.74	0.53	Appropriate
29.	0.63	0.53	Appropriate
30.	0.56	0.47	Best Item
31.	0.58	0.28	Appropriate
32.	0.66	0.39	Appropriate
33.	0.63	0.59	Appropriate
34.	0.56	0.54	Best Item
35.	0.66	0.58	Appropriate
36.	0.61	0.57	Appropriate
37.	0.65	0.51	Appropriate
38.	0.33	0.23	Not Appropriate
39.	0.64	0.49	Appropriate
40.	0.73	0.48	Appropriate
41.	0.66	0.54	Appropriate
42.	0.61	0.45	Appropriate
43.	0.62	0.52	Appropriate
44.	0.34	0.21	Not Appropriate
45.	0.56	0.58	Best Item
46.	0.66	0.49	Appropriate
47.	0.56	0.44	Best Item
48.	0.45	0.31	Appropriate
49.	0.79	0.4	Most easy
50.	0.73	0.49	Appropriate
51.	0.75	0.44	Appropriate
52.	0.25	0.26	Not Appropriate
53.	0.21	-0.18	Worst
54.	0.60	0.51	Best Item
55.	0.72	0.36	Appropriate
56.	0.61	0.31	Appropriate
57.	0.24	-0.17	Worst
58.	0.35	0.36	Appropriate
59.	0.62	0.42	Appropriate
60.	0.63	0.56	Appropriate

Conclusion

Summary of Item Characteristic Curves

Characteristics curves generated by Multilog software explore that majority of items have discrimination power very close to each other and difficulty level within ±1. Even each item of Math achievement test has different difficulty level but test has mean /average difficulty (θ=0).The probability of guessing chance of more than 60% items less than 0.1 and probability of the item no. 1, 5, 6, 7, 8, 9, 11, 14, 15, 19, 24, 25, 27, 28, 46, 49, 51 is in between 0.1 and 0.3 which is appropriate. Item No. 38, 44, 52, 53, and 57 are not good items as shown in Table 5. They should be deleted from final test. So final test will consist of 55 items.

Overall Test Difficulty = 0.62 and Overall Test Discrimination = 0.42. Above table explores that 55 out of 60 items are acceptable. Five items should be deleted from final test. So the final test consisting of 55 items is reliable test. The Overall Test Difficulty of selected 55 items is 0.65. Overall Test Discrimination selected 55 items is 0.45.Thus, over all test is well discriminator but relatively easy test

Recommendations and Suggestions

The recommendations given below are based on the findings and conclusions of the study.

1) This test can be used by teachers of 9th grade during classes as math achievement test for the assessment as summative or it can partially be used as formative assessment.
2) This math achievement test can be used by education department of Punjab, Pakistan and boards of intermediate and the secondary educations for the examinations and assessment of 9th grade student's achievement and their learning problems.
3) This test can be recommended to understand 9th grade students' mathematics achievement.
4) This test can be included in teachers training or in the process of selection of 9th grade teachers.
5) This methodology which is used for the math test achievement can also be used for constructing other subject's tests.
6) This test can be used for a certain period of time to check student's math achievement.
7) For better construction and administration authorities can train teachers accordingly

References

Best, J.W.,&Kahn, J.V.(2006). *Research in Education* (10thed.).New York: Pearson Education Inc..

Breadfield, P.(2007*). Introduction to assessment.* New York: Continuum International Publisher Group.

Earl, L. M. &LeMahieu, P. G. (1997). Rethinking assessment and accountability. In A. Hargreaves (ed.), Rethinking educational change with heart and mind (pp.149-168). Alexandria, VA: Association for Supervision and Curriculum Development.

Earl, L.M. (2003).*Assessment as learning: Using classroom assessment to maximize student learning*. Thousand Oak, CF: Crown Press Inc. Press.

Fischer, G. (1968). *Psychologische test theorie [Psychological test theory*]. Bern: Huber.

Fulcher, G and Davidson, F. (2007). Language Testing andAssessment:An advanced resource book. Routledge

Government of the Pakistan. (2006). *Math National Curriculum*. Islamabad: Ministry of Education.

Mathematics Framework for (2003)*National Assessment of Educational Progress*. National Assessment Governing Board. USA: Department of Education

McMunn, N. D.,&Butler, S. M. (2006). A teacher's guide to classroom assessment: Understanding and using assessment to improve student learning. New York: John Wiley and Sons, Inc.

Messick, S. (1989).*Validity.* In R.L. Linn (Ed.), Educational Measurement (3rded.), (pp 13-103). New York: McMillian.

Messick, S. (1994). The interplay of evidence and consequences in the validation ofperformance assessments. *Educational Researcher, 23*(2), 13-23.

Mousavi, S. (1999). *A dictionary of language testing*. (2nded.). Tehran: Rahnama Publications.

Government of Pakistan. (2006).*National Curriculum for Mathematics I-XII.* Islamabad: Ministry of Education.

Jha, P.K. (2005). Assessment and Evaluation in I-Iigher Education, Delhi: VISTA. International Publishing 1-Louse.

Payne, D.A. (2003). *Applied Educational Assessment*. (2nded.) Toronto: Wadswarth Pearson Education.

Saddiqui,M.H.(2006). *Teaching of Mathematics*. New Dehli:A.P.H. Publishing.

Shah, R.K. (2007).*Educationaltesting and measurement.* Jaipur:Pointer Publishers.

Shah, D.,&Afzaal, M. (2004). *The examination board as educational change agent: The influence of question choice on selective study.*Paper presented at 30thannual IAEA Conference. Philadelphia, USA.

Singh, Y.K. (2005). *Psychology in Education*. New Delhi: A.P.H. Publishing.

Steinberg, L.,&Thissen, D. (1995). *Item response theory in personality research*. In P. E. Shroutand S. T. Fiske (Eds.), Personality research, methods, and theory: A festschrifthonoring (pp. 161-181). Hilldale, NJ: Erlbaum.

Stedman, L. C. (1996). The achievement crisis is real: A review of the manufactured crisis. Retrieved from http://epaa.asu.edu/epaa/v4n1.html on 2-09-2008

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment.*Applied Measurement in Education, 4* (4), 263-273.

Thissen, D., & Orlando, M. (2001). *Item response theory for items scored in two categories*. In D. Thissenand H. Wainer (Eds.). Test scoring(page numbers). Hillsdale, NJ: Erlbaum.

Thissen, D.,& Steinberg, L. (1989). Data analysis using item response theory. *Psychological Bulletin, 104,* 385-395.

Thissen, D. (1991). *Multilog User's Guide. Multiple, categorical item analysis and test scoring using Item Response Theory,*

Waner, V. J. (2004). *Development and validation of a performance-based assessment in work and family life personal development.* Doctoral Thesis. The Ohio State University.